Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research

Venceslas Douillard1, Erick C. Castelli2, Steven J. Mack3, Jill A. Hollenbach4,5, Pierre-Antoine Gourraud1, Nicolas Vince†1* and Sophie Limou1,6† on behalf of the Covid-19|HLA & Immunogenetics Consortium and the SNP-HLA Reference Consortium

1Centre de Recherche en Transplantation et Immunologie, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, Université de Nantes, Nantes, France, 2Unesp—Universidade Estadual Paulista, Botucatu, Brazil, 3Division of Allergy, Immunology and Bone Marrow Transplantation, Department of Pediatrics, School of Medicine, University of California, San Francisco, San Francisco, CA, United States, 4Department of Neurology, University of California, San Francisco, San Francisco, CA, United States, 5Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States, 6Ecole Centrale de Nantes, Department of Computer Sciences and Mathematics in Biology, Nantes, France

The current SARS-CoV-2 pandemic era launched an immediate and broad response of the research community with studies both about the virus and host genetics. Research in genetics investigated HLA association with COVID-19 based on in silico, population, and individual data. However, they were conducted with variable scale and success; convincing results were mostly obtained with broader whole-genome association studies. Here, we propose a technical review of HLA analysis, including basic HLA knowledge as well as available tools and advice. We notably describe recent algorithms to infer and call HLA genotypes from GWAS SNPs and NGS data, respectively, which opens the possibility to investigate HLA from large datasets without a specific initial focus on this region. We thus hope this overview will empower geneticists who were unfamiliar with HLA to run MHC-focused analyses following the footsteps of the Covid-19|HLA & Immunogenetics Consortium.

Keywords: Major Histocompatibility Complex (MHC), HLA, association analysis, imputation, immunogenetics

INTRODUCTION TO HUMAN LEUKOCYTE ANTIGENS: CREATING IMMUNITY FROM DIVERSITY

The classical HLA proteins are expressed on the surface of human cells. Although their primary role is to present exogenous and endogenous peptides, they were first described as “antigens” due to their interaction with T-cells in transplant rejection (Dausset, 1958). Along with other genes in the MHC region, the products of the HLA genes are essential in the adaptive immune response. By presenting peptides to both CD8+ (HLA class I molecules) and CD4+ T cells (HLA class II molecules), HLA proteins initiate an immune response against foreign (non-self) peptides which may be defective products of translation, neo-antigens generated by mutated genes from tumor cells, or pathogenic in origin. In addition, class I HLA proteins interact with the KIR ligands of NK cells, including KIR and LILRB, which are important in innate immunity (Carrington et al., 2008; Kulkarni et al., 2008; Trowsdale and Moffett, 2008). Thus, HLA molecules are key features of both innate and adaptive immune responses. HLA genes central role in immunity against
infectious diseases and their importance for transplantation have made them the subject of much study.

HLA proteins are coded by multiple genes on the short arm of chromosome 6 at the 6p21 locus; this region containing HLA genes is referred to as the Major Histocompatibility Complex (MHC) for its seminal role in transplantation (Dausset, 1981; Montgomery et al., 2018). Although there is a common confusion between the two terms HLA and MHC, HLA specifically refers to the genes involved in antigen processing and presentation whereas the MHC corresponds to a whole locus, with HLA and other immune-related genes such as the complement system. The MHC region is the most gene-dense region of the human genome, with 1% of the human coding genes (>200) found in 0.1% of the genome length (Shiina et al., 2009). The MHC region is commonly defined as a 4 Mb segment on chromosome 6 (MOG 29657002-33192499 COL11A2, GRCh38. p13 assembly) (Beck et al., 1999). However, due to extended patterns of linkage disequilibrium (LD), an extended MHC (xMHC) is often referred to in immunogenomics (25726063-33400556, GRCh38. p13 assembly) (Horton et al., 2004). The MHC region is divided into three regions based on gene sequence similarities and functions, class I, II, and III in which approximately 40% of the genes are immune-related. HLA genes are found in the class I and class II regions and are commonly divided in two categories: classical HLA proteins present peptides to T-cells, whereas non-classical HLA are mostly involved either in peptide presentation with other receptors, with immune modulation, or with various steps of classical HLA formation and loading.

The MHC class I region contains 12 HLA pseudogenes and 6 HLA genes (HLA-A, -B, -C, -E, -F, and -G), including three classical (HLA-A, -B, and -C) that are ubiquitously expressed as a heterodimer with beta-2 microglobulin at the cells’ surface. Class I HLA molecules and their bound peptides are specifically recognized by CD8+ T cells receptors. The non-classical HLA class I molecules (HLA-G, -E, and -F) present different expression patterns. HLA-E and HLA-F are usually ubiquitously expressed in low levels, and they interact with ligands in T and NK cells (such as HLA-E with NKG2A). HLA-G is predominantly expressed at the maternal-fetal interface and has primarily been associated with maternal-fetal tolerance by interacting CD8 from T cells and lLIRB1, lLIRB2, and KIR2DL4 from NK cells (Donadi et al., 2011).

The Class II region comprises four non-classical genes (HLA-DMA, -DMB, -DOA, -DOB), mostly related to peptide loading, and 17 classical HLA genes (e.g., HLA-DRA, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1, and others) that are expressed in Antigen Presenting Cells (APC) such as B cells, monocytes, macrophages, dendritic cells as well as epithelial cells under inflammatory signals (Rock et al., 2016). Unlike class I HLA molecules, class II molecules are heterodimers, consisting of α and β chains, encoded by the corresponding HLA genes (e.g., HLA-DPA1 and HLA-DPB1 produce the HLA-DR molecule), which facilitates molecular diversity. The HLA-DR beta chain can be encoded by nine different genes (DRB1-9) with complex patterns of expression, and gene content adding additional layers of complexity (Faner et al., 2009). Finally, the class III region, located between the class I and II regions, is the most gene-dense region of the MHC; this region contains genes encoding elements of the complement system, chaperone genes, cytokines such as TNF and LTA, but no HLA genes.

