Population Differentiation at the *HLA* Genes

Débora Y. C. Brandt[*,§], **Jônatas César**[*], **Jérôme Goudet**[†,‡,1] **and Diogo Meyer**[*,1]

[*]Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo, SP, Brazil, [§]Current address: Department of Integrative Biology, University of California, Berkeley, USA, [†]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, [‡]Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland

ABSTRACT Balancing selection is defined as a class of selective regimes that maintain polymorphism above what is expected under neutrality. Theory predicts that balancing selection reduces population differentiation, as measured by F_{ST}. However, balancing selection regimes in which different sets of alleles are maintained in different populations could increase population differentiation. To tackle this issue, we investigated population differentiation at the HLA genes, which constitute the most striking example of balancing selection in humans. We found that population differentiation of single nucleotide polymorphisms (SNPs) at the HLA genes is on average lower than that of SNPs in other genomic regions. However, this result depends on accounting for the differences in allele frequency between selected and putatively neutral sites. Our finding of reduced differentiation at SNPs within HLA genes suggests a predominant role of shared selective pressures among populations at a global scale. However, in pairs of closely related populations, where genome-wide differentiation is low, differentiation at HLA is higher than in other genomic regions. This pattern was reproduced in simulations of overdominant selection. We conclude that population differentiation at the HLA genes is generally lower than genome-wide, but it may be higher for recently diverged population pairs, and that this pattern can be explained by a simple overdominance regime.

KEYWORDS HLA; population differentiation; balancing selection

Natural selection is one of the forces shaping the genetic variation within and the differentiation between populations. In the case of a locus where a variant is favored in one population but not in another (*i.e.*, in which selection drives local adaptation), we expect differentiation to exceed that under purely demographic processes (Lewontin and Krakauer 1973). This is the case for well known examples, such as the regulatory variant that generates lactase persistence in adulthood, which is more frequent in Europeans (Bersaglieri *et al.* 2004) and variants of the *EPAS* gene that provide adaptation to high altitude in Tibetans (Xu *et al.* 2011). Purifying selection, on the other hand, is more common and removes most variants which would contribute to differences among populations. Therefore, it is expected to reduce genetic differentiation at the focal locus with respect to a strictly neutral scenario (*e.g.* Barreiro *et al.* (2008)), while differentiation in surrounding genomic regions may increase due to the lower effective population size (Charlesworth *et al.* 1997).

A third regime, balancing selection, is related to diversity and differentiation in more complex ways. By definition balancing selection encompasses all selective regimes that result in increased genetic diversity with respect to neutral expectations. The increased variability can result from a variety of processes, often with different underlying biological properties: frequency dependent selection, heterozygote advantage, selection varying over temporal and geographic scales (Andrés 2011). As a consequence, the expectations regarding population differentiation under balancing selection represent a challenging theoretical and empirical question.

Across human populations, the loci with the strongest evidence for balancing selection are the classical HLA class I and II loci (especially the *HLA-A*, *-B*, *-C*, *-DRB1*, *-DQB1* and *-DQA1* loci), which are the human Major Histocompatibility Complex (MHC) genes. These genes encode proteins that mediate a critical step of the adaptive immune response, which is the binding of peptides for presentation on the surface of the cellular membrane. The HLA-peptide complex is surveyed by T-cell receptors, which may trigger an immune response when a non-self peptide is identified (Klein and Sato 2000). Balancing selection at HLA loci has been strongly supported by a wide variety of methods, with evidence including an excess of alleles at intermediate frequency with respect to neutral expectations (Hedrick and Thomson 1983), higher non-synonymous to synonymous

[1]Diogo Meyer: Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, 05508-090; Jérôme Goudet: DEE, Biophore, Unil-Sorge, CH-1015 Lausanne

substitution rate (Hughes and Nei 1988) and trans-specific polymorphism (Lawlor *et al.* 1988) (Meyer and Thomson 2001, for a review).

Although balancing selection at HLA genes is well documented, the evidence from most studies is compatible with different mechanisms that are difficult to disentangle: heterozygote advantage (Doherty and Zinkernagel 1975; Takahata and Nei 1990; De Boer *et al.* 2004), frequency dependent selection (Slade and McCallum 1992; Borghans *et al.* 2004) and selection that varies over time and space (Eizaguirre *et al.* 2012) have all been proposed to act on the HLA genes.

Interestingly, these selective regimes are theoretically compatible with both increased or reduced population differentiation. For example, the coevolution between HLA and pathogens could create a mechanism of frequency dependent selection, or rare allele advantage. Under this scenario, rare HLA alleles would be advantageous, since few pathogens would have evolved resistance to them (Meyer and Thomson 2001). Rare allele advantage is expected to increase the effective migration rate: migrants will often be rare in the population they arrive to, and thus will be advantageous and increase in frequency in the new population (Schierup *et al.* 2000; Muirhead 2001). Therefore, this regime of balancing selection is expected to reduce population differentiation.

However, in the case of HLA genes, balancing selection may be population-specific, with the different sets of pathogens in each population selecting locally advantageous HLA variants. Under this scenario we expect an increase in population differentiation. Evidence in support of population-specific pathogen selection for humans comes from the finding that HLA and pathogen diversities across populations are correlated (Prugnolle *et al.* 2005), and from theoretical studies showing that population-specific pathogen selection models of balancing selection provide a better explanation for observed HLA variation than heterozygote advantage (Hedrick 2002; Borghans *et al.* 2004).

Pathogen-driven selection implies that specific HLA alleles are more effective in presenting antigens of certain pathogens, to which the population is exposed. Support for this assumption comes from associations between disease susceptibility, resistance or progression with genetic variation at HLA. For example, variants in *HLA-B* are associated to the progression to clinical disease after HIV infection (The International HIV Controllers Study 2010), variants in *HLA-C*, *-DP* and *-DQ* are associated to clearance of hepatitis B virus (Hu *et al.* 2013), alleles in *HLA-DRB1* are associated to susceptibility or resistance to leprosy (Vanderborght *et al.* 2007) and alleles at *HLA-B*, *-DQ* and *-DR* are associated to resistance to severe malaria (Hill *et al.* 1991).

