bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Noname manuscript No.**
(will be inserted by the editor)

# Heterogeneity of $dN/dS$ ratios at the classical HLA class I genes over divergence time and across the allelic phylogeny

**Bárbara Domingues Bitarello, Rodrigo dos Santos Francisco, Diogo Meyer**

**Abstract** The classical class I HLA loci of humans show an excess of nonsynonymous with respect to synonymous substitutions at codons of the antigen recognition site (ARS), a hallmark of adaptive evolution. Additionally, high polymporphism, linkage disequilibrium and disease associations suggest that one or more balancing selection regimes have acted upon these genes. However, several questions about these selective regimes remain open. First, it is unclear if stronger evidence for selection on deep timescales is due to changes in the intensity of selection over time or to a lack of power of most methods to detect selection on recent timescales. Another question concerns the functional entities which define the selected phenotype. While most analysis focus on selection acting on individual alleles, it is also plausible that phylogenetically defined groups of alleles ("lineages") are targets of selection. To address these questions we analyzed how $dN/dS$ ($\omega$) varies with respect to divergence times between alleles and phylogenetic placement (position of branches). We find that $\omega$ for ARS codons of class I HLA genes increases with divergence time and is higher for inter-lineage branches. Throughout our analyses, we used non-selected codons to control for possible effects of inflation of $\omega$ associated to intra-specific analysis, and showed that our results are not artifactual. Our findings indicate the importance of considering the timescale effect when analysing $\omega$ over a wide spectrum of divergences. Finally, our results support the divergent allele advantage model, whereby heterozygotes with more divergent alleles have higher fitness than those carrying similar alleles.

**Keywords** balancing selection, HLA, MHC, $dN/dS$, allelic lineages, antigen recognition site, divergent allele advantage

Address: Departament of Genetics and Evolutionary Biology, University of São Paulo, Rua do Matão, 277, São Paulo. Tel.: +55(11)3091-8092 E-mail: bdbitarello@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



## 1 Introduction

MHC class I and II classical molecules are cell-surface glycoproteins which mediate presentation of peptides to T-cell receptors, and play a key role in triggering adaptive immune responses when the bound peptide is recognized as foreign (Klein and Sato 2000). In humans, they are coded by HLA class I (*HLA-A, -B,* and *-C*) and II (*HLA-DR, -DQ,* and *-DP*) classical genes. The class I and class II HLA classical genes are the most polymorphic in the human genome (Meyer and Thomson 2001), and knowledge about their function in the immune response supports a role for balancing selection in driving the diversity patterns at these loci.

A number of findings suggest MHC genes have experienced balancing selection: unusually high level of heterozygosity with respect to neutral expectations (Hedrick and Thomson 1983); existence of trans-species polymorphisms (Takahata and Nei 1990); high levels of linkage disequilibrium (Huttley et al 1999); site frequency spectra with excess of common variants (Garrigan and Hedrick 2003); high levels of identity-by-descent compared to genomic averages (Albrechtsen et al 2010); positive correlation between HLA polymorphism and pathogen diversity (Prugnolle et al 2005), and significant associations of HLA alleles with the course of infectious diseases (e.g. Apps et al 2013). Information on the crystal structure of MHC molecules (Bjorkman et al 1987) allowed the identification of a specific set of amino acids that make up the antigen recognition site (ARS), which determines the peptides that the molecule is able to bind (Bjorkman et al 1987; Chelvanayagam 1996). The codons of the ARS were shown to have increased nonsynonymous substitution rates (Hughes and Nei 1988, 1989), consistent with the hypothesis that adaptive evolution at HLA loci is driven by peptide binding properties.

Several models of selection are compatible with balancing selection at MHC genes. Heterozygote advantage assumes that heterozygotes have higher fitness values because they are able to mount an immune response to a greater array of pathogens, an idea originally proposed by Doherty and Zinkernagel (1975), who showed that mice which were heterozygous for the MHC had increased immunological surveillance. Heterozygote advantage has received support from experiments in semi-natural populations of mice (Penn et al 2002), which show increased resistance of heterozygotes to multiple-strain infection, and through the finding that among humans infected with HIV, those which are heterozygous for HLA genes have slower progression to AIDS (reviewed in Dean et al 2002). Heterozygote advantage has also received support from substitution rate studies (Hughes and Nei 1988, 1989) as well as simulation-based studies (e.g. Takahata and Nei 1990). A second model for balancing selection at MHC genes is negative frequency dependent selection (or apostatic selection), according to which rare variants have a selective advantage over common ones, because pathogens are more likely to evade presentation by common molecules (Slade and McCallum 1992). Although both are biologically compelling, decades of research have shown that most forms of summarizing genetic observation are incapable of differentiating these two modes of selection (Hughes and Nei 1989; Meyer and Thomson 2001; Spurgin and Richardson 2010), and the functional insights for the action of heterozygote advantage at least partially explain why it is usually favored over negative frequency dependence (Richman 2000).

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



A third model involves selective pressures that are heterogeneous over space and/or time, favoring different alleles in different temporal or geographic compartments, and thus resulting in an overall increase in diversity at MHC loci. This model has been shown to be capable of accounting for features of HLA variation (Hedrick 2002). Many studies have investigated this model by comparing the degree of population differentiation at MHC and putatively neutral loci, with the expectation being that selection that is geographically heterogeneous will result in increased differentiation at HLA genes. As reviewed in Spurgin and Richardson (2010) , the results are mixed, and interpretation is hampered due to differences in the mutational models underlying the evolution of HLA genes and loci used as neutral controls. Although the specific form of selection acting on MHC genes remains an open question, the fact that these genes have evolved in a non-neutral way and are under balancing selection is an undisputed finding, which is robust to complications introduced by demographic history (Harris and Meyer 2006; Hughes and Yeager 1998; Garrigan and Hedrick 2003).

While studies of MHC have documented convincingly a role of selection, certain questions remain unresolved in the context of variation of the human MHC genes (termed HLA loci). The first of these concerns the "timescale" of selection: while most tests for selection have provided strong evidence for selection at classical HLA class I genes in in deep timescales, there is comparatively less support for selection at recent timescales (Garrigan and Hedrick 2003). It has proved difficult to tease apart the possibility that selection differs across timescales from reduced statistical power of tests for recent selection, and thus the question of the timescale of selection on HLA genes remains open.

The second question concerns targets of selection, i.e, which biological entity is targeted by selection in HLA class I genes: individual alleles or groups of similar alleles? Classical MHC genes have many alleles, which can be hierarchically classified into groups of alleles which reflect the phylogenetic relatedness and shared functional attributes of these alleles. Wakeland et al (1990) proposed a mechanism coined "divergent allele advantage", which is a specific case of heterozygote advantage, according to which the fitness values of heterozygotes are proportional to the degree of divergence between the alleles they carry. This model was motivated by the observation that, in MHC class II murine genes, alleles from a given allelic lineage often differ by only minor structural variations in the ARS, while alleles in different lineages have functionally different ARS. The open question is whether individual alleles or allelic lineages are the main targets of selection for HLA genes. Although nucleotide diversity intra-lineages exceeds genome-wide averages, inter-lineage diversity is substantially higher than intra (Takahata and Satta 1998). This raises the question of whether intra-lineage variation is under a different mode and intensity of selection with respect to differences between lineages.

