Immunogenetics (2018) 70:5–27
DOI 10.1007/s00251-017-1017-3

REVIEW

A genomic perspective on HLA evolution

Diogo Meyer[1] · Vitor R. C. Aguiar[1] · Bárbara D. Bitarello[1,3] · Débora Y. C. Brandt[1,2] · Kelly Nunes[1]

Received: 25 October 2016 / Accepted: 16 June 2017 / Published online: 7 July 2017

Abstract Several decades of research have convincingly shown that classical human leukocyte antigen (HLA) loci bear signatures of natural selection. Despite this conclusion, many questions remain regarding the type of selective regime acting on these loci, the time frame at which selection acts, and the functional connections between genetic variability and natural selection. In this review, we argue that genomic datasets, in particular those generated by next-generation sequencing (NGS) at the population scale, are transforming our understanding of HLA evolution. We show that genomewide data can be used to perform robust and powerful tests for selection, capable of identifying both positive and balancing selection at HLA genes. Importantly, these tests have shown that natural selection can be identified at both recent and ancient timescales. We discuss how findings from genomewide association studies impact the evolutionary study of HLA genes, and how genomic data can be used to survey adaptive change involving interaction at multiple loci. We discuss the methodological developments which are necessary to correctly interpret genomic analyses involving the HLA region. These developments include adapting the NGS analysis framework so as to deal with the highly polymorphic HLA data, as well as developing tools and theory to search for signatures of selection, quantify differentiation, and measure admixture within the HLA region. Finally, we show that high throughput analysis of molecular phenotypes for HLA genes—namely transcription levels—is now a feasible approach and can add another dimension to the study of genetic variation.

Keywords HLA (human leukocyte antigen) · MHC (major histocompatibility complex) · Evolution · Genomics · Balancing selection

✉ Diogo Meyer
diogo@ib.usp.br

[1] Department of Genetics and Evolutionary Biology, University of São Paulo, 05508-090 São Paulo, SP, Brazil

[2] Present address: Department of Integrative Biology, University of California, Berkeley, CA, USA

[3] Present address: Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Introduction

The availability of genomic data at the scale of populations is transforming our understanding of the processes shaping human genetic variation. We are now able to answer questions which, little more than 15 years ago, seemed beyond our grasp. We can construct detailed portraits of how natural selection has acted, and identify variants that increased in frequency as a consequence of positive selection (the process that drives advantageous variants to high frequencies) (reviewed in Fu and Akey 2013). In some cases, it is possible to provide mechanistic links between the favored variant and its phenotypic effect, and to estimate the timescale of selection (for example, in the cases of variants involved in pigmentation (Beleza et al. 2013), lactase persistence (Coelho et al. 2005), and adaptation to altitude (Yi et al. 2010)).

There is also increasing interest in developing methods for the cases in which the advantageous variant was already present in the population at the time of onset of selection (i.e., selection on standing variation) (Messer and Petrov 2013). In addition, methods are being developed to identify instances in which selection favors a combination

of genetic variants (polygenic selection), instead of a single advantageous allele (Daub et al. 2013).

Genomic data is helping understand the rate at which we are burdened by deleterious mutations, and the importance of negative selection—which removes deleterious variants from populations—in the human genome (Fu et al. 2013; Henn et al. 2015). Deleterious variants have been hypothesized to play an important role in explaining phenotypic variation, particularly that of common diseases, and population level exome and genome sequencing are being used to tackle this question (with their role remaining controversial, Hunt et al. 2013).

Several studies have also searched for genes under balancing selection, which is the selective regime that maintains several variants in a population at intermediate frequencies, making the persistence time of each allele longer than that of neutral ones. Under this regime, the combination of alleles at a locus is often critical to defining fitness values, and the fitness of an allele may vary over time (reviewed in Key et al. 2014).

Information is also increasingly available for molecular phenotypes, helping understand the functional basis of natural selection. A particularly powerful method is RNAseq, which relies on next generation sequencing of RNA molecules to quantify gene expression. Using such information, Fraser (2013) showed that episodes of recent selection in humans are much more likely to affect gene expression than protein sequence.

Much of the progress in our understanding of how natural selection acts in humans is based on genomewide studies. However, focusing on genes for which we have prior functional knowledge can provide important insights on how natural selection acts. In this review, we integrate knowledge on the function of classical human leukocyte antigen (HLA) genes with population genomic data. We discuss how the genomic perspective both illuminates the study of HLA evolution, and contributes to our understanding of natural selection in the remainder of the genome.

HLA genes code for glycoproteins that bind peptides and present them to T cell receptors. If the bound peptide is non-self (i.e., possibly from a pathogen or a mutated protein), cellular and humoral responses can be mounted (see Box 1). HLA genes also interact with other molecules involved in innate and adaptive immunity. Among these are the killer cell immunoglobulin-like receptors (KIR), for which some HLA class I molecules are ligands (Trowsdale et al. 2001; Parham 2004). When cells are infected or neoplastic, the expression of classical class I loci may decrease, reducing the availability of ligand for KIR molecules. This activates cell lysis by natural killer cells (Yawata et al. 2008).

Research over the last three decades has successfully brought together knowledge on HLA function with advances in theoretical population genetics, allowing evolutionary hypotheses to be tested (in particular through the implementation of neutrality tests, Box 2). There are now several key ideas which are firmly established regarding HLA evolution. First, it is undisputed that HLA genes bear the mark of balancing selection: there are no demographic or genetic factors that can account for the unusually high degree of polymorphism, excess of nonsynonymous variants, or linkage disequilibrium at these genes (Meyer and Thomson 2001; Garrigan and Hedrick 2003; Spurgin and Richardson 2010). Second, there are several lines of support for a role of pathogen-driven selection in shaping HLA variation: HLA genes are associated with susceptibility and resistance to infectious disease (Cagliani and Sironi 2013); experimental studies show that pathogen pressure influences MHC variability (Penn et al. 2002); HLA polymorphism is correlated with pathogen diversity (Prugnolle et al. 2005); variation is highest at sites which define the peptide binding repertoire (Hedrick et al. 1991; Hughes and Nei 1988; Bitarello et al. 2016).

While it is clear that "documenting selection" at HLA genes is no longer a challenge, important questions regarding HLA evolution remain open, and can be addressed using genomic data. First, while it is accepted that balancing selection increases the diversity of HLA genes, there are several types of selection that can produce this effect. Balancing selection is an umbrella term that encompasses heterozygote advantage (or overdominance), selection varying over space or time, and negative frequency-dependent selection (see Box 3). Fleshing out which of these explains the high variability at HLA is a challenge (Spurgin and Richardson 2010), and we discuss the contributions of novel analytical methods and genomewide studies.

Second, the timescale of selection remains an open question. Tests of neutrality used before genomic data became available were only well-powered to detect long-term selection (Garrigan and Hedrick 2003), whereas newer approaches—which rely on dense genetic data spanning thousands of sites—can also detect recent selection (Field et al. 2016; Albrechtsen et al. 2010; Guan 2014). We discuss the findings brought by these approaches, and argue that they indicate that selection on HLA genes can be identified at various timescales.

Third, the increasing understanding of HLA function shows that interactions of HLA genes with other loci—and not just their immediate role in peptide binding—must also be considered in evolutionary studies (Trowsdale and Knight 2013). Further, phenotypic information, including expression levels of the HLA genes, has rarely been incorporated into evolutionary analyses. We discuss the challenges associated to bringing these functional perspectives to the study of HLA evolution.

HLA variation in the age of genome sequencing

Several generations of methods have been used to identify the alleles carried by an individual: PCR-RFLP, SSOP,

immobilized probes, PCR-SSP, and Sanger sequencing (reviewed in Erlich 2012; Carapito et al. 2016). The move to next-generation sequencing (NGS) is actively taking place, and in recent years many protocols have been described for HLA typing and SNP calling (Erlich et al. 2011; Lank et al. 2012; Wang et al. 2012; Danzer et al. 2013; Cao et al. 2013; Major et al. 2013; Langer et al. 2014; Monos and Maiers 2015; Norman et al. 2016; Zhou et al. 2016a).

When deep-sequencing data are available, which is usually the case for HLA-targeted protocols, the tiling of overlapped reads can provide phase information and thus HLA allele sequences (Hosomichi et al. 2013). However, when polymorphisms are on different and non-overlapping reads, statistical approaches to phasing must be used (Castelli et al. 2015, 2017; Lima et al. 2016). Mayor et al. (2015) presented a solution to both the genotype ambiguity and phasing issues by using the PacBio single molecule real time (SMRT) sequencing technology, which generates long reads spanning the entire sequence of individual HLA Class I genes. The method provided accurate and unambiguous HLA genotype calls, representing a promising prospect.

However, an understanding of the role of selection in shaping HLA variation also requires placing it in a genomewide context, so that selective and demographic factors can be disentangled, and genomewide significance testing can be performed. In practice, this requires extracting information on HLA variation from datasets with sequence information for the entire genome. Such data are increasingly generated by exome or whole-genome sequencing, as well as high density SNPs arrays (e.g., The 1000 Genomes Project Consortium 2010; Fu and Akey 2013).

Many genomewide studies, such as Phase I of the 1000 genomes project (The 1000 Genomes Project Consortium 2010), have analyzed HLA polymorphism using standard sequencing pipelines. Given the importance of the 1000 genomes project data to evolutionary research, we previously assessed the reliability of SNP calls which they provide (Brandt et al. 2015). We found that although frequency estimates for HLA SNPs are relatively robust (absolute frequency difference less than 0.1 for 75% of the SNPs), the SNP genotype calls within the HLA loci have alarmingly high error rates (18.6% of calls are incorrect) and are biased toward over-representing the alleles present in the reference genome.

This bias occurs because HLA genes are highly polymorphic, and standard methods align short reads (50 to 250 bp) to a single reference genome. Thus, individuals which are heterozygous at a site, but have one allele which is closer to the reference genome, are likely to only map that variant, with the other one failing to align (Fig. 1). The fact that HLA genes are members of a multi-gene family further complicates the sequencing, since reads from one locus can be incorrectly mapped to another.

An increasingly used strategy to address these challenges is to map short reads, generated by NGS, to multiple MHC/HLA references (e.g., IPD-IMGT/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/)), as opposed to a single reference genome. Recent methods have implemented this idea to efficiently provide more reliable alignments (Castelli et al. 2015, 2017; Lima et al. 2016), HLA allele calls (see Hosomichi et al. 2015 for a review and Bauer et al. (2016) for an evaluation of 12 computational methods), HLA expression estimates (Boegel et al. 2012), or to assemble individual genomes for the MHC region (Dilthey et al. 2015). Encouragingly, Phase III of the 1000 genomes project (The 1000 Genomes Project Consortium 2015) has used this strategy, allowing reads to align to 500 known HLA sequences, in addition to the human reference genome.

A more general solution is to perform genomic alignment using indices which account for the variation across the whole genome, including the MHC. These indices can be built in the form of genome graphs (Novak et al. 2017), an efficient strategy to summarize population-level variation in a graph structure, appropriate for subsequent short-read mapping. Application of such graph indices improves SNP calling in the MHC (Dilthey et al. 2015; Novak et al. 2017), and will likely supplant the use of a single linear reference index in the future. Overall, it appears that a more accurate assessment of HLA variability will come from both the development of new bioinformatic tools, as well as the generation of new data (in particular with long sequencing reads).

Another development is the imputation of HLA alleles based on dense SNP data. Imputation involves using a training set—for which both MHC region SNPs and HLA allele calls are available—to infer the HLA alleles carried by an individual with unknown HLA genotype, but for which SNP data is available (Dilthey et al. 2011; Zheng et al. 2013; Zhou et al. 2016a; Leslie et al. 2008). Zhou et al. (2016a) showed that the concordance rate between imputed HLA alleles and sequencing-based calls can reach 0.93 when using a large reference panel. Imputation is proving to be important in the context of association studies, since it allows an individual's HLA genotype to be included as a variable (Sanchez-Mazas and Meyer 2014), or even to infer specific amino-acids and amino-acid motifs, and quantify their contribution to overall associations (Jia et al. 2013).

