A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations

Sanchez-Mazas Alicia
Department of Genetics and Evolution, Anthropology Unit, University of Geneva, Switzerland

Summary

Human leucocyte antigen (HLA) alleles and single nucleotide polymorphisms (SNPs) lying in the HLA region are known to be associated with several infectious diseases among which acquired immunodeficiency syndrome, hepatitis B, hepatitis C, tuberculosis, leprosy and malaria are highly prevalent in many human populations worldwide. Distinct approaches such as case-control comparisons, immunogenetic analyses, bioinformatic peptide-binding predictions, ancient DNA and genome-wide association studies (GWAS) have contributed to improving this knowledge during the last decade, although many results still need stronger statistical and/or functional support. The present review updates the information regarding the main HLA allele and SNP associations observed to date for six of the most widespread and some other infectious diseases, and provides a synthetic illustration of these findings on a schematic HLA genomic map. It then discusses these results by stressing the importance of integrating information on HLA population diversity in disease-association studies.

Keywords: human leucocyte antigens, HLA, MHC, infectious diseases, AIDS, hepatitis B, hepatitis C, tuberculosis, leprosy, malaria, HIV, HBV, HCV, Mycobacterium tuberculosis, Mycobacterium leprae, Plasmodium falciparum, Plasmodium vivax, population genetics, GWAS, peptide binding predictions, ancient DNA

Introduction

Human leucocyte antigen (HLA) molecules are encoded by a set of 21 protein-coding loci lying amongst many other genes and pseudogenes at region 6p21 of our genome [1, 2]. The HLA transmembrane proteins encoded by the classical (A, B, C, DR, DQ and DP) HLA genes are principally involved in the presentation, at the cell surface, of small pathogen-derived peptides to T cells, which triggers an immune response [3]. HLA class I (A, B, C) molecules are composed of a single α chain non-covalently bound to a small β2-microglobulin polypeptide encoded by another chromosome (15q21). Their α1 and α2 domains form the peptide-binding site where small peptides of 8 to 10 amino acids derived from viral proteins produced by an infected cell are presented to cytotoxic CD8+ lymphocytes, with the effect of destroying that cell. HLA class II (DR, DQ, DP) molecules are heterodimers composed of one α (encoded by DRA, DQA and DPA genes) and one β (encoded by DRB, DQB and DPB genes) chain, the α1 and β1 domains of which form the peptide binding site. These proteins present longer peptides of 13 to 25 amino acids derived from endocytosed antigens to helper CD4+ lymphocytes. In this case, the recognition of the HLA-peptide complex by the lymphocytes, via their T-cell receptors, induces the release of cytokines that will orchestrate a tailored immune response against the pathogen, for instance by helping B cells to secrete high affinity antibodies or by inducing macrophage activation.

All classical HLA genes except DRA and DPA are highly polymorphic at their peptide binding site—encoding exons (exons 2 and 3 for class I and exon 2 for class II genes), which translates into several thousands of currently known HLA alleles and high levels of heterozygosity (often around 90%) in most human populations [1, 4, 5]. This huge diversity is usually explained by the advantage it may confer to heterozygous individuals, which would be capable of presenting a larger variety of peptides than in homozygous individuals, therefore enhancing their T-cell repertoire and protection against pathogens. This heterozygous advantage model, first described by Doherty and Zinkernagel [6], was further refined by the consideration that heterozygotes carrying functionally divergent alleles would have a higher fitness (ability to survive and reproduce) than heterozygotes carrying more similar alleles, hence the naming of divergent allele advantage [7–9]–, also recently extended, under the name of joint divergent asymmetric selection, to the simultaneous action of multiple HLA loci [10].

The extreme genetic variation observed at classical HLA genes in human populations is thus principally considered as the result of an adaptation to pathogen-rich environments during evolution, although other mechanisms have been proposed as well [11]. This population heterozygote advantage is a global effect of HLA allelic variation that does not necessarily indicate allele-specific overdomi-
nance whereby a heterozygote would be more protected than the homozygotes for the same allele [12]. On the other hand, protective or harmful effects of specific HLA alleles or single nucleotide polymorphisms (SNPs) in the HLA region have been suggested through several epidemiological approaches, a number of them relating to infectious diseases, as documented below. Such diseases are characterised by the fact that they can be passed directly or indirectly from one individual to another through several categories of infectious agents (viruses, bacteria, parasites and fungi), with the dramatic consequences that they may quickly affect and kill hundreds of thousands or even millions of individuals each year (table 1). This is why there is an urgent need to understand better how each individual reacts against different kinds of infections, taking into account their HLA genotype. From an evolutionary genetics perspective, a fascinating issue is also to understand how specific HLA markers evolved in relation to the spread of infectious diseases worldwide. For example, in the case of a severe epidemic leading to a reduction in size of a given population (i.e., a population bottleneck), we may expect highly protective alleles to be maintained in the population through positive selection and susceptibility alleles to be eliminated through purifying selection or selective sweep, as revealed in the major histocompatibility complex (MHC) genes of chimpanzees following putative simian immunodeficiency virus epidemics [13–15]. This partly explains why genetic differences are observed at HLA genes between populations (if the latter underwent distinct disease outbreaks during evolution) despite a general effect of population heterozygosity advantage [16]. In the present paper, we first review the associations that have been best documented between markers at classical HLA genes and six of the most prevalent infectious diseases in human populations, and we then discuss these findings in a population genetics and evolutionary perspective.