We address these questions by analysing the temporal and phylogenetic dynamics of $dN/dS$ (or $\omega$) for ARS codons at the class I classical loci ($HLA$-$A$, -$B$ and -$C$) loci, using both pairwise and phylogenetic approaches. These loci are all highly polymorphic and there is an abundance of data available for most exons of their coding sequence, which makes our analyses of non-ARS codons (as a control) possible. Our pairwise comparisons of alleles show that more divergent pairs show higher $\omega$ for ARS codons than closely related pairs of alleles. The phylogenetic analyses support the hypothesis that selection is stronger for inter-lineage branches (i.e, those connecting two clades from the same lineage, as opposed to those who do not), and also which are

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



internal to the phylogeny (when compared to terminal branches), provided that a bias toward overestimating $\omega$ for recent divergence is taken into account (Rocha et al 2006). Although evidence for balancing selection on the intra-lineage scale is weaker than on the inter-lineage scale, our findings show that there is statistical support for deviation from a regime of neutrality for intra-lineage branches of the allelic tree. We conclude that intra-lineage divergence has also evolved under a regime of balancing selection, and that inter-lineage divergence bears an even stronger signature of selection.

## 2 Materials and Methods

### 2.1 Data

Alignments for *HLA-A*, *HLA-B* and *HLA-C* were obtained from the IMGT/HLA Database (Robinson et al 2013). All $dN/dS$ estimates and related analyses were implemented in CODEML (PAML package, Yang 2007). First codon position was considered to be the first codon of exon 2, as indicated by annotation on IMGT alignments. Our initial data sets were comprised of complete coding sequences, i.e, exons 2-7 (for *HLA-A* and *HLA-C*) and 2-6 (*HLA-B*). These data sets were used for the site models (SM) approach. For the pairwise and branch model (BM) approaches, we used two datasets: one with 48 ARS codons (Chelvanayagam 1996) and the other, referred to as "non-ARS", consisting of the remaining codons (Table 1).

In order to be able to use the methods available in CODEML we restricted our analysis to HLA alleles which had complete coding sequences, no stop codons, were expressed in the cell surface and only differed with respect to others by base changes (i.e. no insertions or deletions). Alleles with mutations putatively linked to low or absent cell surface expression were also remove from analyses. The non-ARS data sets were used for estimation of $dS$, used in the pairwise approach as a proxy for allelic divergence, as an internal control for ARS analyses. For the branch models, further pruning of the phylogenetic trees was done, as described below.

### 2.2 Trees and intragenic recombination detection

*Phylogenetic trees* Complete alignments, described above, were used to generate NJ trees for each gene (Saitou and Nei 1987). The program NEIGHBOR, from the PHYLIP package (Felsenstein 1989) was used with the F84 method, $k$ (transition/transversion ratio) $= 2$ and empirical base frequencies for the distance matrices obtained in DNADIST (Felsenstein 1989).

*Recombination detection.* Intragenic recombinants were detected by applying RDP3 (Martin et al 2010) to the complete alignments, followed by manual inspection. The RDP3 program combines several non-parametric recombination detection methods in sequence data, and we used 6 independent tests for recombination detection: RDP; Chimaera; Maxchi; GENECONV, BootScan and SiScan for recombination detection (see Martin et al. 2010 and references therein). Window size was adjusted to 100 for BootScan and SiScan, and to 15 for RDP. The number of variable sites per window was adjusted to 35 and 30 for Maxchi and Chimaera, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



These sizes were chosen based on a test alignment we provided to the software, in which parental and daughter *HLA-B* sequences were known *a priori*. Based on this training set, we adjusted the parameters as described, and for other parameters default values were used. Since these six tests are mostly independent, and have different strengths, we considered a recombination event to be significant when $p < 0.05$ in at least 3 of the above methods, which means we were somewhat conservative in the removal of recombinants from the datasets. "Trace evidence" cases, i.e, those that bear a signal of recombination but are technically not statistically significant, were kept in the data sets. Following this initial procedure, we visually inspected the filtered alignments for the detection of additional recombinant sequences. This procedure generated tow data sets for each locus, one with recombinants and one without ("recombinant" (R), and "non-recombinant" (NR), respectively, Table 1).

*Clade Filter* For the branch models, we used $t$ (expected number of nucleotide substitutions per codon) matrices obtained in pairwise analyses of the non-recombinant non-ARS data sets as input for NEIGHBOR. The trees were visualized for manual pruning and labeling in Mesquite (v2.75, http://mesquiteproject.org/). We imposed that alleles from a given HLA lineage (as defined by the standard HLA nomenclature, which identifies lineage membership by the first field of an allele's name) had to group together in a clade, and alleles which did not group in such manner were manually pruned from trees in order to fulfill this "clade membership criterium". The effect of this filtering on inclusion of alleles is presented in Figure S1 in the Online Resource 1. After pruning of the trees, the corresponding pruned alleles were removed from the NR data sets and these reduced data sets were used for the branch model analyses. Table 1 shows the number of alleles used for each analysis.

## 2.3 CODEML analyses

*Branch models (BM)* With the pruned data sets we compared branch models 0 (one $\omega$ for all branches) and 2 (two or more categories of branches with independent $\omega$) from CODEML. We provided CODEML with a topology based on the non-ARS pruned data set, using branch lengths as starting points for ML estimation (fix_blength=1). For all CODEML analyses (BM, site models and pairwise), the Goldman and Yang (1994) model was used for estimation of substitution rates. Other parameters defined in the control file were as follows: option F3x4 for codon frequency estimation, $\kappa = 2$ and $\omega = 0.4$ as initial values. Tables S14-S16 (Online Resource 1) show likelihood convergence for the branch models, assuming different initial parameter values and codon frequency estimation methods. BM analyses were performed solely for the NR data sets (see tables 2 and 3). Branch models 0 (one omega for all branches) and 2 (two or more omegas) were compared, where branches were labeled either as "intra" or "inter" lineages (Figure 3), or as "terminal" or "internal". The two models were compared *via* a likelihood ratio test (LRT) with one degree of freedom (see below). BM analyses were performed only for the NR (and pruned) datasets. See Figure 3 for an schema of the labels applied to the trees used in the BM analyses.