However, imputation-based estimates will be uninformative with respect to novel variants, or those at very low frequencies. When interest is in identifying novel variants (Klitz et al. 2012), deep sequencing associated with mapping methods that account for variation will be required. In addition, imputation accuracy depends on the availability of reference panels with shared ancestry to the target population, representing an important challenge for studies of highly admixed populations with ancestral components

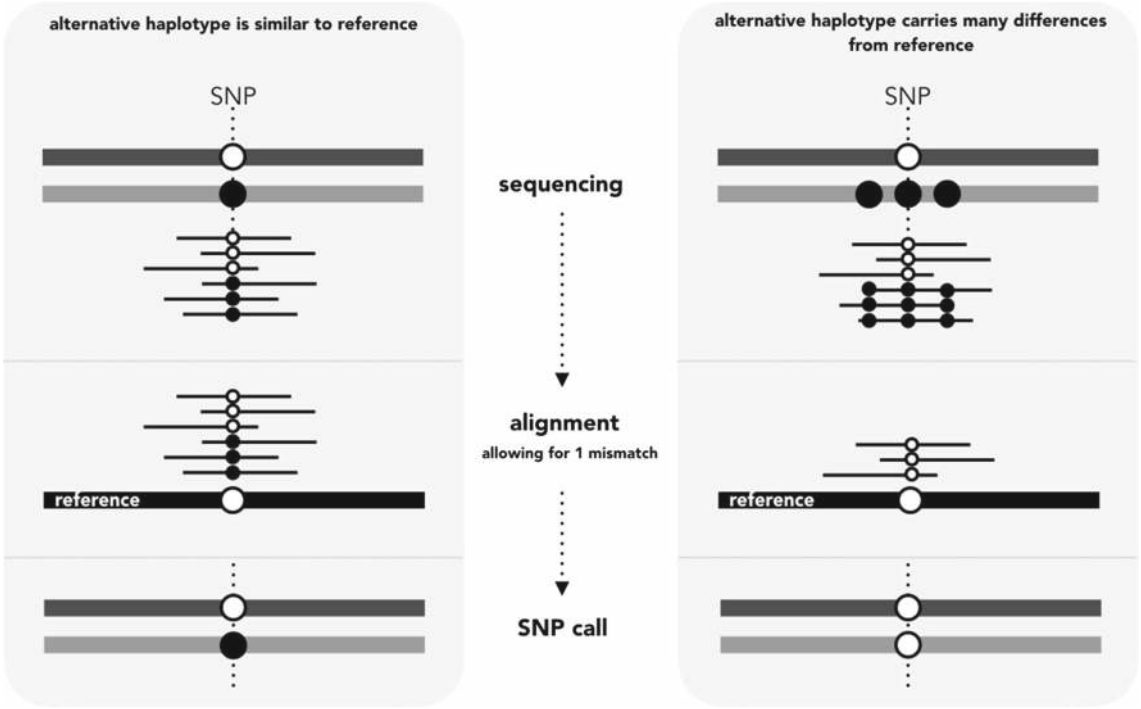

Fig. 1 How genotyping errors arise from the mapping of reads to a single reference genome. The *left panel* represents a case where sequence reads come from an individual who is heterozygous at a SNP, but the rest of the gene is similar to the reference for both haplotypes. The reads from both haplotypes can be aligned to the reference, and the SNP genotype is called correctly (i.e., determined by the analysis software). The *panel on the right* shows a case where one of the haplotypes is different from the reference sequence at more positions than the mismatch threshold (in this simple example, only one mismatch is allowed). Reads from this haplotype will not align to the reference sequence and the genotype will be incorrectly called as homozygous at the SNP of interest. Modified from the *Genes to Genomes* blog, http://genestogenomes.org/the-trouble-with-hla-diversity/

which are relatively poorly studied (Levin et al. 2014; Nunes et al. 2016).

In conclusion, we now have access to a wide array of options for uncovering HLA variation. Whereas genomewide sequencing based on alignment to a reference genome generates biased allele frequency estimates, pipelines that account for known HLA diversity can generate accurate information (Dilthey et al. 2015). Importantly, whole genome sequencing places HLA data in a genomewide context, an ideal scenario for separating demographic and selective contributions to variation, as we discuss in the next section.

Genome scans for balancing selection

The early work on selection at HLA loci was carried out in the "candidate gene" framework, wherein specific HLA loci were tested for selection (see Box 2) (e.g., Hedrick and Thomson 1983, 1986; Hughes and Nei 1988). With genomewide data, on the other hand, it is no longer necessary to *a priori* define which loci will be queried for selection, allowing us to investigate how extreme the evidence for selection at HLA loci is with respect to the remainder of the genome.

Most genomewide scans for selection search for genes that underwent positive selection. The main signatures of this mode of selection are: low variability coupled with extended linkage disequilibrium, caused by the increase in frequency of a favored variant; high population differentiation, due to selection favoring locally adaptive alleles; and an abundance of low frequency variants, due to mutations introducing novel variants into a region recently homogenized by selection (reviewed in Fu and Akey 2013) (see Box 2). Because many of these signatures can also result from non-selective events such as population expansions and bottlenecks, it has become standard for tests of selection to explicitly control for demographic history (e.g., by simulating null distributions under realistic scenarios) (Nielsen et al. 2005). These simulations are parametrized by estimates of the demographic history based on the genomewide data itself. In this way, sets of genes under positive selection have been identified in a robust manner (Akey 2009).

Although there was strong support for positive selection on genes related to immunity (e.g., Nielsen et al. 2005; Tang et al. 2007b; Carlson et al. 2005), few genomic scans found evidence for it in the extended MHC region. Exceptions are the studies of de Bakker et al. (2006) and Sabeti et al. (2006), which identified long range haplotypes in the MHC

region. The weak support for selection on HLA genes across several genomewide studies (Akey 2009) is largely a consequence of the fact that they used tests designed to detect positive—and not balancing—selection (Box 2).

In order to detect balancing selection, it is necessary to develop statistics sensitive to deviations expected under this selective regime. Appropriate tests include searching the genome for regions with ancient shared polymorphisms (e.g., Leffler et al. 2013; Teixeira et al. 2015), extreme patterns of polymorphism relative to divergence (e.g., DeGiorgio et al. 2014; Andrés et al. 2009; Bitarello et al. 2017), an excess of intermediate frequency variants (DeGiorgio et al. 2014; Andrés et al. 2009; Bitarello et al. 2017; Hedrick and Thomson 1983), an excess of identity by descent (IBD) (Albrechtsen et al. 2010), or unusually low differentiation between populations (Hofer et al. 2012; Sanchez-Mazas 2007) (Box 2).

Tests using these approaches have been implemented, and the findings for HLA genes are summarized in Table 1. All studies show hits in the MHC region, with *HLA-B* appearing in five out of the six scans (Andrés et al. 2009; DeGiorgio et al. 2014; Leffler et al. 2013; Teixeira et al. 2015; Hofer et al. 2012; Bitarello et al. 2017). In addition, HLA genes show the most extreme evidence of balancing selection in tests based on ancient shared polymorphisms (Klein et al. 1993; Teixeira et al. 2015; Leffler et al. 2013), and are highly enriched for extreme p-values in tests based on polymorphism and divergence (e.g., DeGiorgio et al. 2014; Andrés et al. 2009; Bitarello et al. 2017). This not only confirms that HLA genes have been under long-term balancing selection but also shows that they are extreme in their patterns of diversity, compared to non-HLA loci.

The MHC region is also the most extreme in a test based on identity-by-descent (IBD), which identifies genomic regions with extensive identity among individuals, consistent with the hypothesis that they descend from an advantageous ancestral variant (Albrechtsen et al. 2010). This signature supports very recent selection (<500 generations, or 10,000 years), which can be positive or balancing. Interestingly, Albrechtsen et al. (2010) showed that the increase in IBD is not expected under heterozygote advantage, leading them to argue that selection at HLA loci may be frequency-dependent, or to fluctuate over time, possibly tracking changes in the evolving pool of pathogens that individuals are exposed to.

Important developments in our understanding of HLA evolution have also come from two recent technological breakthroughs: the ability to sequence ancient samples and the genomic analysis of extremely large samples. Using over 200 ancient genomes, Mathieson et al. (2015) found several loci in modern Europeans which experienced greater changes in allele frequencies (with respect to their presumed ancestors, as inferred using the ancient samples), than expected under drift alone. Within the MHC region of Europeans there are at least seven independent signals for selective changes (consistent with both balancing selection or the occurrence of multiple sweeps). New findings also came from the study of Field et al. (2016) which used the theoretical prediction that recently selected variants should be associated with a less diverse genetic neighborhood than the non-selected variants. Leveraged by very large samples of sequence data, they identified genomic regions where selection has driven advantageous alleles to high frequencies in a time frame as recent as 2 000 years, and found that at least three independent SNPs within the extended MHC region were among the most significant targets (Field et al. 2016). This test is designed to detect recent positive selection, implying that balancing selection should not be seen as the only regime relevant to HLA evolution.

Finally, a recent study sequenced genomes of an extant population from the Northwest coast of North America, along with ancient genomes of individuals presumably from the same group, but from before contact with Europeans (Lindo et al. 2016). The study found that at *HLA-DQA1* there was a shift from past positive to recent negative selection, bringing about marked allele frequency changes. The authors conjecture that this may have resulted from environmental or social changes.

In summary, genomic scans for selection have revealed two important patterns. First, when tests designed to identify balancing selection are used, evidence for selection at HLA genes is strong and extreme with respect to the remainder of the genome, confirming what was known based on candidate gene approaches. Second, two studies have identified selection within the MHC region that is consistent with regimes other than heterozygote advantage, and involving very recent time frames (Albrechtsen et al. 2010; Field et al. 2016). According to these studies, and also a recent ancient-DNA study of Lindo et al. (2016), selection drove recent changes in allele frequencies (e.g., via frequency-dependent selection, or selection in a fluctuating selective environment). This supports the view that several selective regimes account for the patterns of variation of HLA genes.

Disease associations

Identifying HLA variants that contribute to resistance to infectious diseases has important evolutionary implications. Simply put, alleles conferring disease resistance are compelling evidence for past and ongoing selection.

A standard approach for identifying genetic variants that contribute to disease phenotypes is to carry out association studies. These compare the frequencies of genetic variants in groups that differ in a phenotype of interest, such as the occurrence of a specific disease. Thus, for example, if a variant

Table 1 Findings for HLA genes in genome scans for balancing selection

Reference	Method	Selection timescale[d]	Selection at HLA
Andrés et al. (2009)	SFS and polymorphism/divergence ratio	Ancient	*HLA-B*[a]
Albrechtsen et al. (2010)	Excess IBD regions[b]	Recent	Entire MHC region
Leffler et al. (2013)	Long-term shared polymorphism	Ancient	*HLA-B*[c], *HLA-DQA1*, *HLA-DQB1*, *HLA-DPB1*
DeGiorgio et al. (2014)	Composite likelihood	Long-term	*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB5*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPB1*
Teixeira et al. (2015)	Long-term shared polymorphism	Ancient	*HLA-C*, *HLA-DQA1*, *HLA-DPB1*
Bitarello et al. (2017)	SFS and polymorphism/divergence ratio	Long-term	*HLA-B*, *HLA-C*, *HLA-DPA1*, *HLA-DQA1*, *HLA-DPB1*, *HLA-DRB1*, *HLA-DRB5*, *HLA-DQB2*, *HLA-DQB1*, *HLA-G*

IBD identity-by-descent, *SFS* site-frequency spectrum

[a]Out of five HLA genes analyzed

[b]A signature compatible with both positive and balancing selection

[c]The shared polymorphism falling in this gene is a CpG site (has higher mutation rate and could reflect recurrent mutation)

[d]Long-term: more than 1 million years ago; ancient: greater than species-divergence time (6 million years, for humans and chimps)

is significantly less common among those with the disease than those without it, it is said to be associated with protection from the disease (provided that case and control groups are carefully controlled for possible confounding variables). Through much of the 1980s and 1990s, HLA variants were tested for association with resistance or susceptibility to infectious diseases. These studies revealed a large number of associations with infectious diseases, some of the most studied being leprosy, malaria, chronic viral hepatitis, and further into the 90s, HIV/AIDS (see Blackwell et al. 2009, for a thorough review). However, these early studies carried important limitations: samples sizes were modest, typically on the order of hundreds, and *a priori* selected candidate genes were investigated, making it difficult to differentiate between associations which were causal or driven by linkage disequilibrium.

The explosion of data that has occurred in the last decade has brought about important changes. Millions of genetic markers are now queried in extremely large samples, allowing genomewide association studies (GWAS) to identify genes or genomic regions associated with diseases, without having to define beforehand the candidate loci to be queried. These association studies are bringing important contributions to our understanding of how genetic variation at HLA genes is related to response to pathogens. Below, we highlight four insights.

First, the recent generation of GWAS have confirmed that variation at HLA genes is directly associated with the outcome of many infectious diseases. Among these are HIV (Fellay et al. 2007), leprosy (Zhang et al. 2009), hepatitis (Kamatani et al. 2009), and tuberculosis (Sveinbjornsson et al. 2016).