Previous studies which measured population differentiation at the HLA genes found evidence for both increased or reduced population differentiation. For example, Meyer *et al.* (2006) found no significant difference between differentiation at HLA genes and a set of neutral markers used as a control (microsatellites), while Sanchez-Mazas (2007) found lower differentiation at HLA loci than in their genome-wide control (microsatellites and RFLPs). A limitation of these studies is that differences between the neutral genetic markers and the sequence data used for HLA genes introduce confounding variables, making it difficult to determine the roles of selection or characteristics inherent to the marker (*e.g.* mutation rate and diversity). Another study compared differentiation on markers of the same type (microsatellites) located at HLA genes or near them and those located in other genomic regions, which serve as controls (Nunes 2011). This study found increased differentiation in regions near HLA genes. Nonetheless, some issues remain unresolved: Nunes (2011) was mainly interested in native American populations, and used a limited number of markers. Furthermore, the complexity of the mutational mechanism of microsatellites complicates the interpretation of results.

For non-model organisms a similarly wide array of results have been found, with the MHC region (which contains genes homologous to HLA) showing either equal (Miller *et al.* 2010), higher (Loiseau *et al.* 2009; Oliver *et al.* 2009; Cammen *et al.* 2011) or lower (McCairns *et al.* 2011) differentiation than genome-wide averages. These contrasting results could be due to differences in selective regimes among species, or even to variation in selection among genes within a species (Čížková *et al.* 2011).

In summary, it remains unclear whether balancing selection on HLA genes drives increased differentiation due to selection favoring adaptation to locally occurring pathogens, or whether it results in decreased genetic differentiation due to the maintenance of shared polymorphisms among populations.

Here, we revisit the question of population differentiation at the HLA genes through analysis of data from worldwide human populations. We analyze variation at SNPs, which have the advantage of allowing the use of genomic data as an empirical control for HLA SNPs, assuming similar mutation rates for SNPs in the MHC region and the reminder of the genome. Differently from scans that seek genome-wide significance for specific SNPs, we *a priori* define a set of putatively selected SNPs to be surveyed (those within or close to HLA genes). We relate differences in F_{ST} between HLA and non-HLA SNPs to the degree of polymorphism in each of these groups, drawing on recent findings concerning the constraints imposed by allele frequencies on measures of differentiation. Finally, we perform simulations and find a plausible selective regime that reproduces our results.

Materials and Methods

SNP Data

SNP genotypes were acquired from the integrated Variant Call Format (VCF) files from phase 3 of the 1000 Genomes Project (1000G) (The 1000 Genomes Project Consortium 2015), which are available at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/.

This dataset includes variants discovered via high coverage exome targeted resequencing and low coverage whole genome resequencing. Because of the higher coverage in exonic regions, all comparisons between HLA SNPs and non-HLA SNPs (which we treated as a control set) were made within the same functional category (*e.g.* intronic or exonic). To this end each SNP was annotated using ANNOVAR (Wang *et al.* 2010). Our findings for differences in F_{ST} between HLA and non-HLA SNPs were qualitatively the same when using either exonic or intronic regions, so throughout the paper we focus on the results for exonic regions.

The 1000G Phase 3 data contains the genomes of 2504 individuals from 26 populations. After applying filters described in the Extended Materials and Methods section, a total of 2000 individuals in 20 populations were kept (see Table S1).

Estimation of F_{ST}

Population differentiation was calculated as the proportion of variance in allele frequencies among populations (a), relative to

the total genetic variance ($a + b + c$, with b and c referring to the variance components between individuals within populations and between gametes within individuals, respectively):

$$F_{ST} = \frac{a}{a + b + c} \quad (1)$$

To obtain F_{ST} values we used the Weir and Cockerham (1984) estimator implemented in VCFtools v0.1.14 (Danecek *et al.* 2011). The Weir and Cockerham (1984) estimator was chosen because it is unbiased when sample sizes are large and similar, as in the case of our dataset. F_{ST} was calculated per SNP i) over all populations and ii) for pairs of populations.

When summarizing F_{ST} over multiple SNPs, we compared two approaches: i) computing a simple average of F_{ST} at individual SNPs and ii) using the "ratio of averages" approach, suggested by Reynolds *et al.* (1983), in which we first estimate the numerator (a) and denominator ($a + b + c$) of F_{ST} for each SNP, and then compute the averages of a and $a + b + c$ for the desired set of SNPs, and finally compute the ratio of both averages. This second approach provides the least biased estimate of F_{ST}, whereas performing a simple average of the F_{ST}s of each SNP can lead to an underestimation of differentiation, especially in datasets rich in rare variants (Reynolds *et al.* 1983; Bhatia *et al.* 2013). Unless otherwise stated, we used the "ratio of averages" approach to compute F_{ST}. Variance components (a and $a + b + c$) were obtained using a minor modification of the VCFtools source code.

Definition of HLA, peri-HLA and control regions

We define "HLA SNPs" as those contained within the coding sequence of the classical HLA genes *HLA-A*, *-B*, *-C*, *-DRA*, *-DRB1*, *-DQA*, *-DQB1*. Previous studies of *HLA-DPA1* and *-DPB1* found weak or no evidence of balancing selection (Solberg *et al.* 2008; Begovich *et al.* 2001), and even instances of directional selection (Hollenbach *et al.* 2001), making them inappropriate for our question concerning the role of balancing selection on differentiation at HLA loci. Accordingly, our analyses also showed that population differentiation at *HLA-DPA1* and *HLA-DPB1* genes is different from that of other HLA genes (see Results and Figure S1). Therefore, those loci were excluded from our analysis unless otherwise mentioned.

Peri-HLA genes were defined as those that flank the HLA genes and have higher diversity relative to the average of chromosome 6 (Mendes 2013), indicating that their increased polymorphism is driven by hitchhiking to the strongly selected HLA loci (Table S2). These genes are located 119kb to 256kb from the closest HLA locus. All SNPs outside both the HLA and peri-HLA genes comprised the control group.

Controlling for unreliable allele frequency estimates

The use of a single reference genome to map Next Generation Sequencing (NGS) reads creates mapping bias at some HLA SNPs in the 1000 Genomes Project phase 1 dataset (Brandt *et al.* 2015). We therefore excluded the SNPs within the HLA genes which have unreliable frequency estimates in the 1000 Genomes phase 1 dataset, due to mapping bias (Brandt *et al.* 2015). After applying this filter, 38 out of 525 exonic biallelic SNPs in the HLA genes were excluded (Table S3). In total, 487 SNPs were kept in the HLA group, 1193 in the peri HLA group, and 831,174 in the control group.