*Site models (SM)* For the SM approach, the clade filter was not applied, which resulted in minor differences between this data set and the other two (pairwise and branch models approach, see Table 1). We used the

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



site models from CODEML to identify codons with $\omega > 1$ and thus to test if ARS codons bear evidence for adaptive evolution. M0 (one ratio) assumes the existence of only one $\omega$ ratio for all codons, while M1 (neutral) assumes the existence of two categories of sites, one with $\omega_1 = 1$ (sites evolving in a neutral fashion) and the other with $\omega_o < 1$ (sites evolving under purifying selection), while M2 (selection) adds an extra category to M1, where $\omega_2 > 1$, corresponding to sites with evidence for adaptive evolution. M7 (beta) is a flexible null model where the value is sampled from a beta distribution, where $\omega_0 < 1$, and $0 < \omega < 1$ , while M8 adds an extra category to M7, $\omega_2$, which is estimated from the data (Yang 2006). Codons with posterior probabilities $P > 0.95$ of $\omega > 1$ in the Bayes Empirical Bayes (BEB) (Yang et al 2005) approach implemented in CODEML were considered to have significant evidence for adaptive evolution, following criteria described elsewhere (Yang and Swanson 2002; Yang et al 2005). The ARS codon classification proposed by Bjorkman et al. (1987) is referred to as BJOR, while the "peptide binding environments", i.e, the amino acid residues in a fixed neighborhood of the peptide binding residues known from crystal structure complexes (which provide a less restrictive description of the antigen binding sites), are referred to as CHEV (Chelvanayagam 1996). Finally, the list of codons in HLA genes with evidence of $\omega > 1$ from Yang and Swanson (2002) is referred to as YANG (Figure 1 and Online Resource 1, Table S9). M1 *vs* M2 and M7 *vs* M8 models were compared through a LRT with two degrees of freedom. Tables S3-S8 (Online Resource 1) show likelihoods obtained when altering initial CODEML conditions for the SM analyses. SM analyses were performed for R and NR data sets.

Codons with $P > 0.95$ for $\omega > 1$ in M8 (34 in total) were combined for the three loci, and the R and NR data sets, and compared to CHEV, BJOR and YANG. Of these 34 codons, only one was outside of the exons 2 and 3 range (codon 305), which is where all ARS codons are located. Figure 1 shows the overlap between the codons defined as making up the ARS in the BJOR and CHEV classifications, as well as those idenfied as under selection in the YANG set of codons and our analyses.

In order to evaluate if our site model analyses were robust to features of the estimation method, the analyses were repeated with DATAMONKEY, from the HYPHY package (Pond et al, 2005). The substitution model used for construction of the NJ tree was HKY85 (very closely related to F84, used for CODEML analyses). Two criteria for detection significant $dN/dS > 1$ were considered: SLAC and FEL (both with significance level of 0.1), with the former being the most conservative criterion available in the package. Tables S10-S12 report the overlap of sites with evidence for $dN/dS > 1$ for BEB (CODEML), SLAC and FEL.

*LRT* When comparing two nested models the LRT test statistic is given by doubling the log likelihood differece between the more parameter rich model and the less parameter rich model. The difference in parameter number yields the degrees of freedom. It is expected that the use of a chi-square distribution for significance evaluation of this test is a conservative approach (Yang 2006). Both site models and branch models comparisons were performed through LRTs.

*Breslow-Day Test* In order to compare ARS and non-ARS codons with respect to the distribution of synonymous and nonsynoymous changes within and between lineages (or for internal or terminal branches), we used a contingency table approach similar to the one described in Templeton (1996). We estimated the synonymous

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



($S$) and non-nonsynonymous ($N$) changes on each branch in CODEML, using the branch models. Next we counted $N$ (nonsynonyous changes) and $S$ (synonymous changes) for intra/inter or terminal/internal branches for each locus, and for ARS and non-ARS codons (Table 5).

We defined the odds ratio (OR) as:

$$\frac{N_{intra} \cdot S_{inter}}{N_{inter} \cdot S_{intra}}$$

, and used a Breslow-Day test for homogeneity of OR to test the hypothesis that contingency tables from ARS and non-ARS codons have the same OR. We applied the same test to internal/terminal branches. Data from the three loci were combined into the same analysis to increase power.

*Pairwise approach* We also performed analyses where statistics were estimated in comparisons between all pairs of alleles (pairwise analyses, see Table 1) using runmode=-2 in CODEML. This approach does not require a phylogenetic tree. Because IMGT/HLA nomenclature allows information about allelic lineages to be known without a tree, pairs were also classified as intra or inter-lineage. Correlations between allelic divergence and omega values were tested with a Mantel Test using Pearson's correlation index (Online Resource 1, Table S13). We obtained quantiles of the $dS_{\text{non}-\text{ARS}}$ distribution and divided pairwise values according to these quantiles (Online Resource 1, Table S1 for non-ARS data set and Table 4 in main text for ARS data set). Differences in mean $\omega$ values for "intra" and "inter" comparisons were tested for significance by a Wilcoxon rank sum test (Figure 2).

## 2.4 Allele frequencies of *HLA* SNPs in the 1000 Genomes

The IMGT/HLA database contains all HLA alleles described to date, regardless of their population frequencies. Therefore, it is possible that rare variants can contribute disproportionately to patterns identified in the $dN/dS$ analyses. To address this concern, we investigated patterns of variation at the HLA loci in a population (Yoruba, YRI) from the 1000 Genomes Project (1000G), for which frequency of alleles at specific SNP positions is available ($N = 88$ individuals).

To test for a possible enrichment of rare variants in the IMGT data we compared patterns of variation seen in the IMGT and 1000G phase I data (The 1000 Genomes Project Consortium, 2012). To this end, we defined a set of sites, for each locus, which were variable in our IMGT-derived data sets (referred to as the "OVERALL" set of sites). Next, we classified these sites as variable only within a single lineage ("INTRA"), or variable in more than one lineage ("INTER"). For each site, we converted the positions within the HLA locus into a genomic coordinate for *H. sapiens* (hg19).

Next, we verified if these positions are polymorphic in the 1000G Phase I low-coverage dataset (ftp://ftp.1000genomes.ebi.ac.u and recorded the minor allele frequency in the YRI population.

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



## 3 Results

### 3.1 Evidence for selection and assessment of recombination

Before investigating how $\omega$ varies over time and phylogenetic context, we tested (a) whether selection is detectable in our data set with pairwise comparisons and phylogenetic $dN/dS$ approaches; (b) if the presence of *HLA* alleles resulting from intragenic recombination influences our inferences; and (c) if there is agreement between the ARS codons defined by crystal structure (Bjorkman et al 1987; Chelvanayagam 1996) and the codons inferred to have $\omega > 1$ in our data set. The results to these tests are pre-requisites for subsequent analyses addressing the more specific hypotheses about heterogeneity in $dN/dS$ estimated across the allelic phylogeny and divergence time.

We quantified the mean pairwise $dN/dS$ ($\overline{\omega}$), and found $\overline{\omega}_{\texttt{ARS}} > 1$ for all loci (Table 4). We used the non-ARS codons from the same sequences as an internal control, and found that $\overline{\omega}_{\texttt{ARS}}$ is 3.9 (*HLA-A*), 4.0 (*HLA-B*) and 3.2-fold (*HLA-C*) greater than $\overline{\omega}_{\texttt{non-ARS}}$ (Table 4). This effect is not driven by a subset of the pairwise comparisons, since $dN > dS$ for the majority (between 67 and 84%) of ARS pairwise comparisons, in contrast to the non-ARS comparisons, where fewer than 7% show $dN > dS$ (Table 4). Importantly, we find that the result $\overline{\omega}_{\texttt{ARS}} > \overline{\omega}_{\texttt{non-ARS}}$ is due to increased $\overline{dN}$ (3.5 to 14-fold higher for ARS), and not to decreased $\overline{dS}$ (0.5 to 2.8-fold higher for ARS, Table 4). Qualitatively similar results were obtained when we computed the ratio of mean substitution rates, $\overline{dN}/\overline{dS}$ (Table 4). These findings are robust to the presence of recombinants (Online Resource 1, Table S1). Overall, our results document that pairwise comparison of alleles provides strong support for adaptive evolution on ARS codons, as expected.