### HLA allele and SNP associations with infectious diseases

Causal relationships between host genetic variation and infectious disease susceptibility have been found at different genes across the genome, one of the best documented examples being CCR5 Δ32, which impairs the penetration of human immunodeficiency virus (HIV) into the cells [17]. In addition, although precise causal determinants have not been identified, for about half a century HLA genes have been reported to be associated with diseases including many different autoimmune disorders and infections, as well as several kinds of cancers (reviewed in [2, 18–25]). Genome-wide association studies (GWAS), which explore putative associations between genetic markers and particular phenotypes across the whole genome, have identified significant signals of positive selection in that region (e.g., [26]) and a large number of SNPs within the HLA region have been found to be associated with viral infections, as recently reported in several reviews [27–29]. To date, HLA associations have been more particularly documented for acquired immune deficiency syndrome (AIDS), hepatitis B, hepatitis C, tuberculosis, leprosy and malaria, as described below (note that regarding HLA alleles, we focused this review on alleles defined at the second-field level of resolution that corresponds to specific HLA proteins [1] except when presenting the historical backgrounds of some findings).

### Acquired immune deficiency syndrome

One of the best documented associations of markers in the HLA region with viral infections is for the acquired immune deficiency syndrome (AIDS) due to infection by the human immunodeficiency virus (HIV) [30], which is one of the major global public health issues today (table 1). Earliest studies suggested that AIDS progression was accelerated in homozygotes at HLA class I genes compared with heterozygotes [31, 32] and in HLA-B*35 [32, 33] and/or HLA-C*04 carriers [32]. B*35 subtypes differing in their peptide-binding specificity at position 9 of the peptide binding site were also considered, and it was shown that changes at a single amino acid position of the HLA molecules could lead to differences in terms of HIV disease progression [34]. A significant negative correlation was also observed between the rapidity of AIDS progression and the number of viral epitopes recognised by HLA class I types, suggesting a direct relationship between HLA alleles and the progression of the disease [35]. More recently, Pereyra et al. [36] performed GWAS in a multi-ethnic cohort of HIV-1 controllers (infected people who show ability to control HIV replication without any therapy and who do not develop clinical disease) and progressors (used as controls): in the largest European group, they identified 313 SNPs reaching genome-wide significance in HLA and none in other genomic regions. Four independent markers lie in proximity to HLA class I, among which are rs4418214 close to MICA and rs2395029 related to the HLA Complex P5 gene (HCP5) coding for a long non coding regulatory RNA associated with several diseases [37]. The G variant of rs2395029 is itself in linkage disequilibrium with allele HLA-B*57:01 [38–40] protecting against HIV progression to AIDS in populations of European ancestry [41], although also found to confer hypersensitivity to abacavir, an antiretroviral drug used to treat HIV-infected patients [42–45]. However, as reported by Kulski et

---

**Table 1**: Six of the most prevalent infectious diseases according to the World Health Organization.

<table>
<thead>
<tr>
<th>Category</th>
<th>Disease</th>
<th>Infecting agent</th>
<th>Number of cases worldwide*</th>
<th>Number of deaths worldwide*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viral</td>
<td>Acquired immunodeficiency syndrome</td>
<td>Human immunodeficiency virus</td>
<td>37.9 million in 2018</td>
<td>770,000 in 2018</td>
</tr>
<tr>
<td>Viral</td>
<td>Hepatitis B</td>
<td>Hepatitis B virus</td>
<td>257 million in 2015</td>
<td>887,000 in 2015</td>
</tr>
<tr>
<td>Viral</td>
<td>Hepatitis C</td>
<td>Hepatitis C virus</td>
<td>71 million in 2015</td>
<td>399,000 in 2016</td>
</tr>
<tr>
<td>Bacterial</td>
<td>Tuberculosis</td>
<td>Mycobacterium tuberculosis</td>
<td>10 million in 2018</td>
<td>1.5 million in 2018</td>
</tr>
<tr>
<td>Bacterial</td>
<td>Leprosy</td>
<td>Mycobacterium leprae</td>
<td>&gt;200,000 new cases in 2018</td>
<td>–</td>
</tr>
<tr>
<td>Parasitic</td>
<td>Malaria</td>
<td>Plasmodium falciparum and vivax</td>
<td>228 million in 2018</td>
<td>405,000 in 2018</td>
</tr>
</tbody>
</table>