Finally, there are other important non-HLA genes in the MHC, such as TAP1 and TAP2, both related to peptide pumping from the cytoplasm to the endoplasmic reticulum (Praest et al., 2018), MICA and MICB, both induced in viral infections and tumors and activate NK-mediate killing (Ghadially et al., 2017), the tripartite motif (TRIM) family, related to cell cycle progression, autophagy, and viral replication restriction (van Tol et al., 2017), PSORS1C1, conferring susceptibility to psoriasis and systemic sclerosis (Allanore et al., 2011), and others.

In addition to their large number and potential for many combinations, the HLA genes display unparalleled genetic diversity, with more than 27,000 alleles and almost 17,000 unique proteins (June 03, 2021, https://www.ebi.ac.uk/ipd/imgt/hla/stats.html) identified for the five most polymorphic loci (HLA-A, -C, -B, -DRB and -DQB1). This diversity of HLA molecules is concentrated in the peptide-binding groove, which allows the presentation of peptides of various shapes and sizes, hence conferring broad protection against pathogens at the population level. At the same time, the polymorphic nature of HLA is also found on non-coding parts of the genes, such as the promoter and have an impact on expression (Kulkarni et al., 2011; Vince et al., 2016; Lima et al., 2019). Over evolutionary time, together with founder effects, multiple pathogen-challenges have exerted selective pressures on HLA alleles in human populations across the globe (Meyer and Thomson, 2001; Spurgin and Richardson, 2010), shaping allele frequency differences and selecting very specific or even private HLA alleles in some populations (Brandt et al., 2018; Meyer et al., 2018). The progress of genomics, and immunogenomics over the last decade, had deepened our understanding of HLA role in human diseases though the use of genome-wide association studies (GWAS) (Kennedy et al., 2017; Dendrou et al., 2018).

The COVID-19 HLA and Immunogenetics Consortium (CHIC) has been created during the pandemic to coordinate efforts on HLA analysis. The CHIC provided a website with information on HLA data and current projects (The COVID-19 HLA and Immunogenetics Consortium, 2020a). It is supported by a database (The COVID-19 HLA and Immunogenetics Consortium, 2020b) and its role is the centralization of relevant HLA and clinical data for COVID-19 study. It contains HLA data of 2,892 individuals from nine projects. These data are freely available and new data can be easily uploaded upon account creation. In addition, the website allows HLA allele frequencies visualization, and use of HLA data management and analysis tools. An HLA Imputation Portal (HIP) is set up to allow geneticists to infer individuals HLA alleles using SNP genotyping data, relying on multi-ethnic models from Zheng et al. (Zheng et al., 2014). This tool may help leverage SNP data to gain power in HLA association studies. The CHIC also produced a broad review on immunogenetic parameters (e.g., HLA, KIR, complement, cytokines and chemokines receptors) and their role in COVID-19 (Aguir et al., 2021). A more specific review of COVID-19 and HLA
Generating and working with HLA data

Performing immunogenetic studies can be a challenge for those unfamiliar with the specifics of HLA nomenclature. An individual HLA genotype can be obtained through multiple molecular techniques, the complexity of its nomenclature allows the alleles in a genotype to be described in different styles, and these data can be stored in a variety of file standards. Overall, HLA information can take multiple forms, requiring a comprehensive understanding of the nomenclature in order to run proper statistical analyses and find relevant associations.

Generating HLA Data

Originally, immunologists conducted microlymphocytotoxicity assays, testing patients T/B cells (for HLA class I) or B cells (for HLA class II) against different anti-sera or monoclonal antibodies in the presence of complement. Sera or antibodies recognizing the HLA antigens on cells would activate the complement and lyse the cell; this serology staining would reveal the patient HLA serotype (Park and Terasaki, 2000). Serology was however limited by the underlying complexity of HLA and it resulted in poor performances in transplantation (Hurley, 2021). The need to improve this performance and technique evolution, with the advent of PCR, conducted HLA specialists to switch to molecular typing. Molecular techniques were adopted for HLA typing; these methods allowed systematic identification of HLA alleles, based on sequence polymorphisms, providing a 'higher resolution' result that distinguishes many more allele categories than serological methods. This molecular typing consistently improved in resolution throughout the years driving nomenclature evolution along the way. Sequence-specific (PCR-SSO) methods rely on the hybridization of hundreds of labeled SSO probes targeting unique sequences in polymorphic regions. Sequence-specific priming (PCR-SSP) methods directly amplify elements of the HLA genes with PCR primers containing sequence-specific 3' end polymorphisms, resulting in less ambiguity (inability to distinguish alleles with similar nucleotide sequences), than SSO methods (Meral and Beksaç, 2007).

Sanger sequencing-based typing (PCR-SBT) methods initially provided sequences of the exons that encoded the peptide-binding groove, and later overlapping sets of sequences for entire genes. PCR-SBT was the gold standard for HLA genotyping until the development of next-generation sequencing (NGS) methods (Meral and Beksaç, 2007; De Santis et al., 2013). Except for the last one, PCR-SBT, previous methodologies were not suitable to detect new variants, and their goals were detecting known polymorphisms.

The application of NGS was explored in 2012, as part of the 16th International HLA and Immunogenetics Workshop (IHIW), but, given issues with mapping of short reads, allelic imbalance, phasing, and high costs, PCR-SBT remained the gold standard. More recently, the integration of NGS technologies with bioinformatic solutions for immunogenetics has improved the speed and accuracy of NGS HLA genotyping with lower error rates and fewer ambiguities than PCR-SBT (Baier et al., 2019; Jekarl et al., 2021), and the application of NGS was the focus of the 17th IHIW in 2017 (Vayntrub et al., 2020). Moreover, NGS is ideal to detect new HLA variants. Researchers now routinely identify novel HLA alleles (Nilsson et al., 2018; Ralazamahaleo et al., 2019; Loginova et al., 2020; Ananeva et al., 2021a; Ananeva et al., 2021b; Cheranev et al., 2021; Loginova et al., 2021) using NGS and confirm them using SBT with PCR-SBT error often responsible for non-concordance between the two. NGS-based sequencing of multiple exons and introns has led to increases in the growth of the IPD-IMGT-HLA Database collection (Robinson et al., 2015; Robinson et al., 2019). Unfortunately, the total number of new alleles may be underestimated as it is not uncommon for new alleles to be NGS-typed without Sanger validation.