Evidence for adaptive evolution in ARS codons was also strongly supported by phylogenetic methods from CODEML (see Methods), where models allowing for selection (M2 and M8) in a subset of codons were significantly favored over the neutral models M1 and M7 (Online Resource 1, Table S2; $p < 0.01$, LRT). Results were robust to starting conditions for *HLA-A* and *HLA-B* (Online Resource 1, Tables S3-S6), and less so for *HLA-C* (Online Resource 1, Tables S7 and S8).

We next quantified the overlap between codons we inferred to be under selection (using site models from CODEML, "SM") and those defined as ARS based on structural analyses of HLA (Chelvanayagam 1996; Bjorkman et al 1987). Within exons 2 and 3 (which contain all ARS codons) we identified 33 codons with significant $\omega > 1$ for the M8 site model (see Methods and Table S2, Online Resource 1) in at least one locus, of which 27 (82%) are contained within the set that forms the ARS according to the crystal structure-based classification (Bjorkman et al 1987), 25 (76%) are contained within the peptide binding environments (Chelvanayagam 1996), and 25 (76%) overlap with Yang and Swanson's (2002) site models approach to detect codons with $\omega > 1$ in the three classical class I HLA loci (Figure 1 and Online Resource 1, Table S9). The association between ARS and selected sites for all loci is highly significant ($p < 10^{-11}$, chi-square test). There is extensive overlap between the two ARS classifications (Bjorkman et al 1987; Chelvanayagam

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



1996) (Figure 1) and we also find a high overlap of selected sites between the R and NR data sets for each locus (27 out of 33) (Online Resource 1, Tables S10-S12).

Overall, our results show that: (a) the pairwise and phylogenetic site models methods implemented in CODEML strongly support adaptive evolution on the ARS codons of *HLA* loci - as also described by Yang and Swanson (2002) through site models; (b) there is an enrichment of codons with $\omega > 1$ in the CHEV set of codons (see Online Resource 1, Table S9, for the names given to the sets of codons), supporting the use of this classification for our study; (c) although the results were robust to the presence of recombinants, a finding consistent with simulation studies (Anisimova et al, 2003), the estimated values for $\omega$ appear to be sensitive to the inclusion of recombinants. Therefore, where appropriate, in subsequent pairwise analyses, we contrast results of non-recombinant (NR) and recombinant (R) datasets, while for the branch models we use the NR data set exclusively.

In addition, the results of *HLA-C*, although following the same trend observed for *HLA-A* and *HLA-B*, show that absolute divergence values for ARS codons are on average 1/2 of those observed for the other two loci, both for $dN$ and $dS$ (Table 4). This result might be a reflect of the fact that *HLA-C* not only has an antigen presentation function, but has a huge role in interactions with NK receptors (KIR) and that, unlike *HLA-A* and *HLA-B*, all *HLA-C* allotypes form ligands for *KIR* receptors (Hilton et al 2015; Single et al 2007). Because the *KIR* loci have been shown to evolve quite rapidly across primate species, plausibly faster than their MHC class I ligands (Single et al, 2007), it is possible that this important selective pressure is responsible for the lower substitution rates seen for the ARS of *HLA-C*, as well as for the lack of consistency observed in ML estimates

### 3.2 The time-dependence of $\omega$ at HLA class I loci

Having confirmed that selection at ARS sites is detectable with pairwise comparisons and phylogenetic approaches, we investigated if recent evolutionary change (accounting for differences among recently diverged alleles) shows different signatures of selection with respect to changes that occurred over greater timescales. Our first approach consisted in examining the distribution of $\omega_{\texttt{ARS}}$ as a function of the time since divergence between allele pairs. Our estimate of divergence time between allele pairs was based on the values of $dS$ (estimated from non-ARS codons) for each allele pair, thus avoiding statistical non-independence with $\omega_{\texttt{ARS}}$. Because very recently diverged alleles have low synonymous divergence ($dS_{\texttt{non-ARS}}$ ), the corresponding $\omega_{\texttt{ARS}}$ values were often undefined or extremely large. We therefore followed a strategy adopted by Wolf et al (2009) to filter out the allele pairs with $\omega_{\texttt{ARS}} > 5$ (resulting in the removal of 1.1%, 1.4% and 3.9% of $\omega$ values for pairwise comparisons at *HLA-A*, *-B*, and *-C*, respectively).

Pairwise estimates show that $\omega_{\texttt{ARS}}$ increases as a function of divergence time (Table 4). Indeed, $\omega_{\texttt{ARS}}$ and $dS_{\texttt{non-ARS}}$ are positively correlated (Online Resource 1, Table S13; $r_{\texttt{HLA-A}} = 0.17$, $p < 0.001$; $r_{\texttt{HLA-B}} = 0.20$, $p < 0.001$; $r_{\texttt{HLA-C}} = 0.20$, $p < 0.001$; Pearson, significance obtained by Mantel Test). Qualitatively similar results were found for NR data sets and were robust to different correlation measures (Online Resource 1,

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Table S13). We also compared the $\omega$ between allele pairs classified as intra and inter-lineage (Figure 2). For all loci, the median value of $\omega_{\texttt{ARS}}$ is $> 1$ for the inter lineage contrasts, and $< 1$ for the intra-lineage contrasts, and the distribution of $\omega$ is significantly higher for inter-lineage contrasts ($p < 0.001$, Wilcoxon rank sum test; Figure 2) of the ARS codons.

The above pairwise comparison approach suffers from the limitation that allele pairs with $\omega > 5$ were treated as missing data, possibly underestimating $\omega$ for recently diverged alleles. This prompted us to use a phylogenetic model to contrast alleles at different levels of differentiation, which is more robust to the effects of low differentiation between specific allele pairs. We compared a branch model that estimates a single $\omega$ for all branches to one that estimates two values of $\omega$ (inter *versus* intra-lineage; terminal *versus* internal; see Figure 3). For all loci we found higher $\omega_{\texttt{ARS}}$ for inter-lineage branches than for intra-lineage branches, although significance was not attained for these tests (Table 2). For the contrast between internal and terminal branches, we found higher $\omega_{\texttt{ARS}}$ for internal branches at all loci and this result was statistically significant for *HLA-C* (Table 2).

Our results show that both pairwise comparisons and branch models indicate a heterogeneity of $\omega$ throughout the diversification of HLA alleles, with higher $\omega$ values associated to contrasts between more divergent alleles (pairwise approach) or to branches connecting different lineages or that are internal to the phylogeny (BM approach), although the difference was not significant for the "intra-inter" contrasts.

### 3.3 Significantly more nonsynonymous changes inter-lineages at ARS codons

In this study we estimate $\omega$ for allele pairs or branches sampled within a single species, and over varying timescales. Both these features imply in possible biases to the estimation of $\omega$, which we now discuss.

Kryazhimskiy and Plotkin (2008) used analytical and simulation approaches to show that under positive selection the behavior of $\omega$ within a single population is not a monotonic function of the intensity of selection, so that $\omega$ intra a population can be low, even under positive selection. This occurs because, when an advantageous nonsynonymous variant is fixed in a population, nonsynonymous variation can be decreased due to the homogeneity generated by the selective sweep. However, this scenario clearly does not apply to HLA genes, where balancing selection maintains multiple nonsynonymous polymorphisms simultaneously segregating within a population, contributing to $\omega > 1$.