Second, diseases which until recently were impractical to study in a GWAS setting can now be investigated. A remarkable example is the analysis led by the personal genomics company 23andMe, which performed an association study for infectious diseases in a sample of 200,000 customers which had volunteered information on various medical conditions (Tian et al. 2016). The study found that variation at HLA genes or within the MHC region is associated with viral (chickenpox, shingles, cold sores, mononucleosis, mumps, warts caused by papillomavirus, strep throat, scarlet fever, pneumonia) and bacterial (tonsil infections, ear infections) diseases.

Third, because GWAS query SNPs throughout the entire MHC region, it is possible to fine-map associations, i.e., identify associations within a narrower region of the genome. This has shown that several associations involve sites with regulatory function. For example, AIDS progression is associated with a 5' UTR regulatory variant of *HLA-C* (Kulkarni et al. 2011) and hepatitis B recovery is associated with variation at a 3' UTR site which modulates *-DPB1* expression (Thomas et al. 2012). From an evolutionary perspective, this indicates that selection on HLA genes is not restricted to the structural domains involved in peptide binding, but also involves regulatory variants.

Fourth, dense SNP data allows HLA alleles to be imputed (see Section 3) and thus the amino acid sequence coded

by HLA genes to be inferred. In this way, it is possible to study associations at the molecular level, identifying specific changes in a protein that are associated with disease resistance or susceptibility (Nishida et al. 2016; Tian et al. 2016).

Even more activity has taken place in the study of genetic associations with autoimmune diseases. Samples of tens of thousands have routinely been assembled, and copious associations with the MHC region or specific HLA genes have been firmly established, including diabetes, arthritis, celiac disease, lupus, ankylosing spondylitis, multiple sclerosis, psoriasis, and Crohn's disease (reviewed in Trowsdale and Knight 2013). From an evolutionary perspective, the existence of autoimmune conditions associated with relatively common HLA alleles poses an important question: if the disease reduces an individual's chances of survival and reproduction, why have the underlying alleles not been driven to low frequencies?

To answer this question, an influential working hypothesis that the same alleles which conferred resistance to infectious diseases and rose in frequency are also associated with autoimmune conditions (Corona et al. 2010; Sams and Hawks 2014; Abadie et al. 2011). This suggests a trade-off occurs, where the benefits brought by disease resistance outweigh the fitness costs of autoimmunity. A formal test involves asking whether alleles that are associated with autoimmune disease risk have increased evidence of having experienced selection. In the context of non-HLA variants, Fumagalli et al. (2011) found a correlation between the abundance of autoimmune disease predisposing variants and pathogen abundance, an indirect support for the trade-off hypothesis. Specifically for HLA, Abadie et al. (2011) examined whether the *HLA-DQA1* variant which predisposes to celiac disease showed evidence of past selection, but found no support. Corona et al. (2010) surveyed GWAS for complex diseases, and found that for type 1 diabetes strongly predisposing SNPs are also those with strong evidence for positive selection.

Although this approach has not yet delivered a clear picture, the strong evidence of pathogen-driven selection at HLA genes, coupled with the extreme abundance of HLA involvement in autoimmunity, call for further development of evolutionary approaches investigating the possibility that there is a causal connection between evolutionary response to infectious diseases and autoimmunity.

Multilocus effects: epistasis and hitchhiking

There is increasing awareness that many adaptive traits are polygenic, and that searching for allele frequency changes at multiple loci is an important improvement over "single locus" approaches (Daub et al. 2013; Berg and Coop 2014). There are several reasons why we expect adaptation involving HLA genes to be polygenic, which we discuss below.

There is support for epistatic interactions between variants at distinct HLA loci, driving advantageous haplotypes to higher frequencies than expected by chance, and thus explaining the high linkage disequilibrium in the MHC. One reason why a haplotype may be favored is that it carries a combination of alleles that presents a broader range of pathogenic peptides than expected for a random pair of alleles. This hypothesis was recently supported by a theoretical model, as well as data analyses showing that alleles in linkage disequilibrium on average have a lower overlap in the peptide binding repertoire than expected by chance (Penman et al. 2013). Using a simulation-based approach, van Oosterhout (2009) also illustrated that epistasis among HLA loci can play an important role in shaping extant patterns of diversity. Finally, GWAS for HLA loci found multi-locus effects, as is the case of the association of the *DR2* haplotype (*DRB1*1501* and *DRB5*0101*) with multiple sclerosis (Gregersen et al. 2006).

Second, multi-locus interactions have also been documented between HLA genes and those outside the extended MHC (see Box 1). For example, Kirino et al. (2013) found a strong epistatic interaction between *HLA-B*51* and the *ERAP1* locus, with one specific genotype greatly increasing the susceptibility to Behçet's disease. *ERAP1* codes for the protein responsible for trimming the pathogens to be loaded and presented by HLA class I molecules, making interactions between it and HLA genes functionally plausible.

Another case of epistasis involves the interaction between HLA and KIR. KIR molecules can recognize HLA class I molecules carrying HLA-A3, -A11, -Bw4, -B27, -C1, or -C2 epitopes, as well as *HLA-F* and possibly *HLA-G* (reviewed in Parham et al. 2012). In a study of 30 human populations, Single et al. (2007) found a strong negative correlation between the frequency of *HLA-B* alleles of the *Bw4* group, which carry an isoleucine at position 80, and the presence of *KIR3DS1* gene. Because *Bw4* alleles are ligands for *KIR3DS1*, which is an "activator" (a gene whose protein product initiates a cytotoxic response), the combination of high frequencies of ligand and receptors would result in an abundance of excessively activating genotypes, which are prone to autoimmunity. At the other extreme, combinations of low frequencies of ligand and *KIR3DS1* would result in an excessively weak KIR response, increasing the susceptibility to infection. Selection against genotypes at these extremes could account for the observed correlations seen in Single et al. (2007). Using a similar approach, Hollenbach et al. (2013) found strong ($r > 0.79$) and significant correlations between the frequencies of *KIR2DL3* and HLA-C1 in 45 populations.

Support for these interactions also comes from the study of specific populations. In the African KhoeSan, the C2

allotype occurs at an unusually high frequency (63%), whereas in the Yucpa of South America it is the C1 allotype that is common (83%) (Hilton et al. 2015; Gendzekhadze et al. 2009). Strikingly, in both populations, the receptors for these common allotypes show evidence of having been recently selected and driven to high frequencies, with the mutant forms having reduced or complete lack of function. In both cases, these population-specific variants may have been favored due to their ability to restore a balance between C1, C2, and the KIR inhibitory allotypes, providing the benefits of reducing the chances of originating preeclampsia predisposing genotypes (see below). Functional studies provide further support for epistasis, showing that homozygotes for HLA-C1 respond more intensely to a viral infection than those carrying HLA-C2 alleles (Ahlenstiel et al. 2008, see also Augusto et al. 2015, for an example involving the autoimmune disease pemphigus).

The epistatic interactions between KIR and HLA also influence reproduction. For example, mothers homozygous for the KIR haplotype from group A (defined by the presence of four framework genes—*KIR2DL4*, *KIR3DL2*, *KIR3DL3*, and *KIR3DP1*—and *KIR2DL1*, *KIR2DL3*, *KIR2DS4*, and *KIR3DL1*) have an increased rate of miscarriage, pre-eclampsia, and weight restriction at birth when they also carry an HLA-C1 allele and the fetus has an HLA-C2 allele. This results from a less effective remodeling of blood vessels, necessary for placentation (Penman et al. 2016; Hiby et al. 2014; Hiby et al. 2004). On the other hand, individuals with group A KIR haplotypes and HLA-C1 alleles respond to viral infections more efficiently than individuals with group B haplotypes (which carry genes encoding KIRs with decreased or no binding to HLA class I molecules, such as *KIR2DS2*, *KIR2DS3*, and *KIR2DS5*) in combination with HLA-C2 alleles (e.g., hepatitis C and HIV clearance). This tradeoff may result in alternating episodes of reproductive and pathogen-driven selection, explaining the maintenance of polymorphism for KIR haplotypes and for the HLA-C1 and -C2 group alleles in many human populations. This scenario was supported by computer simulations (Penman et al. 2016) and is consistent with patterns of HLA and KIR polymorphism in many human populations (see details in Trowsdale and Moffett 2008; Parham and Moffett 2013; Augusto and Petzl-Erler 2016).

Strong selection at a locus can also influence variation at linked sites through genetic hitchhiking. Under pathogen-driven selection an advantageous variant is driven to higher frequencies at a greater speed than would be expected under drift, and can thus drag linked variants (Charlesworth 2006). This selective regime can increase the frequency of slightly deleterious mutations near the selected gene. Accordingly, Chun and Fay (2011) showed that for regions in the neighborhood of sites with strong evidence for positive selection, there is an enrichment for deleterious polymorphism.

In the context of the MHC region, a natural hypothesis is that genes close to the classical HLA loci will show an enrichment of deleterious variants, with respect to the expectations based on genomewide controls. Mendes (2013) investigated this hypothesis, and in an analysis of the 1000 Genomes data (The 1000 Genomes Project Consortium 2010) found that genes that hitchhike with HLA loci have an increased proportion of putatively deleterious variants (Fig. 2). This hypothesis was also tested by Lenz et al. (2016), who used a larger exome-based dataset to show an excess of intermediate frequency deleterious polymorphism within the MHC. Further, these authors used simulations to show that strong balancing selection—comparable in strength to that seen at HLA genes—makes deleterious variants more common than would be expected without the hitchhiking effect.

These findings are particularly important given the large number of disease associations in the MHC region (including the flanking non-HLA loci), suggesting that balancing selection in HLA genes may drive the accumulation of deleterious variants in their neighborhood, contributing to the associations with disease phenotypes.

To conclude, we emphasize that ongoing research recommends that variation at HLA genes be studied with reference to both the genes they interact with, as well as considering how physical linkage leads to changes in polymorphism at neighboring sites. Placing HLA variation in a genomewide context will be essential in order to achieve these goals.

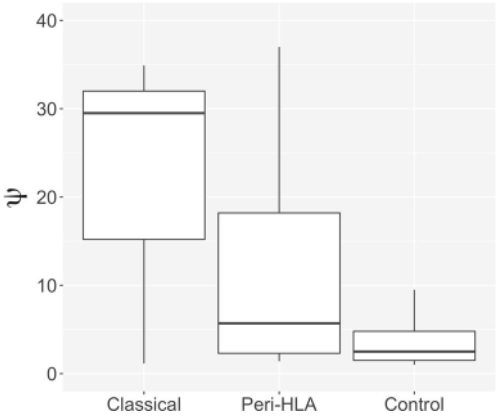

Fig. 2 The value of ψ, a statistic that measures the proportion of deleterious variants, in three sets of SNPs. The statistic is defined by $\psi = \frac{L_S \cdot P_N}{L_N \cdot (P_S+1)}$, where P represents the number of polymorphic sites, L represents the number of potentially mutable sites, and S and N subscripts refer to synonymous and nonsynonymous sites. Higher values of ψ indicate a greater proportion of deleterious (or functional, in the case of the SNPs from the classical HLA genes) variants. Values are shown for exons of classical HLA genes, genes in the immediate neighborhood of the HLA genes ("peri-HLA"), and genes outside the MHC region. Values were computed for sites with a minor allele frequency (MAF) greater than 0.05, to avoid the effect of rare deleterious variants, which are overrepresented in the control set. The peri-HLA genes have higher load (ψ) than the controls

Population differentiation

If distinct populations are under a regime of selection favoring HLA heterozygotes, population differentiation, measured by F_{ST}, is expected to be lower at HLA than at neutral loci (Schierup et al. 2000). This is because balancing selection maintains alleles segregating in populations for longer than expected under neutrality, reducing F_{ST} (Box 2).

An alternative scenario is that selection favors different alleles in distinct populations, driving locally adaptive HLA alleles to higher frequencies, increasing population differentiation.This expectation is consistent with pathogen-driven selection at HLA, for which there is theoretical (Borghans et al. 2004; Hedrick 2002) and empirical support (e.g., Prugnolle et al. 2005; Hedrick 2006). Given the premise that pathogen populations differ between regions, pathogen-driven selection could drive locally adaptive HLA alleles to higher frequencies, and thus cause an increase in population differentiation.

Surprisingly, support for both of these markedly different expectations has been found (Table 2), with some studies showing HLA to be unusually highly differentiated, and others reporting unusually low differentiation at HLA. What is the cause for the inconsistency among studies? Analyses using F_{ST} are sensitive to various aspects of the methodology, all of which can influence the results, as we discuss below.

First, studies which compare different markers, such as HLA alleles and microsatellites, are sensitive to the effects of the mutational mechanism and mean heterozygosity on F_{ST}, making direct contrasts between HLA and non-HLA markers unreliable (a challenge for the studies of Meyer et al. 2006; Sanchez-Mazas 2007). Second, the statistical tests used to define extreme F_{ST} differ among studies, including outlier approaches, tree-based tests, simulation under an various demographic models, among others (Table 2). Third, the power to detect balancing selection may vary depending on the timescale of separation of populations, and features of their demographic histories (reduced HLA differentiation being harder to detect in admixed populations, for which genomewide F_{ST} is lower). Fourth, SNPs with low heterozygosities are constrained to low F_{ST}, implying that HLA and non-HLA SNPs must be compared in a way that accounts for this effect (Bhatia et al. 2013).