* Reported in the last years

---
al. [37], the association of HCP5-rs2395029 with HIV viral load set point has not been confirmed in African populations, where HLA-B*57:03 is found in place of B*57:01. Actually, the closely related HLA-B*57:02, HLA-B*57:03 and HLA-B*58:01 alleles are all found to be protective against HIV disease progression by presenting several p24 Gag-specific epitopes [46], although Env protein-derived epitopes also likely play an important role in HLA-HIV restriction involving HLA-B*57:01 and HLA-B*58:01 [47]. As reviewed by Bardeskar and Mania-Pramanik [48], another SNP located 35 kb upstream of HLA-C, rs9264942 (termed –35Kb), was putatively associated with both HLA-C expression levels and HIV control (variant C being associated with higher expression and slower disease progression) [39, 49], the causal variant being however an in/del located at position 263 of HLA-C 3′-UTR (rs67384697, in linkage disequilibrium with rs9264942) regulating the binding of a micro-RNA (miR-148a) [50–55]. In the African American group, Pereyra et al. [36] identified 33 significant SNPs (none in common with Europeans) among which r2523590 2kb upstream of HLA-B in common with the (so-called by the authors) Hispanic group, as well as the intronic rs2523608, also observed by Pelak et al. [56] to be associated with HLA-B*57:03, which was previously suggested to be involved in slower HIV-1 disease progression in populations of African descent [57]. Pereyra et al. [36] also identified (through amino acid imputation) six significant amino acid positions among which five within the PBR, more particularly position 97 (C-pocket) in HLA-B that would mediate the protection or susceptibility to HIV progression. Finally, besides HLA-B*27:05 [36, 38], HLA-B*27:02 was recently identified as another protective factor in HIV-1 (slow disease progression) [58], which confirms the implication of the formerly proposed HLA-B*27 allele [59]. More recently, bioinformatic peptide-binding predictions have again stressed the putative advantage of HLA heterozygotes in HIV-1 control [60].

**Hepatitis B**

Viral hepatitis includes a set a diseases due to different viruses (A, B, C, D, E) [61], among which HBV and HCV cause 96% of hepatitis mortality in humans [62] (table 1). Former association studies focusing on the HLA region (partly reviewed in [63–65]) suggested that homozygosity, as well as specific alleles at HLA class II loci (DQA1*05:01, DQB1*03:01, DPB1*09:01, DRB1*11:02), increased the risk of HBV infection or progression [66–68], whereas both HLA class I (A*03:01 and class II (DRB1*13:01, DRB1*13:02, DQB1*02:01, DPB1*02:01) alleles were associated with protection or viral clearance [68–72]. More recently, GWAS studies identified HLA-DP as the locus most strongly associated with chronic HBV infection [73, 74]. In the 3′-UTR region of both DPA1 and DPB1, 11 SNPs were found to be associated with chronic hepatitis B in Asian populations, the strongest associations being found for rs3077 and rs9277535, respectively (A alleles conferring protection in both cases) [73]. Based on these results, two protective (DPA1*01:03–DPB1*04:02 and DPA1*01:03–DPB1*04:01) and two risk (DPA1*02:02–DPB1*05:01 and DPA1*02:02–DPB1*03:01) HLA haplotypes were proposed. Further fine-mapping of the MHC region in East Chinese HBV carriers and controls also revealed a strong association with chronic hepatitis B infection of HLA-DPβ1 amino acid positions 84-87 lying in the PBR, themselves in high linkage disequilibrium with SNP rs9277535 at HLA-DPB1 and medium linkage disequilibrium with SNP rs3077 at HLA-DPA1, making these positions possible causal variants involved in peptide presentation [75]. However, decreased expression of HLA-DPA1 and HLA-DPB1 has also been associated with rs3077 and rs9277535, respectively, suggesting that low gene expression may be a risk factor [76], in line with the more general idea that associations between HLA markers and diseases are partly driven by differential expression of HLA alleles [77]. A different SNP located in the 3′-UTR of HLA-DPB1 – rs9277534 (or 496A/G), in complete linkage disequilibrium with rs9277530, rs9277531, rs9277533 and rs9277536 – was found to be associated with either recovery (variant A, in linkage disequilibrium with allele HLA-DPB1*04:01) or persistence (variant G, in linkage disequilibrium with allele HLA-DPB1*01:01) of HBV in both Europeans and African-Americans [78]. Here again, the associations would be due to differences in HLA-DPB1 expression levels rather than in peptide binding properties, but in this case higher levels of HLA-DP expression (496 GG genotype) would confer HBV persistence [78]. A novel SNP near HLA-DPA3 (rs9366816) was also found to be significantly associated with persistent HBV infection in Han-Taiwanese [79] and a SNP near HLA-DPB1 (rs9277542) with protection against chronic HBV and viral clearance in Japanese and Koreans [80] as well as in Argentinians [81]. Mbarek et al. [74] also showed that two SNPs located within the HLA-DQ locus, namely rs2856718 located in the intergenic region between HLA-DQA2 and HLA-DQB1, and rs7453920 located in intron 1 of HLA-DQB2, exerted effects on hepatitis B susceptibility (A conferring protection and G susceptibility in both cases) independently from the effects of the SNPs found at locus HLA-DP in the studied Japanese cohort. Based on this study, HLA haplotypes DQA1*01:02–DQB1*03:03 and DQA1*03:01–DQB1*06:01 were associated with risk and DQA1*01:01–DQB1*05:01 and DQA1*01:02–DQB1*06:04 to protective effects. Two significant SNPs were also identified near the HLA-DQA2 (rs9276370) and HLA-DQB2 (rs7756516) loci [79] and one SNP (rs9268652) in HLA-DRA [29]. Whereas the involvement of HLA-DP SNPs rs3077, rs9277534 and rs9277535 and HLA-DQ SNPs rs2856718 and rs7453920 were confirmed in other studies (e.g., [63, 79–88]), although sometimes contradicted in cohorts of different origins (e.g., [89]), significant associations of SNPs in the HLA-DP/HLA-DQ regions were further expanded by Tao et al. [90], the most significant ones among a total of 76 being again found in HLA-DPB1 3′-UTR. Together with a large meta-analysis confirming the significant associations of SNPs rs3077 and rs9277535 [91], this genomic region thus appears to play a leading functional role in HBV immunity. Some studies also observed significant associations between HBV and SNPs in the HLA class I region, namely rs3130542 near HLA-C (the A allele being associated with a low HLA-C expression and increased risk of
developing chronic infection) [85], further found to be in linkage disequilibrium with HLA*07:02 [88], as well as rs2853953, in linkage disequilibrium with HLA-C*06:02 [88].