Another challenge to the interpretation of $\omega$ arises from that fact that many studies have shown that genes under purifying selection show surprisingly high $\omega$ (often close to 1) when samples with short divergence times are analyzed (e.g., those from a single population or species). For example, Rocha et al (2006) showed that $dN/dS$ between two samples is negatively correlated with their divergence times, and exemplified these predictions with bacterial genomes. Likewise, a decrease of $dN/dS$ with divergence time has been described in Wolf et al (2009), but considering a much deeper timescale. Kryazhimskiy and Plotkin (2008) demonstrated that this pattern is expected even under a regime of purifying selection that is constant over time. Thus, it is plausible that the recent divergence times among alleles within HLA allelic lineages could result in inflated

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



intra-lineage $\omega$ values, explaining the modest differences between intra and inter-lineage $\omega$ values seen in the phylogenetic analyses (Tables 2 and 3). To explore this issue further, we used non-ARS codons as an internal control for this putative build-up of $dN/dS$ are recent timescales, and to do so we compared their patterns of variation to those of ARS codons. We found that non-ARS codons have larger intra-lineage $\omega$ values than inter-lineage values, and also higher $\omega$ for terminal than internal branches ($p < 0.05$ for *HLA-A* in the intra *versus* inter-lineage contrast, and for *HLA-A* and *HLA-C* in the tips *versus* internal contrast; LRT; Table 3). This distribution of $\omega$ values is in the exact opposite direction to that observed for the ARS (Table 2), consistent with an effect of short divergence times inflating the estimates of $\omega$ (Kryazhimskiy and Plotkin 2008).

In order to formally test whether ARS and non-ARS codons have a different distribution of synonymous and nonsynonymous changes intra and inter-lineages (or for internal and terminal branches) we employed a contingency table approach similar to that of Templeton (1996). We used the inferred number of synonymous ($S$) and nonsynonymous ($N$) changes on each branch of the allelic phylogeny from each locus to estimate the total number of each type of change in a specific class of branches (see Figure 3 for a schematic representation of the branch labeling).The odds ratio was defined as presented in the Methods. For all loci, we find that $OR > 1$ for non-ARS codons (proportionally more nonsynonymous on the intra-lineage branches) and $OR < 1$ for ARS codons (proportionally more nonsynonymous changes on the inter-lineage branches), as shown in Table 5. This finding is consistent with the maximum likelihood estimates of $\omega$ for branches (Tables 2 and 3), and the increased pairwise $\omega$ inter-lineage, relative to intra-lineage (Figure 2). To test for differences between ARS and non-ARS codons, we pooled the contingency tables of all loci (due to the fact that several cells for individual loci had low counts) and rejected the null hypothesis that contingency tables from ARS and non-ARS codons have the same $OR$ ($p - value = 0.0069$; Breslow-Day test). Our analysis comparing internal and terminal branches showed the same pattern, with proportionally more nonsynonymous changes in internal branches for ARS codons ($p - value = 0.00013$; Breslow-Day test; Table 5).

In summary, although there is evidence for an excess of inter-lineage nonsynonymous changes (or for terminal branches) for ARS codons, there is also an enrichment for intra-lineage nonsynonymous changes for ARS codons, when compared to non-ARS codons ($P < 0.001$; Fisher's exact test). Next, we discuss possible biases in the data set which could lead to these results.

### 3.4 Comparing $dN/dS$ results with 1000 genomes variation

Our analyses are based on allele sequences available in the IMGT/HLA data base, which is a curated resource to which newly discovered alleles are contributed. This data set is likely to be biased with respect to population frequencies, since very rare HLA alleles are likely to represent a disproportionately larger fraction than in true population samples, since all new alleles which are discovered are encouraged to be submitted to IMGT. We therefore investigated if this bias influenced our findings. Specifically, we were concerned that the enrichment for rare variants could result in an inflation of weakly deleterious nonsynonymous variants for recent divergence,

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



a well documented population genetic signature (Henn et al 2015). This signature could create an artificially inflated value of $\omega$ for intra-lineage variability.

We found that only a subset of variable positions present in our IMGT-derived datasets are present in the 1000 Genomes Phase I low coverage data (Tables 6, 7, and 8 (for *HLA-A*, *HLA-B*, and *HLA-C*, respectively). This is in accordance with the greater degree of sampling of rare variants in the IMGT data set.

We next divided positions into two groups: those which are only variable within a single lineage ('INTRA'), and those variable in more than one lineage ('INTER'). For comparison, a third group, which consists of all variable sites ('OVERALL'), was also defined. We found that, considering all INTRA and INTER positions present in the 1000G data, there is no significant difference in minor allele frequency (MAF) between the two categories (Tables 6, 7, and 8). Furthermore, when we classify the 1000G HLA SNPs into low (MAF<=0.1) and high frequency (MAF>0.1), we do not see an enrichment for low frequency variants within the "INTRA" set of SNPs when compared to the "INTER" set (Wilcoxon test, not shown).

These results reassure us that the intra-lineage variation we observe is not biased in the direction of extremely rare variants, and that our observation that there is evidence for stronger intra-lineage balancing selection for ARS codons than for neutrally evolving regions (non-ARS) is not a spurious result driven by an enrichment for low-frequency SNPs.

## 4 Discussion

Our study documents a positive correlation between $dN/dS$ values and the degree of divergence between allele pairs. This result is supported by phylogenetic analyses, which show higher $\omega$ values for branches connecting different lineages, or branches which are internal to the phylogeny. A heterogeneous nonsynonymous substitution rate ($dN$) for HLA genes was also reported in a study which found that $dN$ for ARS codons is not linearly correlated with divergence time in classical HLA loci (Yasukochi and Satta 2014). By further investigating the temporal dynamics in the *DRB1* gene, these authors showed that this rate heterogeneity is likely the consequence of a reduction in the substitution rates in specific allelic lineages, possibly as a consequence of continuous selective pressure by a specific pathogen. In the present study our goal was to explicitly test for heterogeneity in the $\omega$ ratios over *a priori* defined groups of alleles (the HLA allelic lineages) and for timescales of divergence (low and high divergence). As was the case with the study of Yasukochi and Satta (2014), we find heterogeneity in the intensity of selection, in our case with evidence of increased selection at deeper timescales than at more recent ones, and for greater selection on inter-lineage branches of the allelic phylogeny, with respect to intra-lineage branches. Our findings indicate that long-term balancing selection has resulted in an enrichment for adaptive changes between allelic lineages for HLA class I genes, with proportionally weaker signatures of molecular adaptation for recent (terminal and intra-lineage branches) than for the inter-lineage and for the internal branches.

Although previous studies have shown that low divergence is often associated to inflated $\omega$ estimates (Rocha et al, 2006), the phylogenetic analyses carried out in the present work relied on non-ARS codons as a control

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



to show that low divergence times of intra-lineage contrasts does not explain the $\omega > 1$ values within lineages, at ARS codons. Thus, while we show that inter-lineage selection is stronger than intra-lineage selection, our results also demonstrate that intra-lineage variation bears a signature of balancing selection.