In the present study we analyze this filtered version of the 1000 Genomes phase 3 data instead of phase 1 or Sanger sequencing data analyzed in Brandt *et al.* (2015) because it has a larger sample size and includes more populations. Also, SNPs identified as unreliable in the 1000 Genomes phase 1 data were consistent among different populations, which supports the application of this filter to phase 3 data used here.

Statistical test of F_{ST} differences

When comparing F_{ST} values between HLA and control SNPs, we control for SNPs being located within a small set of genes, resulting in higher linkage disequilibrium (LD) and statistical non-independence than for SNPs in the control group, which comprise a genomewide set. To account for this effect, we designed a strategy to sample the control SNPs so as to approximate the LD structure among the HLA SNPs.

This was done by sampling a random exonic SNP from outside the HLA and peri-HLA genes and selecting all other exonic SNPs from the same gene (resulting in a high LD set of SNPs). We repeated this operation until we obtained a total number of SNPs that matched that of the HLA SNPs, for each MAF bin (see Figure S2). We sampled 1000 such sets of LD-matched control SNPs and compared the F_{ST} distribution of each of those to that of the HLA SNPs, applying a Mann-Whitney test. We recorded the number of comparisons where the difference between F_{ST} distributions was significant ($p < 0.05$).

Haplotype level analyses

In addition to SNP based analyses, we also investigated population differentiation when alleles are defined by the coding sequence of each HLA gene (classically referred to as an "HLA allele"). We treat these analyses as "haplotype level", where haplotypes are defined by a combination of SNPs along an HLA gene (i.e., each allele in these analyses is an intragenic haplotype).

Phasing of SNPs in extremely variable regions like the HLA is very challenging, due to the high SNP density and polymorphism. Therefore, rather than estimate intragenic haplotypes directly from the SNP data, we used a publicly available dataset which provides HLA allele calls for samples in the 1000G data based on Sanger sequencing Gourraud *et al.* (2014), available at the dbMHC website (http://www.ncbi.nlm.nih.gov/gv/mhc/xslcgi.fcgi?cmd=cellsearch).

We restricted these haplotype level analyses to *HLA-A*, *-B*, *-C*, *HLA-DRB1* and *-DQB1*, which are reported in Gourraud *et al.* (2014). HLA allele calls were coded so as to only distinguish alleles with nonsynonymous differences (*i.e.*, only the first two fields of the allele names were used, as described in the HLA nomenclature system) (Marsh *et al.* (2010)).

F_{ST} values of multiallelic HLA haplotypes can't be directly compared to those of biallelic SNPs because multiallelic loci tend to have lower allele frequencies, which constrains the maximum value of F_{ST} (Jakobsson *et al.* 2013). To allow F_{ST} values at the HLA haplotypes to be compared to this null distribution, we recoded each HLA gene as a series of biallelic loci. This recoding was done by treating each allele at each gene as "allele 1", and all other alleles as "allele 2". The Weir and Cockerham (1984) F_{ST} estimator was then computed as described for SNPs, using the *wc* function of the hierfstat R package (Goudet 2005).

Results

Higher F_{ST} in HLA genes

Initially we compared the distribution of F_{ST} among SNPs from the HLA, peri-HLA and control groups (Figure 1).

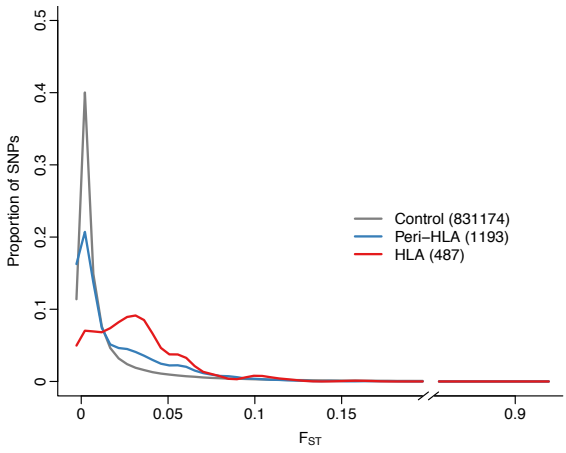

Figure 1 Distribution of F_{ST} values for exonic SNPs from the following groups: outside HLA and peri-HLA regions (control group), within HLA genes and within peri-HLA genes. Number of SNPs in each group is shown in parentheses.

The distribution of F_{ST} of SNPs within HLA genes is shifted towards higher values (median=0.027) compared to control SNPs (median=$9.2 * 10^{-5}$), and the distributions are significantly different (p-value $< 10^{-16}$, two-tailed Mann-Whitney test).

Theory predicts that balancing selection affects only a narrow genomic region, the size of which is defined by the intensity of selection and the recombination rate (Charlesworth *et al.* 1997). To evaluate if the increased differentiation at coding SNPs within HLA genes was also observed in loci that flank the HLA, we applied the same test to the peri-HLA genes. As is the case for HLA SNPs, peri-HLA SNPs have an F_{ST} distribution which is significantly shifted to higher values (median = 0.006, p-value $< 10^{-16}$, two-tailed Mann-Whitney test; Figure 1).

Lower F_{ST} in HLA when accounting for MAF

The effect of balancing selection is to shift the site frequency spectrum (SFS) of selected loci toward an excess of intermediate frequency variants. This is precisely what we see in the data, with the SFS for HLA SNPs showing a shift to intermediate frequencies compared to control SNPs. The peri-HLA SNPs occupy an intermediate position in the SFS (Figure 2).

Constraints imposed by allele frequencies on F_{ST} have been a topic of recent investigation (Roesti *et al.* 2012; Maruki *et al.* 2012; Elhaik 2012; Jakobsson *et al.* 2013; Edge and Rosenberg 2014; Alcala and Rosenberg 2017), and it has been shown that SNPs with very low minor allele frequencies (MAF) are bounded to low F_{ST} values. This relationship between MAF and F_{ST} is empirically illustrated for the 1000 Genomes exome data in Figure 3, which shows that F_{ST} is constrained to low values mainly in the range of low MAF (up to $\sim$ 0.1 in the 1000G dataset). When MAF is above this value the constraint is no longer evident. In File S1, we analytically show the relationship between MAF and the maximum possible value of F_{ST}.

This suggests that the large number of rare variants in the 1000G dataset, and the relative paucity of low MAF variants in the HLA SNPs could generate the observed differences in F_{ST} distributions between those groups (Figure 1).