**Hepatitis C**

Some studies also suggested associations between HLA alleles and hepatitis C virus (HCV) persistence, such as HLA-DRB1*03:01, or spontaneous clearance, such as HLA-A*02:01, A*11:01, B*57:01, B*57:03, C*01:02, DQB1*03:01, DRB1*01:01, DRB1*04:01 and DRB1*11:01, among others [92–99]. However, these associations depend on population origin, as stressed in a recent review and meta-analysis of HCV clearance [100] that reported highly significant associations for alleles HLA-DQB1*02:01, DQB1*03:01, DRB1*07:01, DRB1*11:01 and/or DRB1*12:01, depending on the ethnicity. SNP rs4273729 (C allele) associated with HLA-DQB1*03:01 (which would be expressed at higher levels compared to other DQB alleles) was proposed to be involved in spontaneous resolution of HCV infection in individuals of both European and African ancestry [101]. SNP rs9275572, located in the HLA-DQB1 region, showed a significant association with chronic HCV in Japanese, however giving the same signal as SNP rs1130380, a variant specific to HLA-DQB1*03 that causes an amino acid substitution in the peptide-binding pocket of HLA-DQB1 (position 55) [102]. Note also that alleles HLA-B*57:01 and B*57:03, associated with HCV clearance [92, 95] have also been associated with slow HIV disease progression [41, 57], that SNP rs3077 of the HLA-DPA1 region, mentioned above as associated with HBV infection, has been suggested to increase the risk of chronic HCV infection in a Chinese population [103] and that SNP rs9264942 putatively linked to HIV control (see above) has been found to be associated with persistent HCV seronegativity among C/C genotyped people who inject drugs [104].

Other associations within the HLA region with HBV or HCV progression towards different diseases have recently been documented by Sawai et al. [105] and Omae and Tokunaga [106].

**Tuberculosis**

Tuberculosis is due to infection by the bacteria *Mycobacterium tuberculosis*. Although curable, it is still the major cause of death from infectious diseases on the global scale (table 1). A number of studies have reported significant associations between HLA alleles and either protection or susceptibility to several respiratory diseases [107], including tuberculosis (review and meta-analysis in [108]), but with many inconsistencies [107, 109–111]. GWAS studies have revealed several significant markers in the HLA class II region: according to Sveinbjornsson et al. [112], two markers, rs557011 [T] located between HLA-DQA1 and HLA–DRB1 and rs9272785 located in exon 4 of HLA-DQA1 (a missense mutation p.Ala210Thr which defines HLA-DQA1*03 alleles) would confer susceptibility to *M. tuberculosis* infection in European populations, the former SNP being also a risk factor for the development of the disease in infected individuals (note that about one third of the global population is latently infected by the bacteria, some 10% of them developing the disease); by contrast, rs9271378 [G], also located between HLA-DQA1 and HLA–DRB1, would protect against the development of the disease. The same authors confirmed the susceptibility of HLA-DQA1*03:01 to tuberculosis after imputation of HLA alleles. They suggest that HLA-DQA1*03 molecules would have a reduced stability and would be less able to present *M. tuberculosis* antigens. Other missense mutations classified as probably damaging were found in the Han Chinese population, namely rs41553512 and rs1136744 in HLA-DRB5 and rs41553512 in HLA-DQB1, in addition to a significant association of TB with HLA-DQB1*02:01 [113]. Because six different *M. tuberculosis* strains are unevenly distributed geographically [114], putative associations between HLA polymorphisms and tuberculosis caused by specific strains were investigated, which revealed higher susceptibility conferred by alleles HLA-DRB1*09:01 and HLA-DQB1*03:03 (both in linkage disequilibrium) to modern strains in Southeast Asia [115]. The association of HLA-DQB1*03 with tuberculosis was supported by a recent study [116] that also suggested a protective effect of HLA-B*58:01 through *M. tuberculosis* antigens epitopes CFP-10 (ESAT-6-like protein) binding and a susceptibility effect of HLA-B*58:02, which differs by 3 amino acids from the former.