In order to overcome these issues, we analyzed HLA differentiation among major continental groups, accounting for these effects (Brandt 2015) (Fig. 3). Marker-type effects are accounted for by only analyzing SNP data. The non-HLA SNPs provide expectations due to demographic processes, allowing a statistical assessment of how extreme the differentiation is for SNPs within HLA loci. F_{ST} values for SNPs in the HLA and non-HLA groups are averaged using an approach that controls for the differing heterozygosity distributions in those groups (Reynolds et al. 1983; Bhatia et al. 2013). With these methodological controls in place, the results in Fig. 3 show that SNPs within HLA genes have lower F_{ST} than genomewide SNPs when we compare highly diverged populations (i.e., those from different continents). Population pairs from the same continent have higher differentiation for SNPs in the HLA genes compared to other genomic regions.

How do these findings compare to those of previous studies? Low differentiation among HLA SNPs is consistent with the findings of Hofer et al. (2012), which detected a similar pattern in a dataset including highly divergent human populations. The increased differentiation seen by Bhatia et al. (2011) among African populations is also consistent with this result, since that study analyzed closely related populations.

Table 2 Population differentiation at HLA genes relative to neutral markers

Reference	Neutral marker	HLA marker	Method	F_{ST} in HLA
Akey et al. (2002)	SNP	SNP (genomewide scan)	Empirical outlier	Not an outlier
Meyer et al. (2006)	Microsatellites	HLA allele[a]	Empirical outlier	Not an outlier
Sanchez-Mazas (2007)	Microsatellites and RFLPs	HLA allele[a]	Empirical outlier	Lower in HLA
Bhatia et al. (2011)	SNP	SNP (genomewide scan)	Tree-based test	Higher in HLA
Nunes (2011)	Microsatellites	Microsatellites	Simulation	Higher in HLA
Hofer et al. (2012)	SNP	SNP (genomewide scan)	Simulation	Lower around *HLA-C*
Colonna et al. (2014)	SNP	SNP (genomewide scan)	Empirical outlier + clustering	Not an outlier
Brandt (2015)	SNP	SNP and HLA alleles	Empirical outlier	Lower for HLA SNPs; HLA alleles are not outliers

[a]See Box 1 for the definition of HLA allele

Also, one of the SNPs driving the high differentiation reported in Bhatia et al. (2011) was linked to *HLA-DPA1*, a locus we excluded because it did not show strong evidence of balancing selection in previous studies (Solberg et al. 2008; Begovich et al. 2001), and showed instances of directional selection (Hollenbach et al. 2001). Indeed, for *HLA-DPA1* population differentiation was higher than genomewide in our data as well, consistent with local positive selection. Interestingly, *HLA-DPA1* has one of the strongest signatures of long term balancing selection in Bitarello et al. (2017). A plausible scenario is that *HLA-DPA1* is under a selective regime that varies through time, leaving a signature of past balancing selection and more recent local positive selection.

Given the overall result that natural selection on HLA genes, over long periods of time, results in decreased population differentiation (Fig. 3, *y*-axis), it is natural to consider how to reconcile this with the expectation that pathogens would drive local adaptation, making populations more different from one another at HLA genes. There are two possible ways in which low differentiation at HLA SNPs can be reconciled with a model of local adaptation of HLA alleles.

First, the signal of local adaptation (high differentiation) may only be detectable when comparing closely related populations, such as the ones in the same continent. Indeed, previous studies have detected high differentiation in HLA alleles between populations within the same continent (Cao et al. 2004; Qian et al. 2013), and we have detected higher F_{ST} at HLA SNPs than genomewide SNPs for pairs of populations in the same continent (Fig. 3).

Second, low differentiation at SNPs and high differentiation at HLA alleles may be expected if we consider that HLA alleles are defined by multiple SNPs, and that most SNPs are shared between two or more alleles. The important role that intragenic recombination and gene conversion play in generating HLA allele diversity also contributes to the sharing of SNPs among different HLA alleles (Parham and Ohta 1996). Thus, a plausible scenario is that individual SNPs have low F_{ST}, but the haplotypes which they define may show high divergence. Biologically this amounts to considering that balancing selection favors the maintenance of polymorphism at specific sites, key to defining peptide binding specificities (Bitarello et al. 2016). However, the specific combinations of variants (i.e., the HLA alleles) that become more frequent differ among populations as a function of the pathogens driving the selection.

Selection and admixture

Individuals in admixed populations have genomes which are a mosaic of different ancestries (Winkler et al. 2010). The size and ancestry of segments is determined by factors which are demographic (e.g., proportion of ancestors from each ancestry, timing of admixture) and genetic (e.g., recombination rates). If genetic variants from one of the

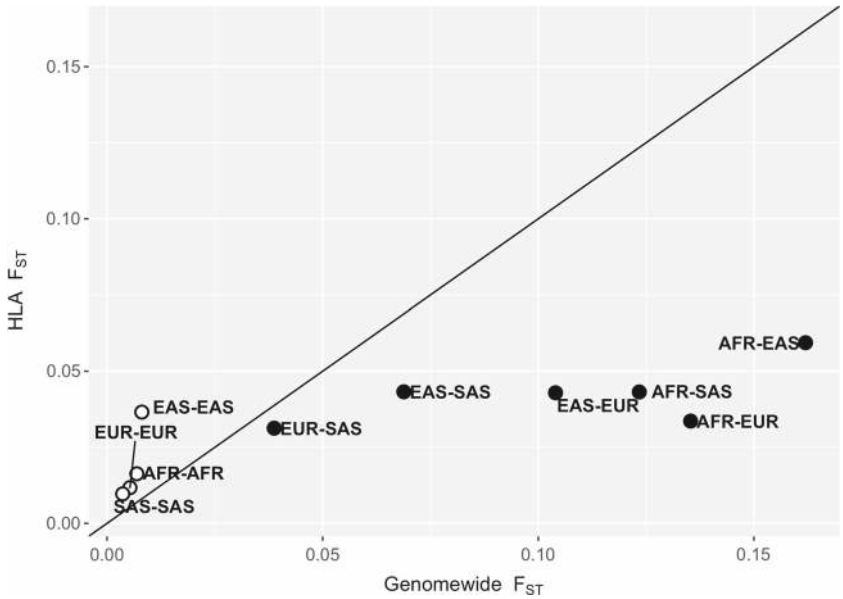

Fig. 3 F_{ST} among pairs of populations. Each point depicts the mean F_{ST} for non-HLA (*x*-axis) and HLA (*y*-axis) SNPs between pairs of populations in each continent (AFR: Africa; EAS: East Asia; EUR: Europe; SAS: Southeast Asia). Pairs of populations from the same continent are represented by white-filled points, and pairs of populations from different continents, by *solid black points*. SNP data was acquired from the 1000 Genomes data phase III (The 1000 Genomes Project Consortium 2015), and HLA SNPs were filtered according to Brandt et al. (2015) to avoid errors due to mapping bias. F_{ST} values were weighted by allele frequency, so that the excess of rare variants in the non-HLA SNPs does not cause a reduction of mean F_{ST} in that class. Notice that HLA differentiation is higher than genomewide for population pairs from the same continent, and lower than genomewide when populations from different continents are compared

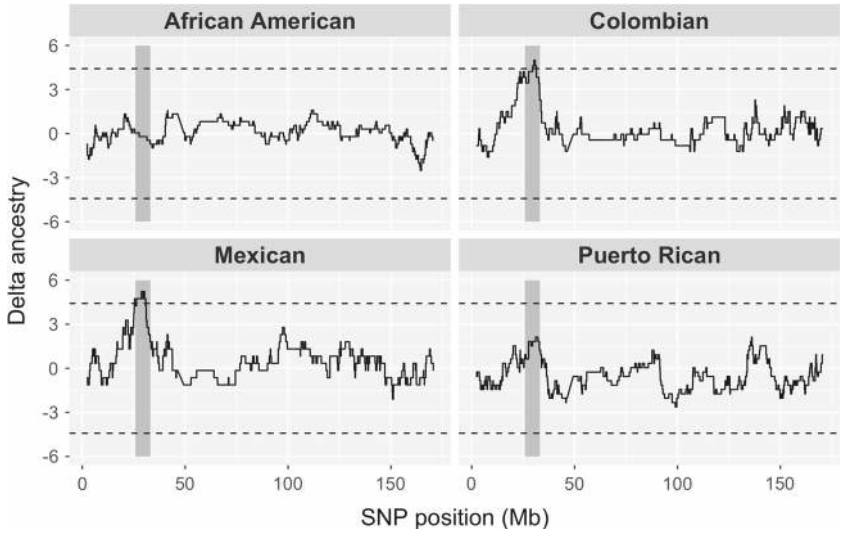

Fig. 4 Deviation from average genomewide ancestry in four admixed populations along chromosome 6. The degree to which local ancestry deviates from genomewide averages is shown for African ancestry (*black lines*). The region encompassing the MHC region is indicated by *gray shading*. Ancestral and admixed populations are from the 1000 genomes project (African and European; ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/), except for the ancestral Native American sample, which is from the HGDP-CEPH (http://www.cephb.fr/hgdp/index.php). Local ancestries were estimated using RFMIX (Maples et al. 2013). The ancestry deviation measure is the difference between ancestry at a given genomic position with respect to the genomewide average, normalized by the standard deviation of the ancestry estimate (thus providing a measure of the number of standard deviations each ancestry departs from its genomewide average)

parental populations are advantageous to individuals in the admixed population, they will rise in frequency and thus cause an over-representation of a specific parental ancestry in the genomic region under selection. Thus, regions of the genome exhibiting ancestry proportions that deviate from the genomewide average provide evidence for recent selection.

To illustrate the power of this approach in understanding selection at HLA genes, we calculated local ancestries (i.e., the ancestry of a specific position of the genome) for individuals from four admixed populations (The 1000 Genomes Project Consortium 2010). For each position in the genome, we quantified how much the ancestry proportions differed from the genomewide average, within each population. For chromosome 6, we find that two of the populations (Colombian and Mexican) have an excess of African ancestry in the MHC region (the threshold of significance set at 4.4 standard deviations, following Seldin et al. 2011) (Fig. 4).

To explore this pattern further, we reviewed the findings of eight studies that investigated the distribution of local ancestries and recorded how often the MHC showed unusual ancestry proportions with respect to genomewide averages. In total, six out of eight studies report an excess of African ancestry in the MHC region for at least one admixed population (Table 3). Interestingly, this effect is seen in populations with different admixture histories, distinct African parental populations and proportion of contributions, and using different methods to estimate local ancestry. Overall, the support for deviation

Table 3 Ancestry proportions in the MHC region vs genomewide

Reference	Admixed population	Method	Observation
Tang et al. (2007a)	Puerto Rican	Frape	Excess African
Johnson et al. (2011)	Mexicans	SABER+	Excess European
Brisbin et al. (2012)	Four Latino populations[a]	PCAdmix	Excess African in Colombian, Puerto Rican and Ecuadorian
Bhatia et al. (2014)	African Americans	RFmix	non-significant increase in African
Guan (2014)	Mexican	ELAI	Excess African
Rishishwar et al. (2015)	Colombian	SUPORTMix	Excess African
Zhou et al. (2016b)	Mexican	ELAI	Excess African
Deng et al. (2016)	Seven Latino populations[b]	Structure and Z-test	Excess African

[a]Dominican Republic, Colombian, Puerto Rican, Ecuadorian

[b]Mexican, Guatemalan, Costa Rican, Colombian, Chilean, Argentinean, Brazilian

in local ancestry for the MHC region is strong and recurrent, prompting us to consider both its possible biological basis as well as the likelihood of methodological artifacts.

A basic concern is whether local ancestry methods are biased by features of the MHC region (other than a true shift in ancestry proportions). For example, Price et al. (2008) pointed out that most deviations in ancestry reported by Tang et al. (2007a) (both within and outside the MHC region) were associated with regions of high linkage disequilibrium (LD). However, new methods for detecting local ancestry control for LD, but still detect an excess of African ancestry in the MHC (Guan 2014; Brisbin et al. 2012) (Table 3). An additional concern is that some ancestry inference methods require phased data, something that is challenging for the MHC, given the high polymorphism. However, ancestry results are consistent across methods that do (e.g., Brisbin et al. 2012) and do not (Guan 2014) require phased data, suggesting this is not the factor driving the findings.