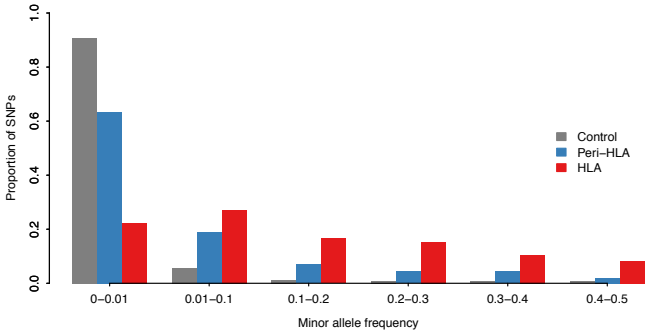

Figure 2 Distribution of minor allele frequency (MAF) for exonic control, peri-HLA genes and HLA genes. MAF of SNPs at the HLA and peri-HLA genes is higher compared to other genes.

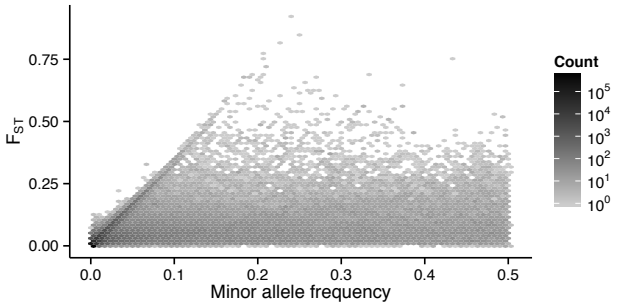

Figure 3 Population differentiation, measured by F_{ST}, as a function of minor allele frequency at biallelic exonic SNPs from the 1000 Genomes Project phase 3 data.

To account for the effects of differences in SFS between HLA and control SNPs when contrasting population differentiation among those groups, we compared the F_{ST} of HLA and control SNPs within bins of MAF values (Figure 4). Contrary to what we observed without controlling for MAF, we now find that HLA and peri-HLA SNPs have significantly lower F_{ST} than at the control SNPs (Mann-Whitney two-tailed test p-value $< 10^{-5}$ for all bins of MAF > 0.01). The bin with MAF < 0.01 shows a similar pattern when further split into smaller bins of MAF (Figure S3). Figure S4 shows F_{ST} distributions including outliers.

This approach used SNPs outside the HLA and peri-HLA genes as controls for the HLA SNPs. However, because the HLA SNPs are located in 5 genes, they are not independent, due to both intra and inter-locus associations. As a consequence, our p-values could be inflated by treating a set of correlated SNPs as independent, and comparing them to a set of control SNPs which are in their majority independent. We controlled for this non-independence using a resampling approach, in which sets of linked SNPs were sampled from our control group, and their F_{ST} values were compared to those from the HLA SNPs, at each MAF bin (see Materials and Methods). After controlling for the non-independence of HLA SNPs, we confirmed that F_{ST} at HLA SNPs was significantly lower than at the resampled SNPs, at all bins of MAF higher than 0.01 (Table S4).

Another strategy to control for the constraint of MAF on F_{ST} is to estimate F_{ST} for multiple loci by computing a "ratio of the

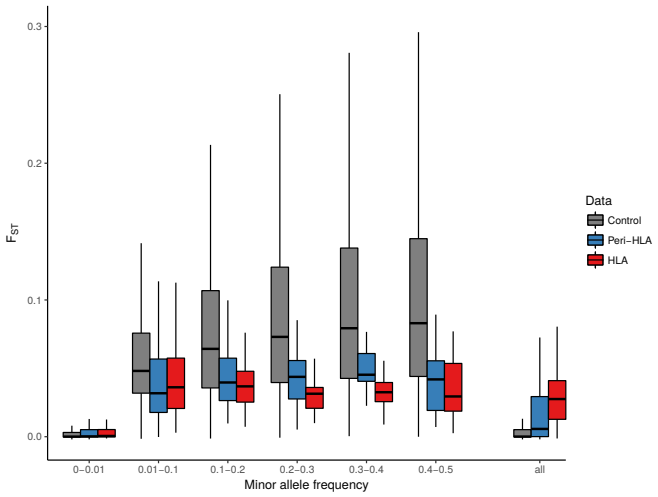

Figure 4 F_{ST} distributions per minor allele frequency (MAF) bin. HLA and Peri-HLA SNPs show lower F_{ST} than control SNPs in all bins with MAF $>$ 0.01. Outliers were removed from figure, but not from statistical test, for better visualization. Figure S4 shows F_{ST} distributions including outliers.

averages" of the numerator and denominator of the AMOVA-based F_{ST} estimator (see Methods). Reynolds *et al.* (1983) showed that this is the least biased estimate of average F_{ST} across multiple loci. The alternative, *i.e.*, computing the average F_{ST}s for individual SNPs (an "average of ratios") creates a bias leading to an underestimation of F_{ST}, the effect of the bias being more pronounced the more rare variants there is in the dataset (Bhatia *et al.* 2013). The ratio of averages approach, on the other hand, proportionally downweights the contribution of variants with low MAF. This results in a higher overall F_{ST} values for datasets rich in rare variants.

We explored how these different averaging methods impact the F_{ST} at HLA genes, and found that the "ratio of averages" approach (which controls for MAF) results in lower average F_{ST} at the HLA SNPs (0.04) than in the control SNPs (0.09). In stark contrast, using the average of individual loci F_{ST}, we found higher F_{ST} values for the HLA SNPs (0.03) than genome-wide (0.01), as in our initial analysis that did not account for MAF. This further emphasizes the importance of accounting for the differences between the site frequency spectrum of HLA and control SNPs when assessing population differentiation.

HLA-DP genes The classical *HLA-DPA1* and *-DPB1* genes were excluded from the previous analysis because they show weak or no evidence of balancing selection (Solberg *et al.* 2008; Begovich *et al.* 2001), and some evidence of directional selection (supported by the observation that within individual populations a small number of alleles are present at a high frequency) (Hollenbach *et al.* 2001).

Consistently with being under a different selective regime, *HLA-DPA1* and *-DPB1* show a pattern of population differentiation which is different from the other classical HLA loci: F_{ST} at these genes is higher than in the control SNPs, even when minor allele frequency is controlled for (Figure S1).