**Leprosy**

Leprosy is due to infection by the bacteria *Mycobacterium leprae*. Although the disease is decreasing at the global level, a high number of new cases are still reported per year (table 1). India, Brazil, Indonesia, the western Pacific and some parts of Africa being the most affected. As reviewed by Jarduli et al. [117], many studies identified both protective and risk class I and class II HLA alleles associated with leprosy. A case-control study in India revealed two SNPs, one in the HLA-DRB1 region (rs9270650) and one within the HLA-DQA1 gene (rs1071630), as major susceptibility markers [118]. Based on genome-wide studies, HLA-DRB1 further appeared as a key region for leprosy susceptibility, with SNPs rs9271366, rs602875 [119], rs9271011 and rs9271100 [120] identified in Chinese Han, the two latter being in strong linkage disequilibrium with HLA-DRB1*15:01, another significant variant being HLA-DQB1*04:01. SNP rs9271011 showed a regulatory effect, the risk allele [T] being responsible of an increased HLA-DRB1 expression. By analysing both ancient and modern DNA, an associated SNP, rs3135388 [T], was found to be significantly more frequent in a sample of 69 M. leprae-positive individuals showing lesions specific for lepromatous leprosy in a medieval population from Denmark than in both ancient and modern randomly sampled individuals from Denmark and north Germany [121]. Based on both direct identification through high-throughput sequencing of HLA-DRB1*15:01 and HLA-DQB1*06:02 alleles in a substantial number of medieval individuals with lepromatous leprosy and a limited *M. leprae* peptide-binding capacity predicted for the HLA-DRB1*15:01 allele, the same study suggested that HLA-DRB1*15:01–DQB1*06:02 was a strong risk haplotype for this disease. HLA class I alleles may also be implicated in leprosy pathogenesis, as suggested by the identification of two class I intergenic SNPs - rs2394885 and rs2922997 - associated with leprosy susceptibility in Vietnamese and
Indians, the former one being in strong linkage disequilibrium with HLA-C*15:05 [122].

Malaria

According to the 2019 report of the World Health Organization [123], 228 million cases of malaria were reported worldwide in 2018 (table 1), among which 213 million (93%) were in Africa, followed by South-East Asia (3.4%). *Plasmodium falciparum* is the most virulent species and the most widespread in Africa (99.7% of cases), South-East Asia (62.8%), the Eastern Mediterranean (69%) and the Western Pacific (71.9%), whereas *Plasmodium vivax* is the predominant parasite in the Americas (74.1%). Several genes across the genome have been shown to display polymorphisms conferring protection against malaria, namely markers related to haemoglobinopathies (sickle-cell trait HbS, HbE, Hbc, α- and β-thalassemias), erythrocyte antigens (ovalocytosis SLC4A1, Duffy DARC), enzymopathies (glucose-6-phosphate dehydrogenase [G6PD], pyruvate kinase [PKLR]) and immunogenetic variants (complement receptor-1 CR1, HLA) (reviewed in [124]). Besides the putative role of HLA-G variation in the regulation of the response to *P. falciparum* infection [125, 126], several alleles of classical HLA genes have also been suggested to be protective to this form of malaria. The best known of them is HLA-B*53, first identified, in addition to the HLA class II DRB1*13:02–DQB1*05:01 haplotype, in a case-control study in Gambia [127]. HLA-B*53 was shown to recognise a conserved nonamer peptide from liver-stage-specific antigen-1 (LSA-1) of *P. falciparum* [128]. Based on their frequency in populations, several HLA alleles were also proposed as resistance (B*53:01, DQB1*05:01, DRB1*01:01 and DRB1*13:02) or susceptibility (A*30:01, A*33:01, DPB1*17:01 and DRB1*04:01) factors [129, 130]. The protective effect of HLA-B*53:01 was further supported by a population genetics modelling showing a clear relationship, after controlling for other genetic, geographic and environmental variables, between the frequency of this allele and the prevalence of *P. falciparum* in Africa [131]. The same study also identified HLA-B*78:01 as another possible candidate. Interestingly, based on peptide binding predictions both HLA-B*53:01 and B*78:01 share a high affinity for peptides with amino acid proline at position 2 (accommodated by pocket B of the peptide binding site), a property that is shared by 50% of the most frequent (frequency above 15% in at least one population) HLA-B alleles found in Africa [131]. Among these alleles, HLA-B*35:01 also revealed a significant protective effect against malaria in Ghana [132], in agreement with a previous finding that two antigenic octamer peptides of *P. falciparum’s* circumsporozoite protein, cp26 and cp29, bind HLAB*35 [133]. Despite the common idea that HLA class II, rather than HLA class I, molecules operate at the peptide-binding level in the case of parasitic infections (although cross-presentation of exogenous antigens on HLA class I is a well described phenomenon in professional antigen presenting cells), these findings suggest that several HLA-B alleles from the standing genetic variation increased in frequency when malaria expanded in Africa, probably with the development of agriculture [134]. This mechanism of soft selective sweep, which is likely to be a common evolutionary process in the human genome [135, 136], does not exclude, however, that pathogen-mediated hard selective sweep also operates at HLA genes, as recently suggested by the very high HLA class II allele and haplotype frequencies, presumably not resulting from rapid genetic drift, observed in a large West African population [137]. Interaction between HLA-B*53:01, carrier of the Bw4 epitope, and killer cell immunoglobulin-like receptor KIR3DL1 has also been suggested to contribute to its protective effect [138]. Genes of the HLA region would also be associated with *P. vivax* infection [139], although, to our knowledge, no specific HLA allele defined at the second field level of resolution nor specific SNP has been proposed so far.