Further problems for local ancestry estimation were raised by Pasaniuc et al. (2013), who found that loci with increased deviation in local ancestry show high polymorphism and increased rates of mendelian inconsistency. These authors also showed that inappropriate parental reference panels (e.g., distantly related from the true parental populations) can introduce errors in the analysis. This fact is of extreme relevance since samples from the true parental populations are not always available.

Further studies will be needed so as to evaluate whether technical artifacts underlie the shifts in ancestry proportions in the MHC region. In this sense, a promising result was reported by Deng et al. (2016), who used simulations under a human demographic model to show that the ancestry deviation in the MHC of Latin American populations is not expected in the absence of selection. In addition, Tang et al. (2007a) showed that an unusual African ancestry proportion in the MHC region of Puerto Rican individuals is found using local ancestry analysis based on SNPs, as well as more traditional admixture estimates using classical HLA markers and microsatellites, providing additional evidence that the shifts in ancestry are not a feature observed with one type of marker or inference method.

Ancestry deviations place the MHC as a striking example of a genomic region under strong recent selection. Nevertheless, even if this general picture is confirmed in new studies, several questions remain to be addressed. First, how many and which HLA alleles are favored by selection, causing the deviation in local ancestry? Second, is the recurrent finding of excess African ancestry explained by higher genomewide diversity in Africans (which indirectly could lead to the harboring of more advantageous variants)? Clearly, a biological understanding of these patterns is still lacking.

Selection favoring alleles of a specific ancestry can also be seen through the analysis of archaic genomes. These studies found evidence for adaptive introgression from archaic groups (Denisova and Neanderthal) into modern humans (reviewed in Racimo et al. 2015), including in the MHC region. Abi-Rached et al. (2011) suggested that a highly divergent allele, *HLA-vB*73*, entered the modern human gene pool through introgression from archaic hominins. In modern populations, *HLA-B*73* is practically absent everywhere except West Asia, and almost all haplotypes carrying *HLA-B*73* also carry *HLA-C*15:05*, which only reaches appreciable frequencies in Asia (Abi-Rached et al. 2011). Simulations showed that introgression from archaic hominins provides a better fit to the data than a model in which the allele arose in Africa before the Out-of-Africa event (Abi-Rached et al. 2011).

Yasukochi and Ohashi (2016) argue that this evidence is circumstancial, noting that *B*73* was not found in any archaic genome and that strong long-term balancing selection could maintain the alleles independently in both species. Also, if Denisova introgression into modern humans occurred in Southeast Asia, that is where *HLA-B*73* should have higher frequency.

On the other hand, Abi-Rached et al. (2011) found even more compelling evidence for adaptive introgression coming from the *HLA-A*11* allele, which occurs at high frequencies in Papua New Guinea and China (but is absent from Sub-Saharan African) and is found in long haplotypes with *HLA-C*15* and *HLA-C*12*, both of which exhibit higher diversity in Asia than in Africa. A likely explanation is that all *HLA-A*11* found in modern humans came from Denisovan introgression, followed by a rise in frequency in Asia. In brief, it may well be that when humans left Africa, they encountered new selective pressures to which archaic hominins were better adapted on a local scale, and strong selection favored those adaptive variants acquired through introgression. However, current evidence for adaptive introgression of HLA alleles should be interpreted with caution because of the technical difficulties in assessing variability of HLA genes, and small sample sizes of archaic species. Also, apparent introgression might result from incomplete lineage sorting, which is particularly likely in the MHC region, where long-term balancing selection results in trans-specific polymorphisms (Klein et al. 1993; Teixeira et al. 2015; Leffler et al. 2013).

From genome to transcriptome

While most studies on selection at HLA genes focus on peptide binding properties, expression levels are also important in determining phenotypes related to disease progression, both for infection and cancer (Blais et al. 2012; Thomas et al. 2012; Apps et al. 2013; Boegel et al. 2014). For example, high expression of *HLA-C* enhances an individual's ability to respond to HIV infection, whereas low expression confers protection against Crohn's disease (Blais et al. 2012; Apps et al. 2013). Additionally, expression varies broadly

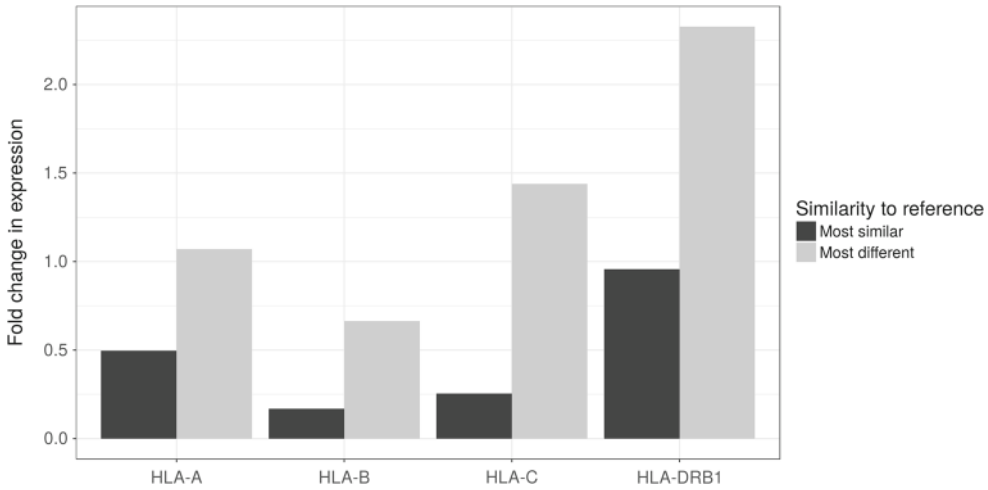

Fig. 5 Fold change in expression estimates obtained by kallisto (Bray et al. 2016) using a supplemented index relative to a standard reference index ($y = 0$). Results are presented for genotypes with different degrees of similarity to the reference genome (*bar colors*). We used 48 CEU individuals for which RNAseq data are available from the Geuvadis consortium (Lappalainen et al. 2013) and HLA genotypes were determined by Sanger sequencing (Gourraud et al. 2014). Genotypes at each locus were divided according to quartiles of differences from the reference allele at that locus. "Most similar" and "Most different" correspond to the first and fourth quartiles respectively (12 individuals each).

among tumor types, ranging from loss/downregulation to high expression (Boegel et al. 2014). Such opposing effects of expression levels may account for the selective maintenance of differential expression across HLA alleles or haplotypes.

Despite the potential importance of HLA expression to evolutionary and medical studies, few datasets with this information have been generated. To a large degree, this results from the the difficulty in quantifying expression for genes which show an unusually high polymorphism and are members of a multi-gene family. For highly polymorphic genes, array-based expression requires probes that avoid polymorphic regions, which if not accounted for can cause differential binding due to genetic variation, biasing expression estimates. The same difficulty applies to quantitative PCR, which needs primers that can bind the entire range of alleles of a specific locus, posing an important challenge when developing the experimental design.

To overcome these difficulties, customized arrays (Vandiedonck et al. 2011) and qPCR primer sets (Ramsuran et al. 2015) have been developed. These account for polymorphism and can provide locus-level expression estimates. However, these studies are limited in the number of samples and population diversity surveyed, and the requirement of custom arrays or primer sets makes repetition of surveys on additional populations and extension to other HLA loci challenging. Further, the expression of each allele cannot be directly estimated, and is instead imputed from the locus-level expression of homozygotes (Ramsuran et al. 2015). This places the quantification of HLA expression as an enterprise still in its infancy, although the studies carried out to date show that HLA expression varies between alleles, loci, and tissues (Boegel et al. 2012; Boegel et al. 2014; Ramsuran et al. 2015; Melé et al. 2015).

The RNAseq technology, which quantifies expression using NGS, is increasingly being used in genomewide studies and has the potential to provide large-scale information on HLA expression, but also has challenges. The technology relies on the mapping of short reads (generated by sequencing the transcriptome) to an index, so as to quantify the abundance of mRNA originating from each gene or exon. In the event that the surveyed individual is highly divergent from the sequences in the index (as is often the case due to the high polymorphism of HLA genes), it is likely that many reads will be discarded due to large numbers of mismatches, failing to document expression, and biasing the estimates toward the overexpression of variants which are more similar to the one in the index. This results in inaccurate and/or biased gene expression estimates and can cause spurious eQTLs to be identified (Panousis et al. 2014). This problem is similar to that of read mapping for HLA genes in NGS, discussed in Section 3 (Brandt et al. 2015).

As a consequence, large studies which surveyed the whole-transcriptome in many individuals (e.g., Lappalainen et al. 2013; Battle et al. 2014) using high-throughput technologies do not provide reliable estimates for the expression of HLA genes. An alternative is the development of bioinformatic tools that use whole-transcriptome RNAseq data to accurately estimate HLA expression. This has the benefit of placing the HLA expression data within the context of genomewide expression levels, and allows the use of RNAseq datasets that are already available (Lappalainen et al. 2013; Battle et al. 2014; Melé et al. 2015).

A promising approach is to use of an index with thousands of HLA sequences reported in databases such as IPD-IMGT/HLA, instead of relying on a single reference genome. For example, *seq2HLA* is a pipeline proposed by Boegel et al. (2012)

which uses a form of *in silico* genotyping to both infer the genotypes at HLA genes as well as estimate the expression of each HLA allele at a locus. Such allele-specific estimates are not obtained when RNAseq data is processed by standard pipelines, which provide expression estimates at the level of genomic features such as annotated genes, exons or isoforms.

The work by Boegel et al. (2012) showed that the use of an appropriate index (i.e., the set of reference sequences to which the short reads generated by the NGS will be aligned) is the key element for the improvement in the estimates. The benefits of this approach are shown in Fig. 5: expression estimates increase when using indices supplemented with many HLA sequences, relative to expression estimated using the single reference genome. This effect is more pronounced for individuals carrying alleles which are most different from the reference. This is expected, since these are the cases where the use of the reference genome leads to the greatest underestimation of expression.

This result suggests that bioinformatic methods tailored to deal with HLA diversity can bring important changes to expression estimates and thus to eQTLs mapped, providing new hypotheses for functional elements which drive HLA expression variation. Promising candidates will include UTR sites, promoter/enhancer polymorphism, transcription factor binding sites, etc, all of which have been documented as enriched categories of eQTLs in standard genomewide studies (e.g., Lappalainen et al. 2013).

It will also be possible to further explore initial findings regarding expression differences among genes and alleles (revealed by qPCR studies). In particular, the pattern of relatively even expression among *HLA-B* alleles (Ramsuran et al. 2017), and variable expression levels among lineages at *HLA-A* (Ramsuran et al. 2015) and *HLA-C* (Apps et al. 2013) will be amenable to investigation on a wider scale.

Conclusions

Our current knowledge of HLA evolution differs with respect to that of a decade ago in many ways. To a large degree, this results from our ability to place HLA variation within the context of the entire genome. Genomewide studies have contributed to our understanding of selection by increasing the power of tests (thanks to the large number of samples and genetic markers) and by allowing variation from the entire genome to be used as a control for complicating factors, including population history. We now have evidence that selection on classical HLA genes extends beyond the heterozygote advantage model and has operated from ancient to very recent timescales (Albrechtsen et al. 2010; Field et al. 2016; Tang et al. 2007a; Mathieson et al. 2015).

By comparing genetic differentiation at HLA genes to that of the remainder of the genome, we have found instances of decreased differentiation (e.g., Hofer et al. 2012), as well as of increased differentiation (Bhatia et al. 2011). Such studies will help investigate which HLA variants represent adaptations to local selective pressures, and which are shared extensively at global scale, as an outcome of long-term balancing selection. We are now also able to investigate patterns of admixture in HLA genes (Tang et al. 2007a; Guan 2014), providing insights into the time frame and mode of selection that occurs when populations of different ancestries meet and interbreed.

We can increasingly test co-evolutionary hypotheses, such as the relation between KIR and HLA polymorphism (e.g., Single et al. 2007), and test hypotheses of epistatic interactions. Genomic data also allows us to test the effect of strong selection on HLA upon linked variants, a process which may be driving the accumulation of deleterious mutations near HLA genes (e.g., Lenz et al. 2016).

A whole new layer of information, namely expression levels, can be generated on a large scale, and integrated with information on genetic variation. This will contribute to association studies, by incorporating a key cellular phenotype—expression level—as a covariate. Such approaches will also help bring functional information to the investigation of HLA evolution (for example, in the form of allelic lineages (Bitarello et al. 2016) or supertype grouping (Francisco et al. 2015)).