Contrasting F_{ST} of HLA SNPs and haplotypes

Next, we explored population differentiation at the haplotype level (*i.e.* with alleles defined by the coding sequence of each HLA gene). Haplotype level analyses were motivated by the idea that the fitness of individuals is more likely to be determined by the combination of SNPs they carry in a gene, rather than by individual SNPs, since it is the combination of SNPs that determines the peptides which HLA molecules present.

To compare F_{ST} for HLA haplotypes to a null distribution, we recoded each HLA gene as a biallelic locus (see Materials and Methods). Population differentiation at the recoded HLA haplotypes was then compared to differentiation at control SNPs and HLA SNPs, while controlling for minor allele frequency, as was done for SNPs.

The boxplots show that differentiation at HLA haplotypes and HLA SNPs are quite similar (Figure 5) and differentiation at HLA haplotypes is not significantly different from that of control SNPs, except for the MAF bin between 0 and 0.08. Thus, despite the existence of haplotypes which are specific to certain world regions, when the average MAF is considered and global F_{ST} is quantified, the degree of differentiation of HLA haplotypes is lower than that of control SNPs, as was found for HLA SNPs.

It is worth noting that ambiguous HLA haplotype calls (i.e., instances where the typing method provided a set of possible allele calls) was resolved by choosing the assignment that minimized population differentiation. Thus, a more reliable assessment of haplotype-level differentiation will require less ambiguous haplotype calls.

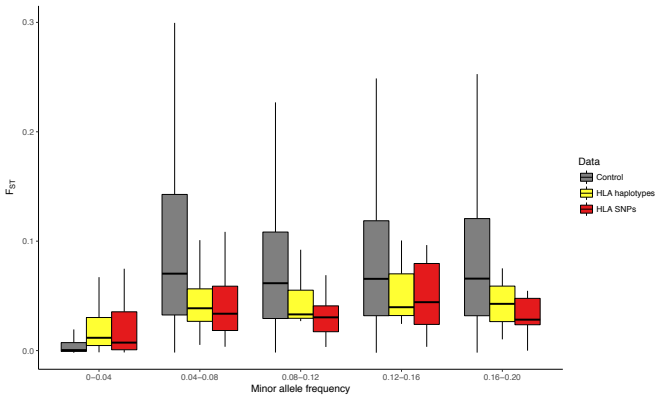

Figure 5 Population differentiation (F_{ST}) as a function of minor allele frequency for the exonic control SNPs (gray), for exonic HLA SNPs (red) and for the haplotype-level *HLA* alleles (yellow) after recoding as biallelic (see Materials and Methods). No haplotypes had frequency higher than 0.2, therefore MAF bins were redefined.

F_{ST} at HLA SNPs depends on divergence times

Our previous analyses examined global F_{ST}, which captures patterns of differentiation among all 20 populations retained from the 1000 genomes full dataset. Next, we asked how specific populations contributed to our findings. In order to investigate this question, and to evaluate how the geographical scale (within and among continents) influences differentiation at the HLA, we analyzed F_{ST} between all pairs of populations.

We found that the lower differentiation at HLA SNPs as compared to control SNPs seen in our previous results (Figure 4) in seen for highly diverged populations (Figure 6) (i.e., contrasts involving populations from different continents).

However, population pairs within the same continent show higher differentiation at the HLA SNPs compared to control

SNPs. This effect was also observed in an independent analysis performed on the populations from the HGDP SNP dataset, with F_{SC} values showing a peak in the MHC region (indicating increased differentiation among populations within continents) (Figure S5).

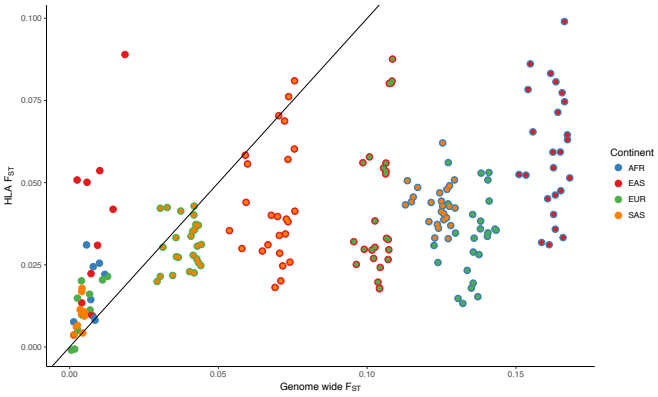

Figure 6 Average F_{ST} at the HLA and at the control SNPs for each pair of populations. Points with a single color represent pairs of populations from the same continent.

Discussion

Population differentiation at HLA SNPs

In an overall analysis of F_{ST} among worldwide populations, we found significantly decreased genetic differentiation at HLA SNPs. We have shown that this result is critically dependent on the use of methods which appropriately account for the properties of the site frequency spectrum of HLA genes.

The decreased differentiation at HLA genes counters the expectation of a model of adaptation to local pathogens driving differentiation. However, we found that the overall pattern of lower differentiation reveals a greater complexity when we compare populations with different divergence times.

Mind the MAF

Our results highlight the importance of accounting for minor allele frequency (MAF) when comparing F_{ST} values for different sets of SNPs, as previously shown by others (Roesti *et al.* 2012; Maruki *et al.* 2012; Elhaik 2012; Jakobsson *et al.* 2013; Edge and Rosenberg 2014). Not accounting for MAF leads to an underestimation of population differentiation for sets of SNPs rich in rare variants. When comparing sets of SNPs from genomic regions with different MAF distributions, this may result in a misleading interpretation of the selective regime acting on each region.

Alcala and Rosenberg (2017) recently arrived at the same results for the upper boundary of F_{ST} as a function of allele frequency in biallelic markers as we present on Supplemental File S1. Their work presents a more general investigation of the constraints of the frequency of the most frequent allele on F_{ST}, using simulations under different migration models. Alcala and Rosenberg (2017) discuss the effect of this constraint in reducing the power of outlier tests which use high F_{ST} as a signature of local positive selection. Here we emphasize the effect of this constraint on interpreting F_{ST} values at regions under balancing selection, where the depletion of rare variants leads to higher overall population differentiation than in other genomic regions. However, when variants with similar MAF are compared, population differentiation in the region under balancing selection is actually lower, as we show for SNPs within HLA genes.