Other infectious diseases

Thanks to GWAS, HLA associations with other infectious diseases have also been suggested. Regarding viral infections, numerous SNPs in the HLA-B region (among which rs114045064) have been associated with herpes zoster (shingles due to varicella-zoster virus, VZV), likely playing a role in viral suppression [140, 141]. A SNP located upstream to HLA-DQBl (rs9357152) has been associated with human papillomavirus (HPV) seropositivity [142], while both risk (HLA-DRB1*15:01–DQB1*06:02) and protective (HLA-DRB1*03:01–DQB1*02:01) HLA class II haplotypes had previously been suggested for this disease (cited by the same authors). Regarding bacterial infections, a significant association has been found at SNP rs7765379 near HLA-DQB1 and HLA-DRB1 for enteric fever due to infection by *Salmonella enterica*, pointing to allele HLA-DRB1*04:05 (after imputation) as a protective factor [143]. In addition, De Lorenze et al. [144] identified three SNPs associated with *Staphylococcus aureus* infection in HLA-DRA (rs4321864) and close to HLA-DRB1 (rs115231074, rs35079132). Among parasitic infections, Leishmaniasis (due to infection by protozoan parasites *Leishmania donovani* or *Leishmania infantum chagasi*) was significantly associated with SNP rs9271858 located between HLA-DRB1 and HLA-DQA1 in three distinct cohorts from India and Brazil [145], more recent studies also showing that HLA-DRB1*15:01 (with specific amino acid residues at positions 11 and 13) and HLA-DRB1*14:04/HLA-DRB1*13:01 were the most significant protective and risk alleles, respectively [146, 147]. By performing GWAS on a large sample of over 200,000 individuals of European ancestry recorded in the database 23andme, Tian et al. [29] also found new significant associations between several infectious diseases and SNPs located within the HLA region: chickenpox with rs9266089 within HLA-B; shingles with rs2523591 upstream to HLA-B (also reported in [141]); cold sores with rs885950 between HLA-A and HLA-C; mononucleosis caused by the Epstein-Barr virus (EBV) with SNP rs2596465 within the HCPI complex upstream to HLA-B; mumps with rs114193679 near HLA-A; plantar warts with rs9270205 within HLA-DQA1; streptococcal sore throat with rs1055821 within HLA-B; scarlet fever with rs36205178 within HLA-DQB1; pneumonia with rs3131623 within the HCPI complex; and childhood ear infections with rs4329147 between HLA-DRB1 and HLA-DQA1.
Interestingly, these authors identified several associations with amino acid polymorphisms in the peptide-binding region, thus involving specific HLA alleles like HLA-A*02:01 associated with chickenpox and shingles, and HLA-A*02:05 with mumps.

A schematic HLA genomic map showing the positions of the main HLA alleles and SNPs (according to the eEnsembl [148] and NCBI dbSNP [149] databases accessed on 9 December 2019 at https://www.ensembl.org and https://www.ncbi.nlm.nih.gov/snp, respectively) associated with the diseases mentioned in this review is provided in figure 1. Interestingly, this figure indicates that, in the current state of knowledge, specific HLA regions have been found to be involved in associations with AIDS (B-C region), tuberculosis (DRB-DQA1), hepatitis C (DQB1-DQA1) and (to a lesser extent) leprosy (mostly DRB-DQA1), whereas putative associations with hepatitis B and malaria are more widespread across the whole HLA region.

Population genetics in HLA and infectious disease association studies

This review shows that, in addition to case-control comparisons and immunogenetic analyses, new kinds of approaches such as GWAS, bioinformatic peptide-binding predictions and ancient DNA studies have helped to improve our knowledge on putative associations between HLA molecular variation and infectious diseases in the last decades. Of course, precise identification of causal markers remains complicated not only because the mechanisms by which the immune responses to infectious diseases is modulated are very diverse, as reviewed by Crux et al. [150] for HIV and HCV infections, but also because the analysis of the HLA polymorphism itself raises specific methodological problems and is especially challenging in a population genetics an evolutionary context [28].