So-called third-generation NGS generates unambiguous, phased HLA genotypes, using instruments like the PacBio SMRT (Mayor et al., 2015) or Oxford Nanopore Technology MinION (De Santis et al., 2020) to avoid multiple molecular techniques. This approach is faster than SBT and generates phased polymorphism with longer reads. Researchers using Oxford Nanopore Technology systems have successfully sequenced 11 HLA loci with low ambiguities in under 6 h (Mosbruger et al., 2020).

HLA Nomenclature

Soon after cell-surface antigens were identified as polymorphic between individuals, the WHO Nomenclature Committee for Factors of the HLA System was formed to develop a specific nomenclature for HLA genes, proteins and allelic variants (Allen et al., 1968). The original "HL-A" factor serologically
typed with multiple antibodies with an individual type (e.g., HLA (1,2/7,8) identifying them as positive for factors 1,2,7,8, and confirmed two distinct haplotypes from parental typing. As dozens of HLA genes and thousands of alleles were identified, the nomenclature was expanded to accommodate new complexity while building on the historical serological vocabulary. In 1987, the nomenclature was updated to accommodate newly available protein and nucleotide sequences (Antigens, 1987). The modern locus names were adopted at this time, and four-digit names were assigned to alleles, which were only defined as protein variants. In 2010, the current field-delimited nomenclature was adopted to account for the growing number of silent and non-coding nucleotide variants (Marsh et al., 2010).

A modern HLA allele name consists of up to four “fields”, each of which includes a two- or more digit number, each separated by a colon (Figure 1).

The first and second fields represent a historical serological group, and a unique protein sequence, respectively. All allele names have at least two fields. Alleles sharing the 1st, and 2nd fields with a different 3rd field encode the same protein but have unique silent-substitution in the exonic sequence, whereas sequence differences contained in the introns are written in the 4th field. The four fields of an allele name can also be suffixed with a single-letter “expression variant”, identifying alleles that are either not expressed, expressed at a low or questionable level, or secreted. For example, HLA-A*02:01:01 represents an exonic sequence shared by e.g., HLA-A*02:01:01:01 and HLA-A*02:01:01:134Q. In the latter case, the expression of HLA-A*02:01:01:134Q is Questionable, due to a potential alternate splicing nucleotide variant in intron 2. Allele names can be truncated to fewer fields for different applications, with each truncation described as a level of “resolution” (e.g., HLA-A*02 is a one-field resolution allele).

In addition to this allele nomenclature, specific groups of alleles have been defined. P and G groups refer to multiple alleles sharing either the same peptide or nucleotide sequence for the peptide-binding groove, respectively. For instance, HLA-A*02:01:01:134Q and HLA-A*02:252 both belong to the A*02:01P P group; the two proteins are globally different but share the same peptide-binding groove. HLA-A*02:01:01:134Q and HLA-A*02:89:01 belong to the A*02:01:01G G group as they share identical peptide-binding groove encoding exon sequences.

HLA supertypes are groups of alleles sharing similar peptide-binding repertoires. Supertypes are defined by “structural similarities, shared peptide-binding motifs, and identification of cross-reacting peptides” (Wang and Claesson, 2014). Using this classification, HLA-A*02:01:01:134Q potentially belongs with HLA-A*02:02, A*02:05, A*69:01 in the A2 supertype. (del
TABLE 1 | Tools for HLA analyses.

| HLA application name | Description | URL |
|----------------------|-------------|-----|
| Alphard-nt (Hayashi et al., 2019) | Identification of somatic mutations in HLA molecules from whole-genome and exome data using Bayesian algorithms | — |
| BIGDAWG (Pappas et al., 2016) | Open-source R package for the case-control analysis of highly-polymorphic data at the allelic, haplotype and amino-acid level | http://software.buhmhan.com/BIGDAWG |
| Easy-HLA (Geffard et al., 2020) | Website with HLA alleles haplotyping, upgrading and inference from HLA genotypes, prediction of HLA-C expression | http://hla.univ-nantes.fr/ |
| HATK (Choi et al., 2021) | Open-source Python pipeline for HLA association studies, including tools for HLA data formatting | https://github.com/WansonChoi/HATK |
| HLA-check (Jeanmougin et al., 2017) | Perl tool evaluating the probability of accurate HLA genotype imputation by comparing it to SNP imputation in the exonic region of HLA. | http://www.HLA-EMMA.com |
| HLA-EMMA (Kramer et al., 2020) | Donor/recipient compatibility assessment based on solvent-accessible amino acids, based on intraducos comparisons | http://www.HLA-EMMA.com |
| HAfX | Open-source R pipeline for HLA association studies. Performing SNP quality control steps, stratification, HLA imputation and representation of the results | — |
| HLAhapV (Osoegawa et al., 2018) | A Java-based HLA Haplotype Validator for quality assessments of HLA typing | https://github.com/mdp-bioinformatics/ImmunogeneticDataTools |
| HLA-NET (Nunes et al., 2014) | Set of tools to manipulate HLA data, infer haplotypes, convert files format, and information about typing | https://hla-net.eu/ |
| HLA-Pers (Aguiar et al., 2020) | Genotyping and quantification of HLA expression from RNA-seq data | https://github.com/genevol-usp/HLA-Pers |
| HLA-TAPAS (Luo et al., 2020) | Open-source Python pipeline for creation of reference panels and HLA association studies | https://CRAN.R-project.org/package=HLA-TAPAS |
| MergeReference (Cook and Han, 2017) | SNP2HLA compatible tool to concatenate multiple reference panels in order to gain accuracy during HLA imputation | http://software.buhmhan.com/MergeReference |
| pYHLA (Fan and Song, 2017) | Association analysis for HLA alleles in Python language | — |

Guercio et al., 1995). Some studies of the HLA molecules’ evolution have interpreted HLA diversity differently. Kaufman et al. (Kaufman, 2018; Di et al., 2021) have proposed promiscuous and generalist HLA categories when Di et al. have challenged the concepts of supertypes and function peptide-binding groove groups.