While the constraint of MAF on F_{ST} strongly affects differentiation at regions under positive and balancing selection, it is less problematic in regions under purifying selection. Under this selective regime, both an enrichment at low frequency variants and low population differentiation are expected. Since population differentiation in low frequency variants is constrained to low values, the relationship between MAF and F_{ST} leads both signatures in the same direction. This effect has been demonstrated in *Drosophila melanogaster* (Jackson *et al.* 2014).

Contrast to previous studies

Two classical HLA genes, *HLA-DPA1* and *HLA-DPB1*, were exceptions to our finding of lower population differentiation at HLA SNPs. SNPs in those genes showed higher population differentiation, even when MAF was accounted for (Figure S1).

A genomic scan performed by Bhatia *et al.* (2011) also found a SNP near a HLA-DP gene with unusually high population differentiation (rs2179915, which is 30kb from *HLA-DPA2*). Similarly, Barreiro *et al.* (2008) found unusually high F_{ST} at the *HLA-DPB2* locus. These results indicate that the HLA-DP genes are not under balancing selection, but rather under directional selection that differs among populations. The studies which identified selection at these loci were designed with an emphasis on the detection of extremely high differentiation, and thus did not capture the pattern of low differentiation which is characteristic of most HLA loci.

Hofer *et al.* (2012), on the other hand, used an approach where regions of adjacent SNPs with extreme F_{ST} were scanned for, and detected the *HLA-C* locus among these, with evidence of unusually low F_{ST}. The method used by Hofer *et al.* (2012) evaluates F_{ST} as a function of the heterozygosity between populations, which for biallelic markers is equivalent to the correction for MAF we applied here. These results show that when a test is designed to account for the possibility of unusually low F_{ST} values, and when the effect of minor allele frequency (or heterozygosity) on F_{ST} is accounted for, a signature of low differentiation which would otherwise not be detected can be found.

Divergence time effect

By taking advantage of multiple populations made available by the 1000 Genomes project, we also examined if the excess of low differentiation at HLA SNPs holds at all timescales of differentiation. Interestingly, we find that for population pairs with low divergence (those from the same continent) F_{ST} at HLA SNPs is equal to or higher than in the control SNPs. For highly diverged population pairs (those from different continents), we consistently find lower differentiation among HLA SNPs. This result shows that a specific set of SNPs may differ in how they deviate from the genomic background depending on the timescale of population divergence.

To understand the process driving the increased differentiation for recently diverged populations (Figure 6) we used a simulation approach. First, we simulated a scenario where an ancestral population is under symmetric overdominant selection, and splits into two daughter populations, both under the same selective regime as the ancestor (details in Supplemental File S2). We refer to this scenario as "shared overdominance", and find that it results in differentiation between the daughter populations being reduced with respect to neutral expectations

(Figure 7A and Figure 8A), in accordance with previous results (Schierup *et al.* 2000).

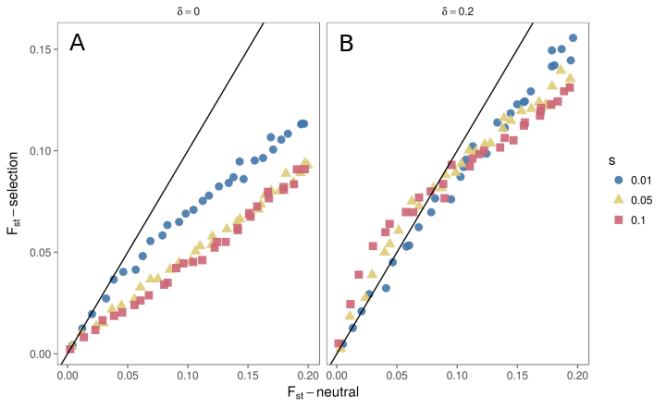

Figure 7 Relation between F_{ST} under neutrality and balancing selection. (A) daughter populations remain under the same regime of overdominance (shared overdominance) as the ancestral population. (B) One of the daughter populations experiences a shift in the fitness values (divergent overdominance), remaining under overdominance but with a new equilibrium value (changed by a value of $\delta = 0.2$. In this case, for recent divergence times we find balancing selection can temporarily increase population differentiation, so long as selection is strong (s=0.05 or greater, $N_e = 1000$).

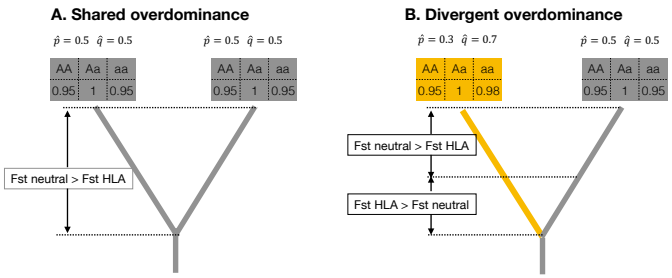

Figure 8 Schematic representation of HLA differentiation under two regimes of overdominance. Each tree represents a population pair experiencing selection according to fitness values presented in the boxes above the tips. In "shared overdominance" both daughter populations share the same fitness values. In "divergent overdominance" the fitness values for one of the homozygotes changes in one of the daughter populations, even though the regime is still one of overdominance. In the divergent overdominance scenario, the equilibrium frequencies at the selected site differs among populations, and drives increased differentiation when divergence is recent (F_{ST} HLA > F_{ST} neutral). For the shared overdominance scenario F_{ST} neutral > F_{ST} HLA throughout the entire history of population divergence.

However, if the two populations remain under overdominant selection but differ for one of the homozygote fitness values (a scenario we refer to as "divergent overdominance"), differentiation can be increased with respect to neutrality for small divergence times (Figure 7B and Figure 8B). This result can be understood if we consider that the populations are extremely similar at the time of the split, so the effect of selection will be to favor changes in allele frequency between them (since homozygote fitness and therefore equilibrium frequencies of selected alleles will differ). As the two populations further diverge, the neutral sites will continue to diverge and will surpass the differentiation for the case of overdominance. Thus, by assuming that the fitness values of an overdominant model can change over time –which is highly plausible if we consider heterogeneity in pathogen populations affecting HLA fitness– increased F_{ST} at HLA for recently diverged populations can be explained.

Consistent with our results, recent studies have identified very recent adaptive change at HLA loci (Field *et al.* 2016; Zhou *et al.* 2016), which could contribute to differentiation at a local scale without erasing signatures of long-term balancing selection.