Recently several papers have drawn attention to the effects of divergence times on $dN/dS$ estimation (e.g. Wolf et al 2009; Stolestki and Eyre-Walker 2011), and the complexities of interpreting these values when data is drawn from a single population (Rocha et al 2006; Kryazhimskiy and Plotkin 2008). Our finding of increased $\omega_{\texttt{ARS}}$ among more divergent alleles (or for inter-lineage branches) is conservative in light of these findings, which predict decreased $\omega$ for more divergent alleles. We accounted for this effect by using non-ARS codons, which have a similar phylogenetic structure to that of ARS codons (after removal of recombinants) to control for the background inflation of omega in recently diverged alleles, and found that ARS codons have very different distribution of $\omega$, with increased inter-lineage evidence for selection, exactly the opposite to what is seen for non-ARS codons.

An important caveat to this interpretation is that the temporal dynamics of $dN/dS$ appears to be sensitive to the selective regime which is assumed to be operating. Thus, while several authors have shown that, under purifying selection, increased $dN/dS$ at low divergence is expected, positive selection can produce a positive correlation with divergence times (Dos Reis and Yang 2013; Mugal et al 2014), which could account for part of the results we describe in this study. However, the case of directional positive selection, involving the sequential substitution of adaptive mutations, is markedly different from the dynamics of a balanced polymorphism, as is the case for HLA genes.

Assuming that balancing selection has been the main selective regime shaping the molecular evolution of HLA genes, and that heterozygote advantage is one (even if not exclusively) of the mechanisms through which selection has acted upon this system, our finding that inter-lineage $\omega_{\texttt{ARS}}$ is greater than intra-lineage is consistent with the divergent allele advantage model, according to which heterozygotes for more divergent alleles have higher fitness than those carrying similar alleles (Wakeland et al 1990). Under this model, excess of inter-lineage nonsynonymous changes in HLA genes would be expected, which is a result we have shown for the ARS data set. This model has been shown to explain patterns of variation in the *DRB* locus in Galapagos sea lions, where local allelic divergence at this locus positively influences fitness directly (Lenz et al 2013), and not mere heterozygosity or number of alleles at the MHC locus. Most likely several selective regimes have shaped the evolutionary history of MHC genes, as suggested by previous observations, and our contribution suggests that these selective regimes could be operating alongside with divergent allele advantage.

Our results suggest that groups of functionally related alleles (in our analysis, the allelic lineages) should be regarded as important targets of selection, rather than individual alleles. In line with our observations, it has been proposed that HLA supertypes - groups of alleles sharing chemical properties at the B and F pockets of the ARS region (Sidney et al 1996) - constitute the level of variation that is the primary target of natural selection in *HLA-B* genes (Francisco et al 2015). Since there is a high overlap between allelic lineage and supertype classifications(Sidney et al 1996), our results indicate that attempts to understand how natural

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



selection acts on HLA variation benefit by comparing the effects of selection on the allelic, allelic lineage or supertype levels of variation.

Electronic Supplementary Material

Supporting tables are available as an additional file.

Competing Interests

The authors declare that they have no competing interests.

Author's Contributions

BDB carried participated in the design of the study, performed analyses, discussed results and drafted the manuscript. RDF performed analyses and discussions. DM conceived of the study, participated in its design and discussion and in the drafting of the manuscript. All authors read and approved the final manuscript.

**Acknowledgements** The authors thank Kelly Nunes for thoughtful comments on the manuscript, Richard Single for comments on the statistical aspects of this work, Aida M. Andrés for general comments and Débora Y.C.Brandt for help with the 1000 Genomes data sets. This work was supported by the São Paulo Research Foundation (grants #2008/09127-8 and #2011/12500-2 to BDB; #08/56502-6 to DM) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (#152676/2011-2 to BDB, #142130/2009-5 to RSF and #308960/2009-2 to DM). The final publication is available at Springer via http://dx.doi.org/DOI: 10.1007/s00239-015-9713-9

Data available in public repositories

`https://github.com/bbitarello/dNdS-hla-allelic-lineages`

## References

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

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164(3):1229–36

Apps R, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, Yuki Y, Del Prete GQ, Goulder P, Brumme ZL, Brumme CJ, John M, Mallal S, Nelson G, Bosch R, Heckerman D, Stein JL, Soderberg Ka, Moody MA, Denny TN, Zeng X, Fang J, Moffett A, Lifson JD, Goedert JJ, Buchbinder S, Kirk GD, Fellay J, McLaren P, Deeks SG, Pereyra F, Walker B, Michael NL, Weintrob A, Wolinsky S, Liao W, Carrington M (2013) Influence of HLA-C expression level on HIV control. Science (80- ) 340(6128):87–91

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987) Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329(6139):506–12

Chelvanayagam G (1996) A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. Immunogenetics 45(1):15–26

Dean M, Carrington M, O'Brien SJ (2002) Balanced polymorphism selected by genetic versus infectious human disease. Annu Rev Genomics Hum Genet 3:263–92

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

Dos Reis M, Yang Z (2013) Why do more divergent sequences produce smaller nonsynonymous/synonymous rate ratios in pairwise sequence comparisons? Genetics 195(1):195–204

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166

Francisco RS, Buhler S, Nunes JM, Bitarello BD, França GS, Meyer D, Sanchez-Mazas A (2015) HLA supertype variation in human populations: new insights about the role of natural selection on the evolution of HLA-A and HLA-B polymorphisms. Immunogenetics, DOI 10.1007/s00251-015-0875-9.

Garrigan D, Hedrick PW (2003) Detecting adaptive molecular polymorphism : Lessons from the MHC. Evolution (N Y) 57(8):1707–1722

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11(5):725–736

Harris E, Meyer F (2006) The Molecular Signature of Selection Underlying Human Adaptations. Yearb Phys Anthropol 130:89-130

Hedrick PW (2002) Pathogen resistance and genetic variation at MHC loci. Evolution (N Y) 56(10):1902–1908

Hedrick PW, Thomson G (1983) Evidence for balancing selection at HLA. Genetics 104(3):449–56

Henn B, Botigué LR, Bustamante C, Clark AG, Gravel S (2015) Estimating the mutation load in human genomes. Nat Rev Genetics 16:333—343

Hilton HG, Guethlein LA, Goyos A, Nemat-Gorgani N, Bushnell DA, Norman PJ, Parham P (2015) Polymorphic HLA-C Receptors Balance the Functional Characteristics of KIR Haplotypes. J Immunol 195:3160-3170

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

Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci U S A 86(3):958–962

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex of vertebrates. Annu Rev Genet pp 415–435

Huttley G, Smith MW, Carrington M, O'Brien S (1999) A scan for linkage disequilibrium accross the human genome. Genetics 152(4):1711–1722

Klein J, Sato A (2000) The HLA system. First of two parts. Adv Immunol 343(10):702–709

Kryazhimskiy S, Plotkin JB (2008) The Population Genetics of dN/dS. PLoS Genet 4(12):10

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Lenz T, Mueller B, Trillmich F, Wolf JBW (2013) Divergent allele advantage at MHC-DRB through direct and maternal genotypic effects and its consequences for allele pool composition and mating. Proc R Soc B 280: 20130714