HLA Data Formats
The modern and legacy nomenclature systems are still in use, which often makes data comparison and meta-analysis difficult. In addition, HLA alleles are stored in multiple formats which impact their use with bioinformatic tools. TSV or CSVs have been used to store HLA genotypes, usually organizing individuals in rows and HLA genes in columns (with two columns for each gene). Such files are often generated manually, but are used by multiple population genetic and disease-association applications (Lancaster et al., 2007; Excoffier and Lischer, 2010; Pappas et al., 2016). More strictly-defined bioinformatic-oriented formats include HLA PED (or HPED) (Choi et al., 2021), an HLA-focused extension of the PED format (Purcell et al., 2007); Variant Call Format (VCF), as used by BEAGLE (Browning et al., 2018), in which HLA allele names are recoded as multiple binary identifiers, and Histoimmunogenetic Markup Language (HML), an XML format developed specifically for exchanging HLA and Killer-cell Immunoglobulin-like Receptor (KIR) genotype data (Milius et al., 2015). The IPD-IMGT/HLA Database releases new and updated reference sequences and allele names every 3 months. Individuals datasets may have been generated under any release version, which is why tools like the Allele Name Translation Tool (ANTT) have been developed to standardize datasets to a common release version. (Mack and Hollenbach, 2010). Development of a standardized means of storing and sharing data is still underway. In 2015, the MIRING reporting guideline (Mack et al., 2015) introduced standardized data elements and a controlled vocabulary for HLA genotype data and meta-data, which were implemented in HML (Milius et al., 2015). An HML message includes information on the IPD-IMGT/HLA Database version, the entity and how they generated the data, as well as references to external sources (e.g., reference sequences and aligned read). HML is used to transmit HLA genotyping data to the National Marrow Donor Program (and other similar registries and donor centers), but has yet to be adopted for genetic–analysis applications. Most of the existent HLA analysis applications require fewer data elements than are included in an HML message.

Given the number of different applications of HLA data, new informatics tools can influence the interpretation of this information. Multiple ancillary tools have been developed for HLA research. Whether they allow researchers to run rapid association analyses, extract new information from data, or link HLA genotypes to novel fields of translational research, all contribute to the advances in the HLA research (Table 1).

INFERRING AND IMPUTING HLA ALLELES: FROM COMPLEX READ-MAPPING TO THE STUDY OF LINKAGE DISEQUILIBRIUM
HLA inference is an umbrella term comprising multiple bioinformatic tools and statistical methods to obtain individuals’ HLA genotypes. Inference implies using missing information to obtain HLA genotypes, this can generally refer
Inference From Whole-Genome Sequencing and Whole-Exome Sequencing

Unlike NGS typing techniques which targets HLA genes (as many commercial kits apply), untargeted sequencing does not focus on HLA. Whole-genome sequencing (WGS) methods aim to identify all genetic variations of an individual genome, while whole-exome sequencing (WES) is designed to target all exons. Initially, these methods did not support the calling of HLA alleles; low coverage and short read-lengths led to poor HLA typing accuracy (Bauer et al., 2016). Low coverage does allow identification of HLA alleles, due to their high levels of polymorphism and extensive conserved sequences among genes, and improvements were needed (Hosomichi et al., 2015). Moreover, general pipelines for analyzing NGS data from WGS do not work for HLA genes; because they present high sequence similarity, it is very common that a short read (a sequence generated in NGS procedures) from one gene aligns to another gene (cross-mapping), leading to genotyping errors (e.g., HLA-A and HLA-H, or HLA-C and HLA-B) (Castelli et al., 2018).

To handle these issues, multiple algorithms have been developed and refined (Klasberg et al., 2019). These include: 1) classic read-mapping with HLA-specific quality control steps or different scores, hla-mapper (http://www.castelli-lab.net/apps/hla-mapper) (Castelli et al., 2018) which also works on KIR genes and provide genotyping and haplotyping at the SNP level, seq2HLA (Boegel et al., 2012) and HLAforest (Kim and Pourmand, 2013), among other tools; 2) population graph reference methods (e.g., HLA*PRGLA), which identify probability edges between polymorphisms nodes and project read data onto these to evaluate the most likely alleles.

Recent reviews and tool comparisons on the optimal methods for non-HLA targeted sequencing data are already available (see (Klasberg et al., 2019; Chen et al., 2021)). In 2020, Chen et al. found that HLA-HD was the most accurate tool for producing HLA genotypes from WGS and WES. However, the study focused on the performance of five tools only. Notably, most of the tools they studied achieved much higher accuracies than previously reported by Bauer et al., 2016, which emphasizes a drastic improvement in read coverage and processing in the MHC region (Bauer et al., 2016). Finally, researchers successfully implemented these tools in association studies, promoting their importance for HLA-centric epidemiological studies (Juhos et al., 2015; Xie et al., 2017; Mimori et al., 2019; Vince et al., 2020a).

HLA Allele Imputation

HLA genotyping data can also be generated using HLA imputation tools, which generate genotypes for individuals on the basis of LD between GWAS-derived SNP data for the MHC region and specific HLA alleles. These methods ultimately rely on reference datasets of HLA and SNP genotypes for the same individuals, and have become increasingly accurate in their predictions as new algorithms are developed.

Following the opportunity brought by SNP to SNP imputation, SNP to HLA imputation algorithms offered a quick and easy way to obtain HLA genotypes from widely available GWAS SNP genotyping data (McCarthy et al., 2016). SNP to HLA imputation relies on reference panels of individuals with known SNPs and HLA genotypes, to generate links between SNPs, haplotypes, and HLA alleles using machine learning algorithms (Figure 2).