Acknowledgments

This research was financially supported by grants from São Paulo Research Foundation (FAPESP), The Brazilian National Council for Scientific and Technological Development (CNPq) and the Swiss National Science Foundation (SNSF). DYCB was funded by FAPESP scholarships #2012/22796-9 and #2013/12162-5, JC was funded by FAPESP scholarship #2015/19990-6 and DM has a FAPESP research grant #12/18010-0 and a CNPq productivity grant #308167/2012-0. JG was supported by grant 31003A-138180 of the SNSF.

Literature Cited

Alcala, N. and N. A. Rosenberg, 2017 Mathematical constraints on F ST : biallelic markers in arbitrarily many populations. Genetics **206**: 1581–1600.

Andrés, A. M., 2011 Balancing Selection in the Human Genome. Encyclopedia of Life Sciences pp. 1–8.

Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci, 2008 Natural selection has driven population differentiation in modern humans. Nature genetics **40**: 340–5.

Begovich, A. B., P. V. Moonsamy, S. J. Mack, L. F. Barcellos, L. L. Steiner, S. Grams, V. Suraj-Baker, J. A. Hollenbach, E. Trachtenberg, L. Louie, P. Zimmerman, A. V. S. Hill, M. Stoneking, T. Sasazuki, V. I. Konenkov, M. L. Sartakova, V. P. Titanji, O. Rickards, and W. Klitz, 2001 Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations. Tissue Antigens **57**: 424–439.

Bersaglieri, T., P. C. Sabeti, N. J. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn, 2004 Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am. J. Hum. Genet. **74**: 1111–1120.

Bhatia, G., N. Patterson, S. Sankararaman, and A. L. Price, 2013 Estimating and interpreting FST: The impact of rare variants. Genome research **23**: 1514–21.

Bhatia, G., N. J. Patterson, B. Pasaniuc, N. Zaitlen, G. Genovese, S. Pollack, S. Mallick, S. Myers, A. Tandon, C. Spencer, C. D. Palmer, A. A. Adeyemo, E. L. Akylbekova, L. A. Cupples, J. Divers, M. Fornage, W. H. L. Kao, L. Lange, M. Li, B. O. Tayo, Q. Yang, S. Mccarroll, P. C. Sabeti, G. Lettre, P. D. Jager, J. Hirschhorn, X. Zhu, R. Cooper, D. E. Reich, J. G. Wilson, and A. L. Price, 2011 Genome-wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. The American Journal of Human Genetics **89**: 368–381.

Borghans, J. a. M., J. B. Beltman, and R. J. De Boer, 2004 MHC polymorphism under host-pathogen coevolution. Immunogenetics **55**: 732–9.

Brandt, D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, and D. Meyer, 2015 Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data. G3:Genes | Genomes | Genetics **5**: 931–941.

Cammen, K., J. I. Hoffman, L. a. Knapp, J. Harwood, and W. Amos, 2011 Geographic variation of the major histocompatibility complex in Eastern Atlantic grey seals (Halichoerus grypus). Molecular Ecology **20**: 740–752.

Charlesworth, B., M. Nordborg, and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical research **70**: 155–74.

Čížková, D., J. G. de Bellocq, S. J. E. Baird, J. Piálek, and J. Bryja, 2011 Genetic structure and contrasting selection pattern at two major histocompatibility complex genes in wild house mouse populations. Heredity **106**: 727–740.

Danecek, P., A. Auton, G. R. Abecasis, C. a. Albers, E. Banks, M. a. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin, 2011 The variant call format and VCFtools. Bioinformatics (Oxford, England) **27**: 2156–8.

De Boer, R. J., J. a. M. Borghans, M. van Boven, C. Keşmir, and F. J. Weissing, 2004 Heterozygote advantage fails to explain the high degree of polymorphism of the MHC. Immunogenetics **55**: 725–31.

Doherty, P. C. and R. M. Zinkernagel, 1975 Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature **256**: 50–52.

Edge, M. D. and N. A. Rosenberg, 2014 Upper bounds on FST in terms of the frequency of the most frequent allele and total homozygosity: The case of a specified number of alleles. Theoretical Population Biology **97**: 20–34.

Eizaguirre, C., T. L. Lenz, M. Kalbe, and M. Milinski, 2012 Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. Nature communications **3**: 621.

Elhaik, E., 2012 Empirical distributions of F(ST) from large-scale human polymorphism data. PloS one **7**: e49837.

Ewing, G. and J. Hermisson, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**: 2064.

Field, Y., E. A. Boyle, N. Telis, Z. Gao, K. J. Gaulton, D. Golan, L. Yengo, G. Rocheleau, P. Froguel, M. I. Mccarthy, and J. K. Pritchard, 2016 Detection of human adaptation during the past 2,000 years. bioRxiv .

Goudet, J., 2005 Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes **5**: 184–186.

Gourraud, P.-A., P. Khankhanian, N. Cereb, S. Y. Yang, M. Feolo, M. Maiers, J. D. Rioux, S. Hauser, and J. Oksenberg, 2014 HLA Diversity in the 1000 Genomes Dataset. PLoS ONE **9**: e97282.

Gravel, S., B. M. Henn, R. N. Gutenkunst, a. R. Indap, G. T. Marth, a. G. Clark, F. Yu, R. a. Gibbs, The 1000 Genomes Project Consortium, and C. D. Bustamante, 2011 Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences .

Hedrick, P. W., 2002 Pathogen resistance and genetic variation at MHC loci. Evolution; international journal of organic evolution **56**: 1902–8.

Hedrick, P. W. and G. Thomson, 1983 Evidence for balancing selection at HLA. Genetics **104**: 449–56.

Hill, A. V., C. E. Allsopp, D. Kwiatkowski, N. M. Anstey, P. Twumasi, P. A. Rowe, S. Bennett, D. Brewster, a. J. McMichael, and B. M. Greenwood, 1991 Common west African HLA antigens are associated with protection from severe malaria.

Hofer, T., M. Foll, and L. Excoffier, 2012 Evolutionary forces shaping genomic islands of population differentiation in humans. BMC genomics **13**: 107.

Hollenbach, J. A., G. Thomson, K. Cao, M. Fernandez-Vina, H. A. Erlich, T. L. Bugawan, C. Winkler, M. Winter, and W. Klitz, 2001 HLA Diversity, Differentiation, and Haplotype Evolution in Mesoamerican Natives. Human Immunology **62**: 378–390.