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P (2010) RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26(19):2462–3

Meyer D, Thomson G (2001) How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet 65(1):1–26

Mugal CF, Wolf JBW, Kaj I (2014) Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol Biol Evol 31(1):212–31

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):11,260–4

Pond SLK, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21(5):676-679

Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F (2005) Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15(11):1022–7

Richman A (2000) Evolution of balanced genetic polymorphism. Mol Ecol 9(12):1953–63

Robinson J, Halliwell Ja, McWilliam H, Lopez R, Parham P, Marsh SGE (2013) The IMGT/HLA database. Nucleic Acids Res 41(Database issue):D1222–7

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239(2):226–235

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sidney J, Grey HM, Kubo RT, Sette A. (1996) Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. Immunol Today 17(6): 261–6

Single RM, Martin MP, Gao X, Meyer D, Yeager M, Kidd JR, Kidd K, Carrington M (2007 Global diversity and evidence for coevolution of KIR and HLA. Nat Genetics 9:1114–1119

Slade R, McCallum H (1992) Overdominant vs. frequency-dependent selection at MHC loci. Genetics 132:861–864

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci 277(1684):979–88

Stolestki N, Eyre-Walker A (2011) The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions. Mol Biol Evol 28(4):1371–1380

Takahata N, Nei M (1990) Allelic Genealogy Under Overdominant and Frequency-Dependent Selection and Polymorphism of Major Histocompatibility Complex Loci. Genetics 124(4):967–978

Takahata N, Satta Y (1998) Footprints of intragenic recombination at HLA loci. Immunogenetics 47(6):430–441

Templeton AR (1996) Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial Cytochrome Oxidase II gene in the hominoid primates.

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Genetics 144(3):1263–1270

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56 —65

Wakeland EK, Boehme S, She JX, Lu Cc, Mclndoe RA, Cheng I, Ye Y, Potts WK (1990) Ancestral Polymorphisms of MHC Class II Genes : Divergent Allele Advantage. Immunol Res 9:115–122

Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H (2009) Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. Genome Biol Evol 1:308–319

Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24(8):1586–1591

Yang Z, Swanson WJ (2002) Codon-Substitution Models to Detect Adaptive Evolution that Account for Heterogeneous Selective Pressures Among Site Classes. Mol Biol Evol 19(1):49 –57

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22(4):1107–1118

Yasukochi Y, Satta Y (2014) Nonsynonymous Substitution Rate Heterogeneity in the Peptide-Binding Region Among Different HLA-DRB1 Lineages in Humans. G3 (Bethesda)

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



| Locus | All alleles[a] | SM (R/NR)[b] | Pairwise (R/NR)[c] | BM pruned data set[d] | Codons | | |
|---|---|---|---|---|---|---|---|
| | | | | | Total | Non-ARS | ARS |
| *HLA-A* | 1193 | 144/107 | 138/104 | 93 | 340 | 292 | 48 |
| *HLA-B* | 1799 | 233/78 | 173/71 | 63 | 324 | 276 | 48 |
| *HLA-C* | 829 | 133/109 | 125/110 | 105 | 341 | 293 | 48 |

**Table 1** Number of alleles and codons for different data sets. a, included all available alleles in release 3.1.0, 2010-07-15., including possible recombinants; b, SM, data set used for site models, i.e, after selection of alleles with complete coding sequences; c, R/NR, with and without recombinants data sets; d, BM (branch models) pruned data set is the NR data set after prunning for alleles which do not cluster intra their respective allelic lineages (see Methods)

| Locus | $\omega^{a}$ | $\omega_{\texttt{inter}}{}^{b}$ | $\omega_{\texttt{intra}}{}^{c}$ | $2\Delta l^{d}$ | $\omega_{\texttt{internal}}{}^{e}$ | $\omega_{\texttt{terminal}}{}^{f}$ | $2\Delta l'$ |
|---|---|---|---|---|---|---|---|
| *HLA-A* | 1.84 | 2.03 | 1.68 | 0.06 | 2.35 | 1.39 | 0.49 |
| *HLA-B* | 0.99 | 1.16 | 0.73 | 0.71 | 1.2 | 0.69 | 0.97 |
| *HLA-C* | 1.89 | 4.14 | 1.19 | 2.61 | 4.91 | 0.95 | 7.36* |

**Table 2** Branch model $dN/dS$ estimations and LRT results (ARS data sets). * significance at 5%; Data sets after removal of recombinants (NR); a, $\omega$ estimate under model 0 (one for all branches); b, $\omega$ inter lineages; c, $\omega$ intra lineages d, negative log-likelihood difference between two nested models; e, $\omega$ for internal branches; f, $\omega$ for terminal branches

| Locus | $\omega^{a}$ | $\omega_{\texttt{inter}}{}^{b}$ | $\omega_{\texttt{intra}}{}^{c}$ | $2\Delta l^{d}$ | $\omega_{\texttt{int}}{}^{e}$ | $\omega_{\texttt{ter}}{}^{f}$ | $2\Delta l'$ |
|---|---|---|---|---|---|---|---|
| *HLA-A* | 0.53 | 0.40 | 0.77 | 2.8 | 0.39 | 0.95 | 4.57* |
| *HLA-B* | 0.42 | 0.40 | 0.55 | 0.34 | 0.39 | 0.66 | 0.86 |
| *HLA-C* | 0.50 | 0.39 | 0.79 | 3.97* | 0.38 | 0.92 | 5.27* |

**Table 3** Branch model $dN/dS$ estimations and LRT results (non-ARS data set). * significance at 5%; Data sets after removal of recombinants (NR); a, $\omega$ estimate under model 0 (one for all branches); b, $\omega$ inter lineages; c, $\omega$ intra lineages; d, negative log-likelihood difference between two nested models; e, $\omega$ for internal branches; f, $\omega$ for terminal branches