The first published algorithms, SNP2HLA (Jia et al., 2013) and HLA*IMP (Ditlhey et al., 2011), were based on different implementations of hidden Markov models; SNP2HLA used BEAGLE (Browning and Browning, 2009), a haplotyping and SNP genotype imputation tool. In 2014, Zheng et al. proposed HIBAG, an attribute bagging method tailored for HLA data (Zheng et al., 2014), which showed better performance than pre-existing tools, and at the time was the only method to provide population-specific reference panels for hundreds of individuals while enabling construction of personalized reference panels building. Initial independent reviews suggested that SNP2HLA performed better on 3,265 samples from BioVU, a de-identified electronic health record database coupled to a DNA biorepository (Karnes et al., 2017). However, later reviews (Kuniholm et al., 2016; Pappas et al., 2018) and studies (Ritari et al., 2020) have favored HIBAG for HLA imputation, notably on more complex HLA data.

In practice, both SNP2HLA and HIBAG are commonly used to conduct HLA imputation or creation of new reference panels. Overall accuracy differences are low for European panels that had been extensively assessed. An important point still under investigation is the impact of population diversity in reference panels. While some researchers advocate for the creation of exhaustive multi-ethnic reference panels (Degenhardt et al., 2019), others have shown that specific populations (e.g., insular or admixed) require more restrained reference panels (Khor et al., 2015; Ritari et al., 2020).

The difficulty in determining if a reference panel is suitable for HLA imputation is related to how well it matches to target data, on the frequency of common alleles and the presence of rare HLA alleles, specific to some population (especially in underrepresented populations). This has led to the creation of reference panels with limited HLA diversity. While accuracy values are often reported as the ultimate answer to a model viability, these values can be misleading. For a rare HLA allele in a validation dataset, a 90% accuracy value can be achieved if that allele should be imputed 20 times out of 2,000 alleles (i.e., 1,000 individuals) but is never predicted. Therefore, other metrics (e.g., sensitivity, specificity, or F1 score (Cook et al., 2021)), must not be overlooked. Admixed populations are formed by individuals from different genetic backgrounds in
variable proportions, and HLA imputation can be sub-optimal if the reference panel is only drawn from one of the ancestral populations. Conversely, a reference panel from an admixed population with a different overall genome proportion from the individuals being imputed may also provide inaccurate results.

To effect worldwide improvement in HLA imputation efforts, we led the creation of an international consortium, the SNP-HLA Reference Consortium (SHLARC), whose aim is to gather data to represent the extreme diversity of HLA alleles, fostering accurate imputation (Vince et al., 2020b). We further advocate for improvements to current HLA imputation tools and for the development of a platform promoting easy access to HLA imputation for immunogeneticists. Though HLA imputation is not yet suited for clinical settings, generalization of HLA association studies offers a new way to investigate immune pathologies (Meyer and Nunes, 2017).

New versions HLA*IMP (Motyer et al., 2016) and SNP2HLA have been released (e.g., MHC*IMP (Squire et al., 2020), CookHLA (Cook et al., 2021), and Deep-HLA (Naito et al., 2021)) that apply new algorithms. These highlight the community intense interest in HLA imputation. CookHLA is an updated version of SNP2HLA (based on the BEAGLE algorithm) that better accounts for LD in the HLA region and makes use of the genetic map option to better impute individuals who are not well represented in the reference panels. For its part, Deep-HLA seems especially promising as deep learning may lead to better imputation of rare alleles.

**BIOINFORMATIC ANALYSES OF HLA INFORMATION**

The pressing challenge of understanding the COVID-19 pandemic, given previous associations with infectious diseases, has led researchers to scrutinize HLA using any available resource. In addition to issues of nomenclature and on-going technological evolution of typing methods, the complexity of HLA analyses is also derived from the multiple forms these analyses can take. On the one hand, HLA allele frequencies and predicted binding affinity of pathogen peptides to HLA alleles allow for a first step in the HLA world, as they are easily available, but are limited to investigate its actual role. On the other hand, the in-depth implication of HLA is revealed when looking at SNP association in the MHC region, and specifically when looking at allele associations, but their realization is hindered by high costs and technical difficulties. The study of HLA is multi-layered, with a continuum of methods peaking with analysis of individual data and multi-locus haplotypes, all of which contributing to a comprehensive understanding of the role of HLA in a given analysis.

**HLA Allele Frequencies**

The diversity of HLA alleles across geographically separated populations is thought to be the result of balancing selection due to local pathogens (Meyer and Thomson, 2001). The allelfrequencies.net database has the most extensive collection of HLA allele frequencies in diverse populations (Middleton et al.,
2003). In addition, HLA typing conducted by bone marrow registries may constitute a local estimation of HLA allele distribution in a population (Sacchi et al., 2019; Schmidt et al., 2020). It is possible to statistically analyze the correlation (e.g., via linear regression or Pearson coefficient) between a quantitative value, such as the number of COVID-19 cases, and the HLA allele frequencies obtained from a different sample in every studied population (e.g., in a database or registry).

However, while these correlations are faster and easier to obtain than new HLA genotypes, they may result in spurious correlations because: 1) most of the HLA alleles (and observed haplotypes) have a low frequency. For example, according to allelefrequencies.net, in the 416,581 individuals from the African-American NMDP population in the United States, two-thirds of the 321 HLA-B alleles at two-field resolution have frequencies below 0.003% (24 or less occurrences). Assuming that reference population samples are representative is not always accurate. A possible solution is to focus on common HLA alleles; 2) statistical tests are often applied without multiple-testing correction, regardless of the number of tests; 3) the confounding variables, both genetic (e.g., ancestry) and environmental (e.g., comorbidities), are often overlooked.

In any case, correlation is not causation. Therefore, the high number of HLA alleles and biased frequencies are bound to create spurious links between their presence and any phenotype. Therefore, to thoroughly investigate the relationship between HLA and phenotype, it is of the utmost importance to conduct studies and control for other genetic factors such as population stratification, linkage disequilibrium, or comorbidities (some linked to HLA polymorphism itself such as diabetes). Statistical bias could also be reduced by working on a higher number of samples and correcting for multiple testing. It is also worth considering different resolution levels of information, from in silico studies to full haplotype information.