Hu, Z., Y. Liu, X. Zhai, J. Dai, G. Jin, L. Wang, L. Zhu, Y. Yang, J. Liu, M. Chu, J. Wen, K. Xie, G. Du, Q. Wang, Y. Zhou, M. Cao, L. Liu, Y. He, Y. Wang, G. Zhou, W. Jia, J. Lu, S. Li, J. Liu, H. Yang, Y. Shi, W. Zhou, and H. Shen, 2013 New loci associated with chronic hepatitis B virus infection in Han Chinese. Nature genetics **45**: 1499–503.

Hughes, A. L. and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**: 167–170.

Jackson, B. C., J. L. Campos, and K. Zeng, 2014 The effects of purifying selection on patterns of genetic differentiation between Drosophila melanogaster populations. Heredity **114**: 163–174.

Jakobsson, M., M. D. Edge, and N. a. Rosenberg, 2013 The relationship between F(ST) and the frequency of the most frequent allele. Genetics **193**: 515–28.

Klein, J. and A. Sato, 2000 The HLA System. N. Engl. J. Med. **343**: 702–709.

Lawlor, D. A., F. E. Ward, P. D. Ennis, A. P. Jackson, and P. Parham, 1988 HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. Nature **335**: 268–271.

Lewontin, R. C. and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74**: 175–95.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science (New York, N.Y.) **319**: 1100–4.

Loiseau, C., M. Richard, S. Garnier, O. Chastel, R. Julliard, R. Zoorob, and G. Sorci, 2009 Diversifying selection on MHC class i in the house sparrow (Passer domesticus). Molecular Ecology **18**: 1331–1340.

Marsh, S. G. E., E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. a. Erlich, M. Fernández-Viña, D. E. Geraghty, R. Holdsworth, C. K. Hurley, M. Lau, K. W. Lee, B. Mach, M. Maiers, W. R. Mayr, C. R. Müller, P. Parham, E. W. Petersdorf, T. Sasazuki, J. L. Strominger, A. Svejgaard, P. I. Terasaki, J. M. Tiercy, and J. Trowsdale, 2010 Nomenclature for factors of the HLA system, 2010. Tissue antigens **75**: 291–455.

Maruki, T., S. Kumar, and Y. Kim, 2012 Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. Molecular biology and evolution **29**: 3617–23.

McCairns, R. J. S., S. Bourget, and L. Bernatchez, 2011 Putative causes and consequences of MHC variation within and between locally adapted stickleback demes. Molecular Ecology **20**: 486–502.

Mendes, F. H. K., 2013 *Seleção natural em genes HLA e seu efeito sobre regiões adjacentes do genoma*. Ph.D. thesis, Universidade de São Paulo.

Meyer, D., R. M. Single, S. J. Mack, H. a. Erlich, and G. Thomson, 2006 Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. Genetics **173**: 2121–42.

Meyer, D. and G. Thomson, 2001 How selection shapes variation of the human major histocompatibility complex: a review. Annals of human genetics **65**: 1–26.

Miller, H. C., F. Allendorf, and C. H. Daugherty, 2010 Genetic diversity and differentiation at MHC genes in island populations of tuatara (Sphenodon spp.). Molecular Ecology **19**: 3894–3908.

Muirhead, C. a., 2001 Consequences of population structure on genes under balancing selection. Evolution; international journal of organic evolution **55**: 1532–41.

Nunes, K., 2011 *Populações ameríndias da América do Sul: estudo multi-locus e inferência histórico-demográfica e seletiva*. Ph.D. thesis, Universidade de São Paulo.

Oliver, M. K., X. Lambin, T. Cornulier, and S. B. Piertney, 2009 Spatio-temporal variation in the strength and mode of selection acting on major histocompatibility complex diversity in water vole (Arvicola terrestris) metapopulations. Molecular Ecology **18**: 80–92.

Prugnolle, F., A. Manica, M. Charpentier, V. Guernier, M. Population, G. Group, D. Street, U. Kingdom, and E. E. Group, 2005 Pathogen-Driven Selection and Worldwide HLA Class I Diversity. Current Biology **15**: 1022–1027.

Reynolds, J., B. S. Weir, and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics **105**: 767–79.

Roesti, M., W. Salzburger, and D. Berner, 2012 Uninformative polymorphisms bias genome scans for signatures of selection. BMC evolutionary biology **12**: 94.

Sanchez-Mazas, A., 2007 An apportionment of human HLA diversity. Tissue antigens **69 Suppl 1**: 198–202.

Schierup, M. H., X. Vekemans, and D. Charlesworth, 2000 The effect of subdivision on variation at multi-allelic loci under balancing selection. Genetical research **76**: 51–62.

Slade, R. W. and H. I. McCallum, 1992 Overdominant vs. Frequency-Dependent Selection at MHC Loci. Genetics **132**: 323.

Solberg, O. D., S. J. Mack, A. K. Lancaster, R. M. Single, Y. Tsai, A. Sanchez-Mazas, and G. Thomson, 2008 Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Human immunology **69**: 443–64.

Takahata, N. and M. Nei, 1990 Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124**: 967–78.

The 1000 Genomes Project Consortium, 2015 A global reference for human genetic variation. Nature **526**: 68–74.

The International HIV Controllers Study, 2010 The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. Science **330**: 1551–1557.

Vanderborght, P. R., A. G. Pacheco, M. E. Moraes, G. Antoni, M. Romero, A. Verville, V. H. Thai, N. T. Huong, N. N. Ba, E. Schurr, E. N. Sarno, and M. O. Moraes, 2007 HLA-DRB1*04 and DRB1*10 are associated with resistance and susceptibility, respectively, in Brazilian and Vietnamese leprosy patients. Genes and immunity **8**: 320–4.

Wang, K., M. Li, and H. Hakonarson, 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. **38**: 1–7.

Weir, B. S. and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38**: 1358–1370.

Xu, S., S. Li, Y. Yang, J. Tan, H. Lou, W. Jin, L. Yang, X. Pan, J. Wang, Y. Shen, B. Wu, H. Wang, and L. Jin, 2011 A genome-wide search for signals of high-altitude adaptation in Tibetans. Molecular biology and evolution **28**: 1003–11.

Zhou, Q., L. Zhao, and Y. Guan, 2016 Strong Selection at MHC in Mexicans since Admixture. PLoS Genetics .