Our perspective is that, increasingly, we will see the immunogenetics community working closely with researchers in genomics. Placing HLA within the genomic context is key to understanding HLA genes; complementarily, immunogenetics expertise will be key to interpreting genomewide studies, within which HLA genes are frequent and striking findings (be it in GWAS, selection, admixture or expression studies). In addition, lessons and challenges associated with studying a highly polymorphic region under intense balancing selection, as is the case for the MHC, can be carried over to the study of other genes or genomic regions under balancing selection (Leffler et al. 2013; Teixeira et al. 2015; DeGiorgio et al. 2014; Bitarello et al. 2017).

Acknowledgements We thank three anonymous reviewers for their helpful comments. We are grateful to Maria Luiza Petzl-Erler, Glenys Thomson, Danillo Augusto, Erick Castelli, Richard Single, and Cibele Masotti for their thoughtful reading and useful suggestions and criticisms. Scholarships provided by the São Paulo Research Foundation (FAPESP):# 2014/12123-2 (VRCA), #12/22796-9 (DYCB), #11/12500-2 (BDB), #12/09950-9, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) #1645581 (KN). FAPESP research grant 12/18010-0 (DM) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grant 305888/2015-3.

Appendix

BOX 1. HLA terminology

The vocabulary within the immunogenetic literature is often quite challenging to non-specialists. Here is a brief summary of key terms used in this review.

MHC region. The MHC (Major Histocompatibility Complex) genomic region, located on the short arm of chromosome 6 (6p21.3), which contains the HLA (Human Leukocyte Antigen) genes, as well as at least 200 other genes (approximately 40% of which are involved in some aspect of the immune response), spanning around 4 Mb (reviewed in Shiina et al, 2009). It is the most gene-dense region of the genome (with on average one gene every 16 kb) and, due to the presence of HLA loci, shows unusually high levels of polymorphism. It is also referred to as the HLA region, in humans.

Extended MHC region. The extended MHC comprises around 8 Mb of chromosome 6, containing over 250 protein-coding genes. This region expands the MHC borders to segments with high linkage disequilibrium with the MHC and contains additional genes involved in the immune response (Horton et al, 2004; Shiina et al, 2009).

HLA genes (or loci). **Classical HLA genes** are the highly polymorphic loci that code for the proteins that present peptides to the T-cell receptors (*HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1, -DRA1*), and distinguish themselves from the **Non-Classical genes**, which have reduced polymorphism and do not have a role in peptide presentation.

HLA allele. This refers to a specific DNA sequence at an HLA gene, and can be thought of as a haplotype of SNP variants. There are extraordinarily large numbers of HLA alleles, with more than 12,000 for class I alleles and more than 4,000 for class II. (see details in http://www.ebi.ac.uk/ipd/imgt/hla/stats.html).

HLA haplotype. The combination of alleles at several HLA genes on a single chromosome of a given individual.

HLA function. **HLA class I** molecules, expressed in most cells, bind peptides of intracellular origin and present them on the cell surface to the T-cell receptors of CD8+ T-cells, initiating a cytotoxic response if they are recognized as foreign. **HLA class II** molecules, expressed in antigen presenting cells, typically bind peptides of extracellular origin and present them to the TCRs of CD4+ cells, triggering a signaling process that leads to the multiplication of T-helper cells, leading to the stimulation of B-cells, which produce antibodies to the antigen that triggered the response.

BOX 2. Detecting genomic regions with signatures of natural selection

What are "neutrality tests"?

Since the proposition of the neutral theory of molecular evolution, a myriad of tests have been proposed for the null hypothesis of neutrality (i.e., that evolutionary change results exclusively from mutation and genetic drift). When the neutral model is rejected, plausible alternative hypotheses include positive or balancing selection. Because the null hypothesis is based on the neutral model of evolution, these tests are called 'neutrality tests'.

Tests based on differentiation among populations. If the intensity of positive selection at a locus varies among populations, an allele can become common exclusively in the population where it is advantageous. As a consequence, differentiation at this selected locus will be greater than for the rest of the genome (Lewontin and Krakauer, 1973). However, for loci under balancing selection, differentiation is expected to be reduced because variants will be kept at intermediate frequencies in all populations. Tests based on population differentiation typically use F_{ST} to compare differentiation at focal loci to simulated expectations or genomic controls (i.e., all other genes). These tests can detect relatively recent selection.

Tests based on linkage disequilibrium and diversity. Variants under strong positive selection rise in frequency rapidly, reducing diversity at neutral loci in their vicinity. For balancing selection, because two or more variants are maintained, local genetic variation increases in the region. These tests are powerful to detect recent and strong positive selection, and less powerful for long-term balancing selection.

Tests based on patterns of allelic frequency distribution (or "Site Frequency Spectrum", SFS). These tests compare the distribution of observed allelic frequencies to those expected under neutrality. These include Tajima's *D*, Fu and Li's *D*, Fu and Li's *F*, Fay and Wu's *H* tests, the T_2 test (DeGiorgio et al, 2014) and the *NCD* statistics (Bitarello et al, 2017). Positive selection is expected to increase the number of low frequency variants, whereas balancing selection increases the number of intermediate frequency variants around the selected site. These tests can detect selection at comparatively deeper timescales than those based on LD or F_{ST} and are influenced by demographic parameters. For example, a bottleneck will preferentially remove low frequency variants, thus shifting the SFS to more intermediate frequencies even in the absence of balancing selection (Gattepaille et al, 2013). A crucial point of SFS-based tests is that the expectations should be evaluated for a neutral scenario that includes demographic parameters, since various demographic trajectories can mimic the effects of selection (e.g., population structure results in many of the signatures expected under balancing selection).

Tests contrasting intra and inter-specific variation. These include the dN/dS (non-synonymous to synonymous substitution rates ratio), HKA and MK tests, as well as the T_1 and T_2 tests (DeGiorgio et al, 2014) and the *NCD2* statistic (Bitarello et al, 2017). While both positive and balancing selection are compatible with elevated dN/dS levels, they produce opposing signatures for HKA and MK tests (both of which compare polymorphism and divergence levels): positive selection causes lower polymorphism levels compared to divergence within the target species, balancing selection causes higher polymorphism relative to divergence. These tests are most powerful to detect long-term (> 1 million years) selection in humans.

Shared polymorphisms between species. If species divergence is not extremely recent and population sizes are not extremely large, we expect lineages to share their most recent common ancestor with other lineages from the same species. In the case of humans, the probability that two lineages do not coalesce before the split between humans and chimpanzees from their common ancestor, in the absence of selection, is on the order of 10^{-4} (Leffler et al, 2013). Considering the size of the human genome, one would still expect to find $\approx$ 100 of such sites. However, by also requiring the same to occur for chimpanzees, this sharing is extremely unlikely for neutrally evolving sites, thus providing the basis for a test of neutrality (Klein et al, 1993). If a polymorphism is maintained by balancing selection for a sufficiently long time, it may be found in two sister species, such as human and chimpanzee. This test is appropriate to detect ancient balancing selection, operating at a deeper timescale than the divergence between humans and the sister species (e.g. $\approx$ 6 myr for humans and chimpanzees).

"Outlier" vs "model-based" approaches

Commonly, summary statistics are calculated for genomic windows or biologically defined entities (such as genes). There are two main strategies that the statistic can be used to assess if this genomic region conforms to neutral expectations.

Simulation-based tests. Using simulations that assume neutral evolution, a null distribution is generated for the test statistic. Empirical data that are extreme with respect to this distribution are considered to reject the null hypothesis of neutrality. This approach relies on an appropriate demographic model, and allows the quantification of the proportion of the genome that has extreme signatures of selection compared to neutral expectations.

Empirical outliers. In this approach the genomic regions with the most extreme values for the chosen test-statistic are regarded as regions of interest, which are potentially under selection. While this approach allows the exploration of a few, very extreme, candidate regions/genes, it does not allow the quantification of the pervasiveness of extreme signatures in the genome.

BOX 3. Mechanisms of balancing selection

Balancing selection is a term that encompasses a broad range of selective regimes, all of which generate high levels of adaptive genetic variation. An important challenge in the study of HLA genes is teasing apart which of these forms of selection is operating.

Heterozygote advantage (or overdominance). Occurs whenever the fitness of the heterozygous genotype is higher than that of both homozygous genotypes. Under this scenario, two or more alleles can be maintained indefinitely in a population, eventually reaching a frequency equilibrium. Assuming genotype fitnesses are constant through time and that marginal frequencies are the same, a polymorphic equilibrium can be achieved, with both alleles being kept in a population. This has been proposed as one (non-exclusive) biologically plausible mechanism through which variants are maintained in the HLA class I and class II classical loci because heterozygote individuals would be able to present a vaster repertoire of antigens than homozygote individuals (Doherty and Zinkernagel, 1975).

Variable selection (over time and space). If selection coefficients (and hence the fitnesses of the genotypes) vary over time or space, a population may maintain levels of polymorphism which are greater than those expected under neutrality. The observed levels of diversity may be similar to those expected under heterozygote advantage, even if the heterozygote is not constantly the fittest genotype. Instead, the time or space-averaged fitness of heterozygotes must be higher than that of homozygotes (Gillespie, 2004).

Negative frequency-dependent selection. Under this selective regime the fitness of a genotype is inversely proportional to its frequency in the population. This model is biologically plausible, since pathogens are likely to evolve escape mutations to the most common HLA alleles. This reduces these HLA alleles' fitness, and thus their frequency. Relaxation of selection upon the pathogen may make this variant once again capable of conferring resistance, causing it to rise in frequency, thus driving a cycle which results in the maintenance of a polymorphic state (Spurgin and Richardson, 2010).

References

Abadie V, Sollid LM, Barreiro LB, Jabri B (2011) Integration of genetic and immunological insights into a model of celiac disease pathogenesis. Annu Rev Immunol 29:493–525. doi:10.1146/annurev-immunol-040210-092915

Abi-Rached L, Jobin MJ, Kulkarni S et al (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334(6052):89–94. doi:10.1126/science.1209202

Ahlenstiel G, Martin MP, Gao X, Carrington M, Rehermann B (2008) Distinct KIR/HLA compound genotypes affect the kinetics of human antiviral natural killer cell responses. J Clin Invest 188(2):1017–1026. doi:10.1172/JCI32400

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19(5):711–722. doi:10.1101/gr.086652.108

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12(12):1805–1814. doi:10.1101/gr.631202

Albrechtsen A, Moltke I, Nielsen R (2010) Natural selection and the distribution of identity-by-descent in the human genome. Genetics 186(1):295–308. doi:10.1534/genetics.110.113977

Andrés AM, Hubisz MJ, Indap A et al (2009) Targets of balancing selection in the human genome. Mol Biol Evol 26(12):2755–2764. doi:10.1093/molbev/msp190

Apps R, Qi Y, Carlson JM et al (2013) Influence of HLA-c expression level on HIV control. Science 340(6128):87–91. doi:10.1126/science.1232685

Augusto DG, Petzl-Erler ML (2016) KIR And HLA under pressure: evidences of coevolution across worldwide populations. Hum Genet 134(9):929–940. doi:10.1007/s00439-015-1579-9

Augusto DG, O'Connor GM, Lobo-Alves SC et al (2015) Pemphigus is associated with KIR3DL2 expression levels and provides evidence that KIR3DL2 may bind HLA-a3 and a11 in vivo. Eur J Immunol 45(7):2052–2060. doi:10.1002/eji.201445324

de Bakker PI, McVean G, Sabeti PC et al (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 38(10):1166–1172. doi:10.1038/ng1885

Battle A, Mostafavi S, Zhu X et al (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res 24(1):14–24. doi:10.1101/gr.155192.113

Bauer DC, Zadoorian A, Wilson LO, The Melbourne Genomics Health Alliance, Thorne NP (2016) Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. Briefings in Bioinformatics Epub ahead of print:1–9. doi:10.1093/bib/bbw097

Begovich A, Moonsamy P, Mack S et al (2001) Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations. Tissue Antigens 57(5):424–439. doi:10.1034/j.1399-0039.2001.057005424.x

Beleza S, Santos AM, McEvoy B et al (2013) The timing of pigmentation lightening in Europeans. Mol Biol Evol 30(1):24–35. doi:10.1093/molbev/mss207

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. PLoS Genet 10(8):1–25. doi:10.1371/journal.pgen.1004412

Bhatia G, Pasaniuc B, Zaitlen N et al (2011) Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am J Hum Genet 89(3):368–381. doi:10.1016/j.ajhg.2011.07.025

Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: the impact of rare variants. Genome Res 23(9):1514–1521. doi:10.1101/gr.154831.113

Bhatia G, Tandon A, Patterson N et al (2014) Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. Am J Hum Genet 95(4):437–444. doi:10.1016/j.ajhg.2014.08.011

Bitarello BD, Francisco RdS, Meyer D (2016) Heterogeneity of dN/dS ratios at the classical HLA class I genes over divergence time and across the allelic phylogeny. J Mol Evol 82(1):38–50. doi:10.1007/s00239-015-9713-9