**In silico Peptide Binding**

HLA molecules present endogenous and exogenous peptides, however, affinities for these peptides vary greatly depending on the peptide conformation and the peptide cleft topology and chemistry. Whether an HLA allele presents several or few peptides derived from one specific pathogen is one mechanism potentially explaining the strong immune response or tolerance towards it. Researchers can use prediction tools, such as NetMHCpan (Nielsen and Andreatta, 2016; Jurtz et al., 2017), trained on binding affinity and elution assays, to evaluate the number of potentially bound peptides for any HLA class I allele. The “pan” methods, contrary to the “allele-specific” methods, use similarities in sequence data to predict the peptide binding capacity of HLA alleles for which no information is available. Other tools exist and have been reviewed by Mei et al., in 2020 (Mei et al., 2019). Such predictions, coupled with HLA genotye data of individuals, give a theoretical insight into the possible adaptive immune response of a person. In these tools, the peptidome of the studied pathogen is informatically divided into peptide sequences of limited size (8–12 residues to account for the size of peptides presented by class I molecules), and the number of alleles predicted to bind a large number of peptides is inferred to represent better presentation to T cells, and a protective role against the pathogen. However, the only way to definitely determine peptide binding affinity is through laboratory experiments.

**Genome-wide Association Studies**

Genotyping data obtained with SNP arrays has proven to be fast and inexpensive for investigating the genetic component of complex traits and diseases (Claussnitzer et al., 2020), compared to more thorough and exhaustive sequencing technologies. Without assumptions regarding the region potentially involved in the studied trait, GWAS helped discover protective and risk alleles, particularly in the HLA region (Kennedy et al., 2017). Contrary to the use of independent HLA allele frequencies for studying a pathology, association studies assess the difference between affected individuals and unaffected individuals or the distribution of a particular quantitative trait. Both genetic and phenotypic data are individual and not population-based, reducing biases. The statistically significant SNP (aka, top hits) are linked to genes by proximity, and investigation by pathway analysis can reveal additional biological information on their effect. More recently, transcriptome-wide association studies have allowed more accurate investigation of the impact of a SNP on the expression of genes (Wainberg et al., 2019). In addition, some SNPs can be highly correlated to an HLA allele (e.g., rs2395029 and HLA-B*57:01 have been described multiple times as in complete linkage disequilibrium (de Bakker et al., 2006)), and therefore provide additional functional information for biological interpretation. Finally, statistical regression models can take into account potential confounding factors (e.g., genetic ancestry and population stratification, sex, age, comorbidities) to control for limiting biases.

Given the complex LD patterns across the MHC region, SNP association analyses are not usually precise enough to identify specific disease-associated HLA alleles. LD patterns may differ between populations. For example, the rs2395029 tags HLA-B*57:01 in Europeans but displays reduced LD in African-Americans (Colombo et al., 2008). The complex LD patterns and the high number of genes in the MHC region, make it difficult to pinpoint an SNP to a specific HLA allele in most cases.

**HLA Allele Association Studies**

Association studies of HLA alleles offer a more relevant biological explanation, based on peptide presentation. HLA allele data can come from different sources, including various epochs of HLA typing and HLA imputation from SNPs (see above). These data can be analyzed as is, or low resolution HLA data can be “upgraded” using the HLA-Upgrade tool from the Easy-HLA website, which statistically impute the most probable two-field genotype based on a haplotype database (Geffard et al., 2020). Once HLA data from multiple sources have been standardized for allele content and resolution, a frequency cut-off value is usually applied to test only those alleles with sufficient occurrences in the dataset to guarantee statistical power in the analysis. HLA alleles being highly polymorphic, they often display lower frequencies, and a larger sample size is
usually required to obtain significant results compared to SNP analyses.

Regression models, which are commonly used for SNP association, are the most versatile and common statistical models implemented to test associations between individual HLA alleles and phenotypes of interest (linear models for continuous and logistic for discrete phenotypes, respectively). Regression models can work with multiple covariables, allowing the disentanglement of the HLA effect and confounding factors such as population stratification, sex, gender, and others. Similar to GWAS SNP analyses, HLA alleles are tested individually as biallelic markers for each HLA gene, as each individual can exhibit 0, 1, or 2 occurrences of a given allele. As HLA molecules are expressed co-dominantly (Hughes and Nei, 1988), the dominant genetic model is commonly preferred to allelic or recessive models to assess HLA allele associations. However, it should be mentioned that different alleles might present different expression levels due to promoter and 3′UTR variations and final protein stability. Indeed, this is another HLA world: the effect of variants in the expression levels, which sometimes are directly linked with disease susceptibility (Kulkarni et al., 2011).

As in GWAS analysis, the overall performance of a statistical model can be evaluated with a Quantile-Quantile (QQ) plot, representing the observed \( p \)-value distribution for each HLA allele compared to the expected distribution under the null hypothesis. Any deviation from this distribution is highlighted by a deviation from a straight line. (Murdoch et al., 2008). Different scenarios can be described: 1) observed \( p \)-values mostly follow the null hypothesis, indicating that the statistical model accurately fits the data; 2) observed \( p \)-values deviate below the null hypothesis line, indicating that the statistical model is probably underpowered; 3) observed \( p \)-values deviate above the null hypothesis line, indicating that the statistical model may not be well parameterized and some confounding factors are not enough considered. Once the robustness of the analysis is confirmed, it is important to obtain a comprehensive visualization of the results with Manhattan plots, for instance, displaying \(-\log_{10}(p\text{-value})\) along with the list of test HLA alleles ordered numerically (as seen in Vince et al. (Vince et al., 2020a)). Volcano plots can also display the significativity of alleles along with their effect size, allowing a global view of their impact. Finally, the significance threshold accounting for multiple testing can be determined with the Bonferroni correction (5% a threshold divided by the number of tests) or other corrections such as the FDR, or permutations.

**Easy-HLA: Going Beyond HLA Alleles to HLA Genes Haplotypes, HLA Expression Levels, Specific HLA Amino Acids, KIR Ligand Groups**

New tools have been developed to facilitate the analysis of additional immunogenetic parameters (e.g. KIR ligands, see Figure 3).