First, HLA genes are particularly polymorphic. SNPs in the HLA region are often characterised by multiple non-synonymous substitutions at peptide-binding regions [151], such as rs1130380 located within HLA-DQB1 and associated with hepatitis C, and many SNPs are also multiallelic in non-coding or intergenic HLA regions, such as rs9271100, rs557011, rs9275572 and rs7453920 associated with leprosy, tuberculosis, hepatitis C and hepatitis B, respectively (major and minor alleles of each SNP can be found at https://www.ncbi.nlm.nih.gov/snp/). As a consequence, HLA variation is still difficult to characterise in particularly dense polymorphic regions (e.g., SNP array probes may be difficult to design), unless long-read sequencing techniques are applied, which represents a strong limiting factor for disease-association studies. Although imputation methods (i.e., methods that infer HLA variation from SNPs genotyped at sites flanking the highly polymorphic regions) are currently used [152], their accuracy varies greatly among studies because of numerous factors such as SNP density in the studied HLA region, as well as which reference panels and bioinformatics methods are used for imputation [153]. Moreover, the heterogeneous recombination patterns prevailing in the HLA genomic region [154–157] are a source of confounding effects, as many genetic markers in linkage disequilibrium may cosegregate in association to the same disease phenotypes without all being instrumental as protective or susceptibility factors. As an example, the rs9264942 (or -35Kb) SNP was considered as modulating HLA-C expression levels with direct consequences on HIV control before the linked rs67384697 in/del SNP was identified as a micro-RNA binder acting as a real regulatory factor. Also, nonadditive effects such as interactions among loci within the HLA region, recently investigated for a number of autoimmune diseases (e.g., [158–160]) may complicate the identification of one among several alternative models explaining the effects of HLA variation on disease phenotypes. Finally, the extreme HLA molecular variation observed within populations requires very large sample sizes to be satisfactorily characterised and for the statistical tests to reach sufficient power, avoid false positives and allow replication of the results, in particular because the large number of comparisons needed to consider all possible alleles and SNPs may fail to reach significance after correction for multiple tests (e.g., odds ratios are often close to 1 and/or p-values borderline, e.g., in [66]).

Besides the specificity of the HLA region in terms of high molecular variation within populations, HLA allele or SNP frequencies, as well as linkage disequilibrium across the HLA region, greatly differ between populations from different geographic regions [4, 5, 161]. Examples are given in figure 2, which displays global frequencies of the main HIV and malaria protective alleles, as well as those of two DQB1*03 subtypes that have been associated with hepatitis B, hepatitis C and/or tuberculosis (all data were taken from published articles and the allelefrequencies.net database [162]). This partly explains why disease association studies are often not observed in populations of distinct origins (e.g., for HBV-HCV [163]). A dramatic example is that of African populations, which are known to exhibit much higher levels of genetic diversity and much lower levels of linkage disequilibrium than populations from other continents both at the genome level [164] and across the HLA region [161]. This necessarily handicaps the discovery of significant disease associations, whereas Africa is one of the world regions where the greatest burden of infectious disease exists today (e.g., this region accounted for 68% of the world’s people living with HIV in 2018) while being still largely underrepresented in genetic variation studies [165–168]. A better characterisation of HLA molecular diversity in populations worldwide and the integration of population genetics data in disease-association studies are clearly essential.

HLA genetic differences between populations result from both selective effects due to the prominent role of HLA genes in immunity and to human history of migration since the emergence of modern humans from Africa, that likely happened between 300,000 and 200,000 years ago [169, 170]. Disease-association studies thus need to be controlled for population structure, for example by comparing HLA with genome-wide neutral variation in the same populations [16, 171]. If selective and demographic factors are correctly disentangled, differences in HLA frequencies between populations living in distinct environments may be more easily interpreted in terms of local adaptations to pathogens. However, selective signals may also be weak and difficult to identify because multiple alleles at a given HLA locus (e.g., through soft selective sweep) and/or sev-
eral (HLA and non-HLA) loci across the genome may jointly confer susceptibility or protection to the same diseases, malaria resistance probably being the best example of polyallelic and polygenic adaptation [127, 131, 172–174]. Moreover, host-pathogen interaction is a dynamic process where host molecules may drive changes in pathogen molecules and vice versa, leading to nonpermanent protective or susceptibility effects as a result of fluctuating or frequency-dependent selection [175]. Simultaneous analyses of host genome versus pathogen genome diver-

Figure 1: Schematic map of the HLA genomic region showing the positions of the HLA alleles and SNPs associated to six of the most prevalent infectious diseases in human populations. (A): AIDS, in red; (B): Hepatitis B, in blue; (C): Hepatitis C, in grey; (L): Leprosy, in orange; (M): Malaria, in pink; (T): Tuberculosis, in green. Protective HLA alleles are in bold, susceptibility HLA alleles are in italics. The genomic regions shown at the right as putatively involved in disease associations are only based on SNP information, except for malaria (dashed line). The genomic positions (GRCh38) of the HLA genes and SNPs have been determined by using the Ensembl [148] and NCBI dbSNP [149] databases accessed on 9 December 2019 at https://www.ensembl.org/index.html and https://www.ncbi.nlm.nih.gov/snp/, respectively.
sity in relation to specific diseases (e.g., [176]) may thus help to better understand the evolution of the HLA polymorphism in relation to its pathogenic environment.

Of course, the evolutionary timeframe is not the same for all disease. Some infectious diseases are known to be much more recent than others, making more difficult the exploration of host-pathogen co-evolution through time. For example, it has been suggested that the last common ancestor of HIV-1 pandemic strains dates back to the beginning of the 20th century [177], with a likely transmission of the disease through zoonotic infections from simian reservoirs and an early spread among humans from the region of Kinshasa in the Democratic Republic of Congo [178]. This corresponds to about four or five human generations, a very short time for protective factors to be positively selected and for susceptibility variants to be eliminated un-