Bitarello BD, de Filippo C, Teixeira JC et al (2017) Signatures of long-term balancing selection in human genomes. BiorXiv doi:10.1101/119529

Blackwell JM, Jamieson SE, Burgner D (2009) HLA And infectious diseases. Clin Microbiol Rev 22(2):370–385. doi:10.1128/CMR.00048-08

Blais ME, Zhang Y, Rostron T et al (2012) High frequency of HIV mutations associated with HLA-c suggests enhanced HLA-c–restricted CTL selective pressure associated with an AIDS-protective polymorphism. J Immunol 188(9):4663–4670. doi:10.4049/jimmunol.1103472

Boegel S, Löwer M, Schäfer M et al (2012) HLA Typing from RNA-seq sequence reads. Genome Med 4(102):1–12. doi:10.1186/gm403

Boegel S, Löwer M, Bukur T, Sahin U, Castle JC (2014) A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. Oncoimmunology 3(8):e954,893. doi:10.4161/21624011.2014.954893

Borghans JA, Beltman JB, De Boer RJ (2004) MHC Polymorphism under host-pathogen coevolution. Immunogenetics 55(11):732–739. doi:10.1007/s00251-003-0630-5

Brandt DY (2015) Population differentiation at genes under strong balancing selection: a case study on the HLA genes. Master's, University of São Paulo. http://www.teses.usp.br/teses/disponiveis/41/41131/tde-25092015-104711/publico/Debora_Brandt_SIMPL.pdf

Brandt DY, Aguiar VR, Bitarello BD et al (2015) Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. G3: Genes—Genomes—Genetics 5(5):931–941. doi:10.1534/g3.114.015784

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal RNA-seq quantification. Nat Biotechnol 34(5):525–527. doi:10.1038/nbt.3519

Brisbin A, Bryc K, Byrnes J et al (2012) PCADmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol 84(4):343–364. doi:10.3378/027.084.0401

Cagliani R, Sironi M (2013) Pathogen-driven selection in the human genome. Int J Evol Biol 2013:1–6. doi:10.1155/2013/204240

Cao H, Wu J, Wang Y et al (2013) An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human MHC region using targeted high-throughput sequencing. PLoS ONE 8(7):1–9. doi:10.1371/journal.pone.0069388

Cao K, Moormann A, Lyke K et al (2004) Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens 63:293–325. doi:10.1111/j.0001-2815.2004.00192.x

Carapito R, Radosavljevic M, Bahram S (2016) Next-generation sequencing of the HLA locus: methods and impacts on HLA typing, population genetics and disease association studies. Human Immunology In Press doi:10.1016/j.humimm.2016.04.002

Carlson CS, Thomas DJ, Eberle MA et al (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15(11):1553–1565. doi:10.1101/gr.4326505

Castelli EC, Mendes-Junior CT, Sabbagh A et al (2015) HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. Hum Immunol 76(12):945–953. doi:10.1016/j.humimm.2015.06.016

Castelli EC, Gerasimou P, Paz MA et al (2017) HLA-G variability and haplotypes detected by massively parallel sequencing procedures in the geographicaly distinct population samples of Brazil and Cyprus. Mol Immunol 83:115–126. doi:10.1016/j.molimm.2017.01.020

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2(4):e64. doi:10.1371/journal.pgen.0020064

Chun S, Fay JC (2011) Evidence for hitchhiking of deleterious mutations within the human genome. PLoS Genet 7(8):e1002,240. doi:10.1371/journal.pgen.1002240

Coelho M, Luiselli D, Bertorelle G et al (2005) Microsatellite variation and evolution of human lactase persistence. Hum Genet 117(4):329–39. doi:10.1007/s00439-005-1322-z

Colonna V, Ayub Q, Chen Y et al (2014) Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol 15(6):R88. doi:10.1186/gb-2014-15-6-r88

Corona E, Dudley JT, Butte AJ (2010) Extreme evolutionary disparities seen in positive selection across seven complex diseases. PLoS ONE 5(8):1–10. doi:10.1371/journal.pone.0012236

Danzer M, Niklas N, Stabentheiner S et al (2013) Rapid, scalable and highly automated HLA genotyping using next-generation

sequencing: a transition from research to diagnostics. BMC Genomics 14(1):221. doi:10.1186/1471-2164-14-221

Daub JT, Hofer T, Cutivet E et al (2013) Evidence for polygenic adaptation to pathogens in the human genome. Mol Biol Evol 30(7):1544–1558. doi:10.1093/molbev/mst080

DeGiorgio M, Lohmueller KE, Nielsen R (2014) A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet 10(8):e1004,561. doi:10.1371/journal.pgen.1004561

Deng L, Ruiz-Linares A, Xu S, Sijia W (2016) Ancestry variation and footprints of natural selection along the genome in Latin American populations. Sci Rep 18(6):21,766. doi:10.1038/srep21766

Dilthey A, Cox C, Iqbal Z, Nelson MR, McVean G (2015) Improved genome inference in the MHC using a population reference graph. Nat Genet 47(6):682–688. doi:10.1038/ng.3257

Dilthey AT, Moutsianas L, Leslie S, McVean G (2011) HLA*IMP–An integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics 27(7):968–972. doi:10.1093/bioinformatics/btr061

Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice heterozygous at the h-2 gene complex. Nature 256(5512):50–52. doi:10.1038/256050a0

Erlich H (2012) HLA DNA Typing: past, present, and future. Tissue Antigens 80(1):1–11. doi:10.1111/j.1399-0039.2012.01881.x

Erlich RL, Jia X, Anderson S et al (2011) Next-generation sequencing for HLA typing of class I loci. BMC Genomics 12(42):1–13. doi:10.1186/1471-2164-12-42

Fellay J, Shianna KV, Ge D et al (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317(5840):944–947. doi:10.1126/science.1143767

Field Y, Boyle EA, Telis N et al (2016) Detection of human adaptation during the past 2,000 years. Science. doi:10.1126/science.aag0776. http://science.sciencemag.org/content/early/2016/10/12/science.aag0776

Francisco RdS, Buhler S, Nunes JM et al (2015) HLA Supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-a and HLA-b polymorphisms. Immunogenetics 67(11-12):651–663. doi:10.1007/s00251-015-0875-9

Fraser HB (2013) Gene expression drives local adaptation in humans. Genome Res 23(7):1089–96. doi:10.1101/gr.152710.112

Fu W, Akey JM (2013) Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet 14(1):467–489. doi:10.1146/annurev-genom-091212-153509

Fu W, O'Connor TD, Jun G et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493(7431):216–220. doi:10.1038/nature11690

Fumagalli M, Sironi M, Pozzoli U et al (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet 7(11):e1002,355. doi:10.1371/journal.pgen.1002355

Garrigan D, Hedrick PW (2003) Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. Evolution 57(8):1707–1722. doi:10.1554/02-732

Gattepaille L, Jakobsson M, Blum M (2013) Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. Heredity 110(5):409–419. doi:10.1038/hdy.2012.120

Gendzekhadze K, Norman PJ, Abi-Rached L et al (2009) Co-evolution of KIR2DL3 with HLA-C in a human population retaining minimal essential diversity of KIR and HLA class I ligands. Proc Natl Acad Sci USA 106(44):18692–97

Gillespie JH (2004) Population genetics: a concise guide, 2nd edn. The Johns Hopkins University Press

Gourraud PA, Khankhanian P, Cereb N et al (2014) HLA Diversity in the 1000 genomes dataset. PLoS One 9(7):e97,282. doi:10.1371/journal.pone.0097282

Gregersen JW, Kranc KR, Ke X et al (2006) Functional epistasis on a common MHC haplotype associated with multiple sclerosis. Nature 443(7111):574–577. doi:10.1038/nature05133

Guan Y (2014) Detecting structure of haplotypes and local ancestry. Genetics 196(3):625–642. doi:10.1534/genetics.113.160697

Hedrick PW (2002) Pathogen resistance and genetic variation at MHC loci. Evolution 56(10):1902–1908. doi:10.1111/j.0014-3820.2002.tb00116.x

Hedrick PW (2006) Genetic polymorphism in heterogeneous environments: the age of genomics. Annu Rev Ecol Evol Syst 37(1):67–93. doi:10.1146/annurev.ecolsys.37.091305.110132

Hedrick PW, Thomson G (1983) Evidence for balancing selection at HLA. Genetics 104(3):449–456. http://www.genetics.org/content/104/3/449

Hedrick PW, Thomson G (1986) A two-locus neutrality test: applications to humans. E. coli and lodgepole pine. Genetics 112(1):135–156. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1202687/

Hedrick PW, Whittam TS, Parham P (1991) Heterozygosity at individual amino acid sites: extremely high levels for HLA-a and -B genes. Proc Natl Acad Sci U S A 88(13):5897–5901. doi:10.1073/pnas.88.13.5897

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S (2015) Estimating the mutation load in human genomes. Nat Rev Genet 16(6):333–343. doi:10.1038/nrg3931

Hiby S, Apps R, Chazara O et al (2014) Maternal KIR in combination with paternal HLA-c2 regulate human birth weight. J Immunol 192(11):5069–5073. doi:10.4049/jimmunol.1400577

Hiby SE, Walker JJ, O'Shaughnessy KM et al (2004) Combinations of maternal KIR and fetal HLA-c genes influence the risk of preeclampsia and reproductive success. J Exp Med 200(8):957–965. doi:10.1084/jem.20041214

Hilton HG, Norman PJ, Nemat-Gorgani N et al (2015) Loss and gain of natural killer cell receptor function in an african hunter-gatherer population. PLoS Genetics 11(8):1–19

Hofer T, Foll M, Excoffier L (2012) Evolutionary forces shaping genomic islands of population differentiation in humans. BMC Genomics 13(107):1–13. doi:10.1186/1471-2164-13-107

Hollenbach J, Thomson G, Cao K et al (2001) HLA Diversity, differentiation, and haplotype evolution in Mesoamerican atives. Hum Immunol 62(4HTC+01):378–90

Hollenbach JA, Augusto DG, Alaez C et al (2013) Report from the 16th international histocompatibility and immunogenetics workshop (IHIW) component: population global distribution of killer immunoglobulin-like receptor (KIR) and ligands. Int J Immunogenet 40(1):39–45. doi:10.1111/iji.12028

Horton R, Wilming L, Rand V et al (2004) Gene map of the extended human MHC. Nat Rev Genet 5(12):889–899. doi:10.1038/nrg1489

Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I (2013) Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics 14(355):1–16. doi:10.1186/1471-2164-14-355

Hosomichi K, Shiina T, Tajima A, Inoue I (2015) The impact of next-generation sequencing technologies on HLA research. J Hum Genet 60(11):665–673. doi:10.1038/jhg.2015.102

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335(6186):167–170. doi:10.1038/335167a0

Hunt KA, Mistry V, Bockett NA et al (2013) Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 498(7453):232–235. doi:10.1038/nature12170

Jia X, Han B, Onengut-Gumuscu S et al (2013) Imputing amino acid polymorphisms in human leukocyte antigens. PLoS ONE 8(6):e64,683. doi:10.1371/journal.pone.0064683

Johnson NA, Coram MA, Shriver MD et al (2011) Ancestral components of admixed genomes in a Mexican cohort. PLoS Genet 7(12):e1002,410. doi:10.1371/journal.pgen.1002410

Kamatani Y, Wattanapokayakit S, Ochi H et al (2009) A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. Nat Genet 41(5):591–595. doi:10.1038/ng.348

Key FM, Teixeira JaC, de Filippo C, Andrés AM (2014) Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev 29:45–51. doi:10.1016/j.gde.2014.08.001

Kirino Y, Bertsias G, Ishigatsubo Y et al (2013) Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. Nature Genetics 45(2):202–207. doi:10.1038/ng.2520

Klein J, Satta Y, O'hUigin C (1993) The molecular descent of the major histocompatibility complex. Annu Rev Immunol 11:269–295. doi:10.1146/annurev.iy.11.040193.001413

Klitz W, Hedrick P, Louis EJ (2012) New reservoirs of HLA alleles: pools of rare variants enhance immune defense. Trends Genet 28(10):480–486. doi:10.1016/j.tig.2012.06.007

Kulkarni S, Savan R, Qi Y et al (2011) Differential microRNA regulation of HLA-c expression and its association with HIV control. Nature 472(7344):495–8. doi:10.1038/nature09914

Langer V, Böhme I, Hofmann J et al (2014) Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics 15(63):1–11. doi:10.1186/1471-2164-15-63

Lank SM, Golbach BA, Creager HM et al (2012) Ultra-high resolution HLA genotyping and allele discovery by highly multiplexed cDNA amplicon pyrosequencing. BMC Genomics 13(1):378. doi:10.1186/1471-2164-13-378