HLA genotypes can be used to infer additional immunogenetic parameters that can further be analyzed (see Figure 3.) to get a clearer understanding of the relationship between immunity and pathologies. While one HLA allele already represents a haplotype of SNPs within a gene, as it is a collection of polymorphisms in the gene of interest, researchers have demonstrated the importance of looking at multiple HLA alleles on the same chromosome, which is referred to as an HLA haplotype. Association studies can be done on haplotypes, but many haplotype frequencies can be even lower than constituent allele frequencies. In a clinical setting, the collection of haplotype information is also useful, notably in HSCT transplants, for identifying haploidentical individuals. These haplotypes can be inferred using the HLA-2-Haplo tool from Easy-HLA website (Geffard et al., 2020), for instance. A straightforward, reliable, but expansive strategy to get HLA gene haplotypes is the analysis of trios (mother, father, and offspring) or third-generation long-read sequencing such as PacBio SMRT.

Easy-HLA also infers HLA-C expression levels, HLA alleles amino acids, and KIR ligand groups. Recently, high HLA-C
expression levels were associated with better control of HIV (Apps et al., 2013; Vince et al., 2016). Class I HLA alleles have also been grouped according to their dependence on tapasin, a major actor in peptide loading, which proved to be an interesting subdivision for studying HIV-1 control (Bashirova et al., 2020). Moreover, testing HLA allele amino acids may indicate a specific function of a given residue across several alleles, as with this study by McLaren et al., again in HIV control (McLaren et al., 2012). Finally, studying KIR ligand groups along with KIR typing as previously described (Martin and Carrington, 2013; Vince et al., 2014) can reveal the binding patterns of specific HLA alleles. For example, HLA-A and HLA-B molecules bearing the Bw4+ motif bind specifically to KIR3DL1. Similarly, HLA-C group 1 (C1) allele-encoded molecules carry an asparagine at position 80 and specifically bind KIR2DL2/3, as opposed to group 2 (C2) allele-encoded molecules, which carry a lysine and specifically bind KIR2DL1 (Parham et al., 2012). Grouping HLA alleles according to different functional parameters can increase the power of detecting a true positive signal and represent an opportunity to come closer to the biological cause behind HLA genetic association with diseases.

CONCLUSION

However intricate it may be, the MHC region, and HLA in particular, is the perfect candidate to investigate infectious or auto-immune diseases, as its primary biological role is to present antigen to the immune system. HLA research was able to grow in different directions from in silico studies on peptide binding to association studies of HLA alleles, giving leads on HLA involvement in pathologies. That said, HLA-focused analysis requires special care because its immense diversity and low-frequency distribution may potentially result in spurious associations when tested incorrectly or in a small cohort. Fortunately, many tools have been and are still developed to obtain high-quality HLA information for a low cost with statistical inference, through HLA inference from NGS data or HLA imputation from SNP GWAS data or HLA resolution upgrading from HLA genotypes. Researchers considering to explore HLA should take advantage of existing resources and mobilize them when taking on new challenges, such as with the SARS-CoV-2 research.

AUTHOR CONTRIBUTIONS

VD contributed in writing the review and produced figures. EC, SM, JH, P-AG, NV, and SL contributed in writing and editing various sections of the review.

FUNDING

NV has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 846520. This work is supported by the ATIP-Avenir Inserm program, the Region Pays de Loire ConnectTalent. This work was also supported by United States, National Institutes of Health (NIH) National Institute of Allergy and Infectious Disease (NIAID) grants R01AI128775 (JH, SJM), and R01AI158861 (JH).

REFERENCES

Aguiar, V. R. C., Augusto, D. G., Castelli, E. C., Hollenbach, J. A., Meyer, D., Nunes, K., et al. (2021). An Immunogenetic View of COVID-19. Genet. Mol. Biol. 44 (1), 1–24. doi:10.1590/1678-4685-gmb-2021-0036

Aguiar, V. R. C., Masotti, C., Camargo, A. A., and Meyer, D. (2020). ”HLApers: HLA Typing and Quantification of Expression with Personalized Index,” in Methods in Molecular Biology. Editor S. Boegel (New York, NY: Springer US), Vol. 2120, 101–112. doi:10.1007/978-1-0716-0327-7_7

Allanore, Y., Saad, M., Dieudé, P., Avouac, J., Distler, J. H. W., Amouyel, P., et al. (2021). An Immunogenetic View of COVID-19. Genet. Mol. Biol. 44 (4), 44–53. doi:10.1590/1678-4685-gmb-2021-0036

Ananeva, A., Leksina, Y., Andryushkina, A., and Shagimardanova, E. (2021). The influence of HLA-C Expression Level on HIV Control. Science (80–) 340 (6128), 87–91. doi:10.1126/science.1232685

Baier, D. M., Hofmann, J. A., Fischer, H., Rall, G., Stolze, J., Ruhner, K., et al. (2019). 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 Transpl. 54 (6), 928–930. doi:10.1038/s41409-018-0411-2

Bashirova, A. A., Viard, M., Naranbhai, V., Grifoni, A., Garcia-Beltran, W., Akdag, M., et al. (2020). HLA Tapasin independence: Broader Peptide Repertoire and HIV Control. Proc. Natl. Acad. Sci. USA 117 (45), 28232–28238. doi:10.1073/pnas.2013554117

Bauer, D. C., Zadoorian, A., Wilson, L. O. W., and Thorne, N. P. (2016). Evaluation of Computational Programs to Predict HLA Genotypes from Genomic Sequencing Data. Brief Bioinform. 19 (2), bbw097. doi:10.1093/bib/bbw097

Beck, S., Geraghty, D., Inoko, H., Rowen, L., Aguado, B., Bahram, S., et al. (1999). Complete Sequence and Gene Map of a Human Major Histocompatibility Complex. The MHC Sequencing Consortium. Nature 401 (6756), 921. doi:10.1038/44853

Boegel, S., Löwer, M., Schäfer, M., Bukur, T., de Graaf, J., Boisguérin, V., et al. (2012). HLA Typing from RNA-Seq Sequence Reads. Genome Med. 4 (12), 102. doi:10.1186/gm403