Figure 2: Worldwide population frequency maps of some of the HLA-B alleles suggested to confer protection against AIDS (B*57:01, B*57:02, B*57:03, B*58:01) and malaria (B*53:01, B*35:01) as well as of two HLA-DQB1*03 alleles suggested to be associated to hepatitis B susceptibility (DQB1*03:01 and DQB1*03:03), to hepatitis C protection (DQB1*03:01) and to tuberculosis susceptibility (DQB1*03:03) (see text for references). In each map, the sizes of the pie charts are proportional to the frequencies of the allele in the corresponding populations, but pie charts are not at the same scale in the different maps. The maximum size of the pie charts in each map correspond to the following frequencies: B*57:01 = 0.082 (Tamil from South India; South Asia); B*57:02 = 0.042 (Xhosa from South Africa, sub-Saharan Africa); B*57:03 = 0.057 (Lusaka from Zambia, sub-Saharan Africa); B*58:01 = 0.170 (Han, China, East Asia); B*53:01 = 0.244 (Angolan, Sao Tome Island, sub-Saharan Africa); B*35:01 = 0.485 (Seri, Mexico, North America); DQB1*03:01 = 0.896 (Gila River Amerindian from Arizona, North America); DQB1*03:03 = 0.313 (Jehai from Malaysia, Southeast Asia). In each map, the crosses represent HLA-tested populations where the allele was not observed. All HLA population frequency data were taken from both published articles and the allelefrequency.net database [162].
der selective pressure, in contrast to the long timeframes for diseases that disseminated in much earlier times (e.g., before or during the Neolithic), such as tuberculosis, leprosy or malaria [179]. This probably explains, for example, why AIDS protective alleles are found at low frequencies (<10%), whereas some malaria protective alleles reach higher frequencies (>20%) in regions where these diseases are highly prevalent (fig. 2), even though soft rather than hard selective sweep mechanisms were at work.

Actually, distinct signals of selection resulting from different evolutionary processes and corresponding to either recent or ancient selective pressures are expected to be observed through population genetic studies. For example, long-term balancing selection likely decreased population differentiation, such as at the HLA-B locus [131, 180], whereas local adaptations to specific diseases probably drove adaptive HLA markers to high frequencies and led to marked genetic differences among populations, as at some HLA class II loci [137]. Such signatures are detectable through comparisons of HLA with neutral genomic variation [16, 181], computer simulations mimicking the evolution of the HLA polymorphism during human migration history with variable intensities of selection [182, 183] and other population genomics approaches taking into account the functional properties of the HLA molecules [184]. This is what makes population genetics approaches powerful to decipher the signatures of pathogen-driven selection on HLA molecular variability. Clearly, the identification of genetic markers involved in disease protection or susceptibility in the HLA region would largely benefit from a better integration of population genetics, with its demographic and evolutionary perspectives, in disease association studies.

Acknowledgments
This work was supported by the Swiss National Science Foundation, grant #310030_188820 (HLA-AFRICA). We also thank Da Di for his help with the allelic frequency maps and his very helpful comments, and José Manuel Nunes for stimulating discussions and two anonymous reviewers for their constructive comments.

Disclosure statement
No financial support and no other potential conflict of interest relevant to this article was reported.

References
genom-091212-153455
24 Remarque E. The dissection of complex susceptibility to infectious diseases: Current status and future perspectives from genome-wide ap-
The emerging role of HLA-C in HIV-1 infection


Zuckerman AJ. Hepatitis viruses. In: Medical Microbiology. Baron S, editor. Galveston, TX: University of Texas Medical Branch at Galveston; 1996.


Review article: Biomedical intelligence

Swiss Med Wkly. 2020;150:w20214

Published under the copyright license “Attribution – Non-Commercial – No Derivatives 4.0”.


The missing diversity in human T-cell repertoire

Tissue Antigens

2017

Waterboer SJ

J Immunol Res

2017

Kaul R

A high-resolution HLA and SNP haplotype map for disease associ-
ation studies in the extended human MHC. Nat Genet.

2006


Singh B, Kaur R, Kaul A, Khan K. A comparative review of HLA asso-
ciations with hepatitis B and C viral infections across global popula-


Tshabalala M, Meller J, Pepper MS. Human Leukocyte Antigen Diversi-

Dodda RG, Weale MR, Ardlie KG, Pask KM, Innes ALG, Tishkoff SA, et al. A comparative review of HLA asso-
ciations with hepatitis B and C viral infections across global popula-


Singh B, Kaur R, Kaul A, Khan K. A comparative review of HLA asso-
ciations with hepatitis B and C viral infections across global popula-


Tshabalala M, Meller J, Pepper MS. Human Leukocyte Antigen Diversi-

Dodda RG, Weale MR, Ardlie KG, Pask KM, Innes ALG, Tishkoff SA, et al. A comparative review of HLA asso-
ciations with hepatitis B and C viral infections across global popula-


Singh B, Kaur R, Kaul A, Khan K. A comparative review of HLA asso-
ciations with hepatitis B and C viral infections across global popula-


Tshabalala M, Meller J, Pepper MS. Human Leukocyte Antigen Diversi-

Dodda RG, Weale MR, Ardlie KG, Pask KM, Innes ALG, Tishkoff SA, et al. A comparative review of HLA asso-
ciations with hepatitis B and C viral infections across global popula-


Singh B, Kaur R, Kaul A, Khan K. A comparative review of HLA asso-
ciations with hepatitis B and C viral infections across global popula-