Lappalainen T, Sammeth M, Friedländer MR et al (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501(7468):506–511. doi:10.1038/nature12531

Leffler EM, Gao Z, Pfeifer S et al (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339(6127):1578–1582. doi:10.1126/science.1234070

Lenz TL, Spirin V, Jordan DM, Sunyaev SR (2016) Excess of deleterious mutations around HLA genes reveals evolutionary cost of balancing selection. Mol Biol Evol 33(10):1–30. doi:10.1101/053793

Leslie S, Donnelly P, McVean G (2008) A statistical method for predicting classical HLA alleles from SNP data. J Hum Genet 82(1):48–56. doi:10.1016/j.ajhg.2007.09.001

Levin AM, Adrianto I, Datta I et al (2014) Performance of HLA allele prediction methods in African Americans for class II genes HLA-DRB1, -DQB1, and -DPB1. BMC Genet 15(1):72. doi:10.1186/1471-2156-15-72

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74(1):175–195. http://www.genetics.org/content/74/1/175

Lima TH, Buttura RV, Donadi EA et al (2016) HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample. Hum Immunol 77(10):841–853. doi:10.1016/j.humimm.2016.07.231

Lindo J, Huerta-Sánchez E, Nakagome S et al (2016) A time transect of exomes from a native american population before and after european contact. Nat Commun 7:13,175 EP. doi:10.1038/ncomms13175

Major E, Rigó K, Hague T, Bérces A, Juhos S (2013) HLA Typing from 1000 genomes whole genome and whole exome Illumina data. PLoS ONE 8(11):1–9. doi:10.1371/journal.pone.0078410

Maples BK, Gravel S, Kenny EE, Bustamante CD (2013) RFMIx: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet 93(2):278–288. doi:/10.1016/j.ajhg.2013.06.020

Mathieson I, Lazaridis I, Rohland N et al (2015) Genome-wide patterns of selection in 230 ancient Eurasians. Nat Genet 528 (7583):499–503. doi:10.1038/nature16152

Mayor NP, Robinson J, McWhinnie AJ et al (2015) HLA Typing for the next generation. PLoS ONE 10(5):1–12. doi:10.1371/journal.pone.0127153

Melé M, Ferreira PG, Reverter F et al (2015) The human transcriptome across tissues and individuals. Science 348(6235):660–665. doi:10.1126/science.aaa0355

Mendes FH (2013) Natural selection on HLA and its effects on adjacent regions of the genome. Master's, University of São Paulo. http://www.teses.usp.br/teses/disponiveis/41/41131/tde-02082013-161104/pt-br.php

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28(11):659–669. doi:10.1016/j.tree.2013.08.003

Meyer D, Thomson G (2001) How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet 65(1):1–26. doi:10.1046/j.1469-1809.2001.6510001.x

Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G (2006) Signatures of demographic history and natural selection in the human major histocompatibility complex loci. Genetics 173(4):2121–2142. doi:10.1534/genetics.105.052837

Monos D, Maiers MJ (2015) Progressing towards the complete and thorough characterization of the HLA genes by NGS (or single-molecule DNA sequencing): consequences, opportunities and challenges. Hum Immunol 76(12):883–886. doi:10.1016/j.humimm.2015.10.003

Nielsen R, Bustamante C, Clark AG et al (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3(6):e170. doi:10.1371/journal.pbio.0030170

Nishida N, Ohashi J, Khor SS et al (2016) Understanding of HLA-conferred susceptibility to chronic hepatitis B infection requires HLA genotyping-based association analysis. Sci Rep 19(6):24,767. doi:10.1038/srep24767

Norman PJ, Hollenbach JA, Nemat-Gorgani N et al (2016) Defining KIR and HLA class I genotypes at highest resolution via high-throughput sequencing. Am J Hum Genet 99(2):375–391. doi:10.1016/j.ajhg.2016.06.023

Novak AM, Hickey G, Garrison E et al (2017) Genome graphs. bioRxiv doi:10.1101/101378

Nunes K (2011) Native populations in South America: a multi-locus study of demographic and selective history. Phd thesis, University of São Paulo. http://www.teses.usp.br/teses/disponiveis/41/41131/tde-25042012-153528/pt-br.php

Nunes K, Zheng X, Torres M et al (2016) HLA Imputation in an admixed population: an assessment of the 1000 Genomes data as a training set. Hum Immunol 77(3):307–312. doi:10.1016/j.humimm.2015.11.004

van Oosterhout C (2009) A new theory of MHC evolution: beyond selection on the immune genes. Proc R Soc B 276(1657):657–665. doi:10.1098/rspb.2008.1299

Panousis NI, Gutierrez-Arcelus M, Dermitzakis ET, Lappalainen T (2014) Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. Genome Biol 15(467):1–8. doi:10.1186/s13059-014-0467-2

Parham P (2004) Killer cell immunoglobulin-like receptor diversity: balancing signals in the natural killer cell response. Immunol Lett 92(1–2):11–13. doi:10.1016/j.imlet.2003.11.016

Parham P, Moffett A (2013) Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. Nat Rev Immunol 13(2):133–144. doi:10.1038/nri3370

Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. Science (Washington D C) 272(5258PO96):67–74

Parham P, Norman PJ, Abi-Rached L, Guethlein LA (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 19(367):800–811. doi:10.1098/rstb.2011.0266

Pasaniuc B, Sankararaman S, Torgerson DG et al (2013) Analysis of Latino populations from GALA and MEC studies reveals

genomic loci with biased local ancestry estimation. Bioinformatics 29(11):1407–1415. doi:10.1093/bioinformatics/btt166
Penman BS, Ashby B, Buckee CO, Gupta S (2013) Pathogen selection drives nonoverlapping associations between HLA loci. PNAS 110(48):19,645–19,650. doi:10.1073/pnas.1304218110
Penman BS, Moffett A, Chazara O, Gupta S, Parham P (2016) Reproduction, infection and killer-cell immunoglobulin-like receptor haplotype evolution. Immunogenetics 68(10):755–764. doi:10.1007/s00251-016-0935-9
Penn DJ, Damjanovich K, Potts WK (2002) MHC Heterozygosity confers a selective advantage against multiple-strain infections. Proc Natl Acad Sci U S A 99(17):11,260–11,264. doi:10.1073/pnas.162006499
Price AL, Weale ME, Patterson N et al (2008) Long-range LD can confound genome scans in admixed populations. Am J Hum Genet 83(1):132–135. doi:10.1016/j.ajhg.2008.06.005
Prugnolle F, Manica A, Charpentier M et al (2005) Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15(11):1022–1027. doi:10.1016/j.cub.2005.04.050
Qian W, Deng L, Lu D, Xu S (2013) Genome-wide landscapes of human local adaptation in Asia. PLoS One 8(1):1–10. doi:10.1371/journal.pone.0054224
Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015) Evidence for archaic adaptive introgression in humans. Nat Rev Genet 16(6):359–371. doi:10.1038/nrg3936
Ramsuran V, Kulkarni S, O'huigin C et al (2015) Epigenetic regulation of differential HLA-a allelic expression levels. Hum Mol Genet 24(15):4268–4275. doi:10.1093/hmg/ddv158
Ramsuran V, Hernández-Sanchez PG, O'hUigin C et al (2017) Sequence and phylogenetic analysis of the untranslated promoter regions for HLA class I genes. J Immunol 198(6):2320–2329. doi:10.4049/jimmunol.1601679
Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105(3):767–79. http://www.genetics.org/content/105/3/767
Rishishwar L, Conley AC, Wigington CH et al (2015) Ancestry, admixture and fitness in Colombian genomes. Sci Rep 21(5):12,376. doi:10.1038/srep12376
Sabeti P, Schaffner S, Fry B et al (2006) Positive natural selection in the human lineage. Science 312(5780):1614–1620. doi:10.1126/science.1124309
Sams A, Hawks J (2014) Celiac disease as a model for the evolution of multifactorial disease in humans. Hum Biol 86(1):19–36. doi:10.3378/027.086.0102
Sanchez-Mazas A (2007) An apportionment of human HLA diversity. Tissue Antigens 69(s1):198–202. doi:10.1111/j.1399-0039.2006.00802.x
Sanchez-Mazas A, Meyer D (2014) The relevance of HLA sequencing in population genetics studies. J Immunol Res 2014:1–12. doi:10.1155/2014/971818
Schierup MH, Charlesworth D, Vekemans X (2000) The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. Genet Res 76(1):63–73. http://journals.cambridge.org/article_S0016672300004547
Seldin MF, Pasaniuc B, Price AL (2011) New approaches to disease mapping in admixed populations. Nat Rev Genet 12(8):523–528. doi:10.1038/nrg3002
Shiina T, Hosomichi K, Inoko H, Kulski JK (2009) The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 54(1):15–39. doi:10.1038/jhg.2008.5
Single RM, Martin MP, Gao X et al (2007) Global diversity and evidence for coevolution of KIR and HLA. Nat Genet 39(9):1114–1119. doi:10.1038/ng2077
Solberg O, Mack S, Lancaster A et al (2008) Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum Immunol 69(7):443–464. doi:10.1016/j.humimm.2008.05.001
Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc R Soc Lond B Biol Sci 277(1684):979–988. doi:10.1098/rspb.2009.2084
Sveinbjornsson G, Gudbjartsson DF, Halldorsson BV et al (2016) HLA Class II sequence variants influence tuberculosis risk in populations of European ancestry. Nat Genet 48(3):318–322. doi:10.1038/ng.3498
Tang H, Choudhry S, Mei R et al (2007a) Recent genetic selection in the ancestral admixture of Puerto Ricans. Am J Hum Genet 81(3):626–633. doi:10.1086/520769
Tang K, Thornton KR, Stoneking M (2007b) A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol 5(7):e171. doi:10.1371/journal.pbio.0050171
Teixeira JC, de Filippo C, Weihmann A et al (2015) Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees and bonobos. Mol Biol Evol 32(5):1186–1196. doi:10.1093/molbev/msv007
The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073. doi:10.1038/nature09534
The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526(7571):68–74. doi:10.1038/nature15393
Thomas R, Thio CL, Apps R et al (2012) A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection. J Virol 86(12):6979–6985. doi:10.1128/JVI.00406-12
Tian C, Hinds DA, Hromatka BS et al (2016) Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. biorxiv doi:10.1101/073056
Trowsdale J, Knight JC (2013) Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet 14:301–323. doi:10.1146/annurev-genom-091212-153455
Trowsdale J, Moffett A (2008) NK Receptor interactions with MHC class I molecules in pregnancy. Semin Immunol 20(6):317–320. doi:10.1016/j.smim.2008.06.002
Trowsdale J, Barten R, Haude A et al (2001) The genomic context of natural killer receptor extended gene families. Immunol Rev 181(1):20–38. doi:10.1034/j.1600-065X.2001.1810102.x
Vandiedonck C, Taylor MS, Lockstone HE et al (2011) Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. Genome Res 21(7):1042–1054. doi:10.1101/gr.116681.110
Wang C, Krishnakumara S, Wilhelmya J et al (2012) High-throughput, high-fidelity HLA genotyping with deepsequencing. Proc Natl Acad Sci U S A 109(22):8676–8681. doi:10.1073/pnas.1206614109
Winkler CA, Nelson GW, Smith MW (2010) Admixture mapping comes of age. Annu Rev Genomics Hum Genet 11:65–89. doi:10.1146/annurev-genom-082509-141523
Yasukochi Y, Ohashi J (2016) Elucidating the origin of HLA-b*73 allelic lineage: did modern humans benefit by archaic introgression? Immunogenetics Epub ahead of print:1–5. doi:10.1007/s00251-016-0952-8
Yawata M, Yawata N, Draghi M et al (2008) MHC Class I-specific inhibitory receptors and their ligands structure diverse human NK-cell repertoires toward a balance of missing self-response. Blood 112(6):2369–2380. doi:10.1182/blood-2008-03-143727
Yi X, Liang Y, Huerta-Sanchez E et al (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329(5987):75–8. doi:10.1126/science.1190371

Zhang FRR, Huang W, Chen SMM et al (2009) Genomewide association study of leprosy. N Engl J Med 361(27):2609–2618. doi:10.1056/NEJMoa0903753

Zheng X, Shen J, Cox C et al (2013) HIBAG-HLA Genotype imputation with attribute bagging. Pharmacogenomics J 14(2):192–200. doi:10.1038/tpj.2013.18

Zhou F, Cao H, Zuo X et al (2016a) Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. Nat Genet 48(7):740–746. doi:10.1038/ng.3576

Zhou Q, Zhao L, Guan Y (2016b) Strong Selection at MHC in Mexicans since admixture. PLoS Genet 10(12):e1005,847. doi:10.1371/journal.pgen.1005847