Brantd, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (2015). Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. G3 (Bethesda) 5 (5), 931–941. doi:10.1534/g3.j3.15015784
Parham, P., Norman, P. J., Abi-Rached, L., and Guethlein, L. A. (2012). Human-specific Evolution of Killer Cell Immunoglobulin-like Receptor Recognition of Major Histocompatibility Complex Class I Molecules. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367 (1590), 800. doi:10.1098/rstb.2011.0266

Park, I. and Terasaki, P. (2000). Origins of the First HLA Specificities. *Hum. Immunol.* 61 (3), 185–198. doi:10.1016/S0198-8859(99)00154-8

Præst, P., Luteijn, R. D., Brak-Boer, I. G. J., Lanfermeijer, J., Hoelen, H., Ijgosse, L., et al. (2018). The Influence of TAP1 and TAP2 Gene Polymorphisms on TAP Function and its Inhibition by Viral Immune Evasion Proteins. *Mol. Immunol.* 101 (May), 55–64. doi:10.1016/j.molimm.2018.05.025

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PHILNK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Razafimahaleo, M., Andreani, M., Giustiniani, P., Guidicelli, G., and Visentin, J. (2019). Characterization of the Novel HLA-DQA1*01:01:05 Allele by Sequencing-based Typing. *Hla* 94 (2), 172–173. doi:10.1016/j.humimm.2018.11.003

Ritari, J., Hyvärinen, K., Clancy, J., Partanen, J., and Koskela, S. (2020). Increasing Accuracy of HLA Imputation by a Population-specific Reference Panel in a FinnGen Biobank Cohort. *NAR Genomics Bioinforma* 2 (2), 1–9. doi:10.1093/nargab/leaa030

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., and Marsh, S. G. E. (2019). IPD-IMGT/HLA Database. *Nucleic Acids Res.* 48 (D1), D948–D955. doi:10.1093/nar/gkx950

Robinson, J., Hallwell, I. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. (2015). The IPD and IMGT/HLA Database: Allele Variant Databases. *Nucleic Acids Res.* 43, D423–D431. doi:10.1093/nar/gku161

Rock, K. L., Reits, E., and Neefjes, J. (2016). Present Yourself! by MHC Class I and Antigen (HLA) Supertypes, with xHLA. *Cell* 167 (1), 247–259. Available at: http://www.xhla.org/

Sidney, J., del Guercio, M. F., Southwood, S., Engelhard, V. H., Appella, E., Rammensee, H. G., et al. (1995). Several HLA Alleles Share Overlapping Peptide Specificities. *J. Immunol.* 154 (1), 247–259. Available at: http://www.ncbi.nlm.nih.gov/pubmed/75327812.

Spurgin, L. G., and Richardson, D. S. (2010). How Pathogens Drive Genetic Diversity: MHC, Mechanisms and Misunderstandings. *Proc. Biol. Sci.* 277, 979–988. doi:10.1098/rspb.2009.2084

Squire, D. M., Motyer, A., Ahn, R., Nititham, J., Huang, Z.-M., Oksenberg, J. R., et al. (2020). MHC-IMP - Imputation of Alleles for Genes in the Major Histocompatibility Complex. *bioRxiv*. doi:10.1101/2020.01.24.919191

The COVID-19 HLA and Immunogenetics Consortium (2020a). HLA|COVID-19. Available at: http://www.hla covid19.org/

The COVID-19 HLA and Immunogenetics Consortium (2020b). HLA|COVID-19 Database. Available at: https://database.hla covid19.org/

Trowsdale, J., and Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annu. Rev. Genom. Hum. Genet.* 14 (1), 301–323. doi:10.1146/annurev-genom-091212-153455

Trowsdale, J., and Moffett, A. (2008). NK Receptor Interactions with MHC Class I Molecules in Pregnancy. *Semin. Immunol.* 20 (6), 317–320. doi:10.1016/j. semimunol.2008.06.002

van Tol, S., Hage, A., Giraldo, M., Bharaj, P., and Rajasbaum, R. (2017). The TRIMeDous Role of TRIMs in Virus-Host Interactions. *Vaccines* 5 (3), 23. doi:10.3390/vaccines5030023

Vinay, T. A., Mack, S. J., and Fernandez-Via, M. A. (2020). Preface: 17th International HLA and Immunogenetics Workshop. *Hum. Immunol.* 81, 52–58. doi:10.1016/j.humimm.2020.01.008

Vince, N., Bashiroida, A. A., Lied, A., Gao, X., Dorrell, L., McLaren, P. J., et al. (2014). HLA Class I and KIR Genes Do Not Protect against HIV Type 1 Infection in Highly Exposed Uninfected Individuals with Hemophilia A. *J. Infect. Dis.* 210 (7), 1047–1051. doi:10.1093/infdis/jiu214

Vince, N., Douillard, V., Geffard, E., Meyer, D., Castelli, E. C., Mack, S. J., et al. (2020). SNP-HLA Reference Consortium (SHLARC): HLA and SNP Data Sharing for Promoting MHC-centric Analyses in Genomics. *Genet. Epidemiol.* 44 (7), 733–740. doi:10.1002/gepi.22334

Wang, M., and Claesson, M. H. (2014). "Classification of Human Leukocyte Antigen (HLA) Supertypes," in Methods in Molecular Biology (Clifton, NJ), 309–317. doi:10.1007/978-1-4939-1115-8_17

Xie, C., Yeo, Z. X., Wong, M., Piper, J., Long, T., Kirkness, E. F., et al. (2017). Fast and Accurate HLA Typing from Short-Read Next-Generation Sequence Data with xHLA. *Proc. Natl. Acad. Sci. USA* 114 (30), 8059–8064. doi:10.1073/pnas.1707945114

Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., et al. (2014). HIBAG-HLA Genotype Imputation with Attribute Tagging. *Pharmacogenomics* 14 (12), 192–200. doi:10.1080/13214714.2013.18

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Douillard, Castelli, Mack, Hollenbach, Gourraud, Vince and Limou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.