| | | non-ARS | | | | | ARS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Locus | Quantile[a] | $\overline{dN}$ | $\overline{dS}$ | $\overline{\omega}^{b}$ | $\overline{dN}/\overline{dS}$ | $dN > dS^{d}$ | $\overline{dN}$ | $\overline{dS}$ | $\overline{\omega}$ | $\overline{dN}/\overline{dS}$ | $dN > dS$ |
| *HLA-A* | | **0.02**[c] | **0.05** | **0.35** | **0.35** | **628(6.64%)** | **0.12** | **0.07** | **1.36** | **1.74** | **7364(77.90%)** |
| | 1 | 0.00 | 0.01 | 0.35 | 0.42 | 628 | 0.05 | 0.04 | 1.08 | 1.41 | 2132 |
| | 2 | 0.02 | 0.05 | 0.398 | 0.397 | 0 | 0.12 | 0.06 | 1.47 | 1.94 | 2347 |
| | 3 | 0.02 | 0.06 | 0.37 | 0.37 | 0 | 0.14 | 0.09 | 1.34 | 1.55 | 2316 |
| | 4 | 0.02 | 0.08 | 0.29 | 0.29 | 0 | 0.15 | 0.08 | 1.50 | 1.97 | 2339 |
| *HLA-B* | | **0.01** | **0.04** | **0.33** | **0.30** | **470(3.16%)** | **0.14** | **0.11** | **1.33** | **1.26** | **9908(66.59%)** |
| | 1 | 0.01 | 0.02 | 0.46 | 0.46 | 470 | 0.10 | 0.09 | 1.17 | 1.08 | 2405 |
| | 2 | 0.01 | 0.03 | 0.35 | 0.35 | 0 | 0.15 | 0.12 | 1.25 | 1.21 | 2460 |
| | 3 | 0.01 | 0.05 | 0.27 | 0.27 | 0 | 0.15 | 0.13 | 1.28 | 1.18 | 2229 |
| | 4 | 0.02 | 0.06 | 0.25 | 0.25 | 0 | 0.17 | 0.11 | 1.58 | 1.59 | 2814 |
| *HLA-C* | | **0.02** | **0.05** | **0.38** | **0.37** | **474(6.12%)** | **0.07** | **0.02** | **1.22** | **3.04** | **6514(84.05%)** |
| | 1 | 0.00 | 0.01 | 0.44 | 0.46 | 474 | 0.04 | 0.02 | 0.99 | 1.71 | 1303 |
| | 2 | 0.01 | 0.04 | 0.31 | 0.31 | 0 | 0.07 | 0.02 | 1.04 | 3.52 | 1791 |
| | 3 | 0.02 | 0.06 | 0.41 | 0.41 | 0 | 0.08 | 0.02 | 1.63 | 3.95 | 1810 |
| | 4 | 0.03 | 0.08 | 0.37 | 0.37 | 0 | 0.09 | 0.03 | 1.55 | 3.35 | 1682 |

**Table 4** Pairwise estimations for substitution rates (data sets prior to the removal of recombinants). a, quantiles of divergence ($dS_{\text{non-ARS}}$); b, average pairwise $dN/dS$; c, bold refers to the average pairwise values for each locus; d, percentages correspond to the proportion of pairs for which $dN > dS$ in relation to the total number of pairwise comparisons

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



| Data set | Substitution | Branch category | | | |
|---|---|---|---|---|---|
| | | intra | inter | terminal | internal |
| non-ARS | $N$ | 118.8 | 148 | 89.3 | 158.5 |
| | $S$ | 39.4 | 106 | 24.9 | 115.1 |
| | | $OR = 2.15$ | | $OR = 2.96$ | |
| ARS | $N$ | 172.7 | 230.7 | 144.3 | 291.2 |
| | $S$ | 18.5 | 17.5 | 17.4 | 17.7 |
| | | $OR = 0.71$ | | $OR = 0.21$ | |
| | | $p = 6.9 \times 10^{-3}*^{b}$ | | $p = 1.3 \times 10^{-4}*$ | |

**Table 5** Distribution of changes for ARS and non-ARS codons. Counts correspond to the total (combined) values for *HLA-A*, *-B* and *-C*; *significant at 1%; $N$, nonsynonymous change; $S$, synonymous change; intra, intra lineage; inter, inter lineage; terminal, terminal branches; internal, internal branches

| Set of SNPs | Var. Pos | Var. Pos. 1000g | MAF <= 0.1 | MAF > 0.1 | $\overline{MAF}$ |
|---|---|---|---|---|---|
| Intra | 68 | 29 | 5 | 24 | 0.15 |
| Inter | 88 | 55 | 12 | 43 | 0.14 |
| Overall | 156 | 84 | 17 | 67 | |

**Table 6** *HLA-A*: MAFs for SNPs in the 1000 Genomes dataset. Overall, set of variable positions considering all sequences in the site models dataset after removal of recombinants. Intra, subset of the 'Overall' set which is variable only within one allelic lineage for the locus. Inter, subset of the 'Overall' set which is variable within more than one allelic lineage. Var.Pos, set of all variable positions in the site models dataset. Var.Pos.1000g, subset of Var.Pos which is a SNP in the 1000G low coverage Phase I data. MAF, minor allele frequency. For details, see Methods.

| Set of SNPs | Var. Pos | Var. Pos. 1000g | MAF <= 0.1 | MAF > 0.1 | $\overline{MAF}$ |
|---|---|---|---|---|---|
| Intra | 44 | 24 | 6 | 18 | 0.30 |
| Inter | 59 | 38 | 8 | 30 | 0.39 |
| Overall | 103 | 62 | 14 | 48 | |

**Table 7** *HLA-B*: MAFs for SNPs in the 1000 Genomes dataset. MAFs for SNPs in the 1000 Genomes dataset. Overall, set of variable positions considering all sequences in the site models dataset after removal of recombinants. Intra, subset of the 'Overall' set which is variable only within one allelic lineage for the locus. Inter, subset of the 'Overall' set which is variable within more than one allelic lineage. Var.Pos, set of all variable positions in the site models dataset. Var.Pos.1000g, subset of Var.Pos which is a SNP in the 1000G low coverage Phase I data. MAF, minor allele frequency. For details, see Methods.

| Set of SNPs | Var. Pos | Var. Pos. 1000g | MAF <= 0.1 | MAF > 0.1 | $\overline{MAF}$ |
|---|---|---|---|---|---|
| Intra | 78 | 27 | 8 | 19 | 0.26 |
| Inter | 68 | 55 | 19 | 36 | 0.24 |
| Overall | 146 | 82 | 27 | 55 | |

**Table 8** *HLA-C*: MAFs for SNPs in the 1000 Genomes dataset. MAFs for SNPs in the 1000 Genomes dataset. Overall, set of variable positions considering all sequences in the site models dataset after removal of recombinants. Intra, subset of the 'Overall' set which is variable only within one allelic lineage for the locus. Inter, subset of the 'Overall' set which is variable within more than one allelic lineage. Var.Pos, set of all variable positions in the site models dataset. Var.Pos.1000g, subset of Var.Pos which is a SNP in the 1000G low coverage Phase I data. MAF, minor allele frequency. For details, see Methods.

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



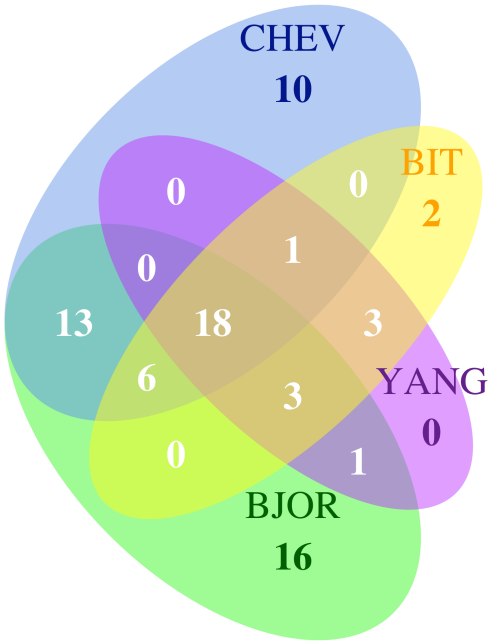


**Fig. 1** Overlap between two ARS classifications and two site models studies. BJOR and CHEV are ARS classifications (Bjorkman et al 1987; Chelvanayagam 1996); YANG is a list of codons with significant in *HLA* genes; BIT is the set of codons with from our SM (site models) approach (see Materials and Methods for details)

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



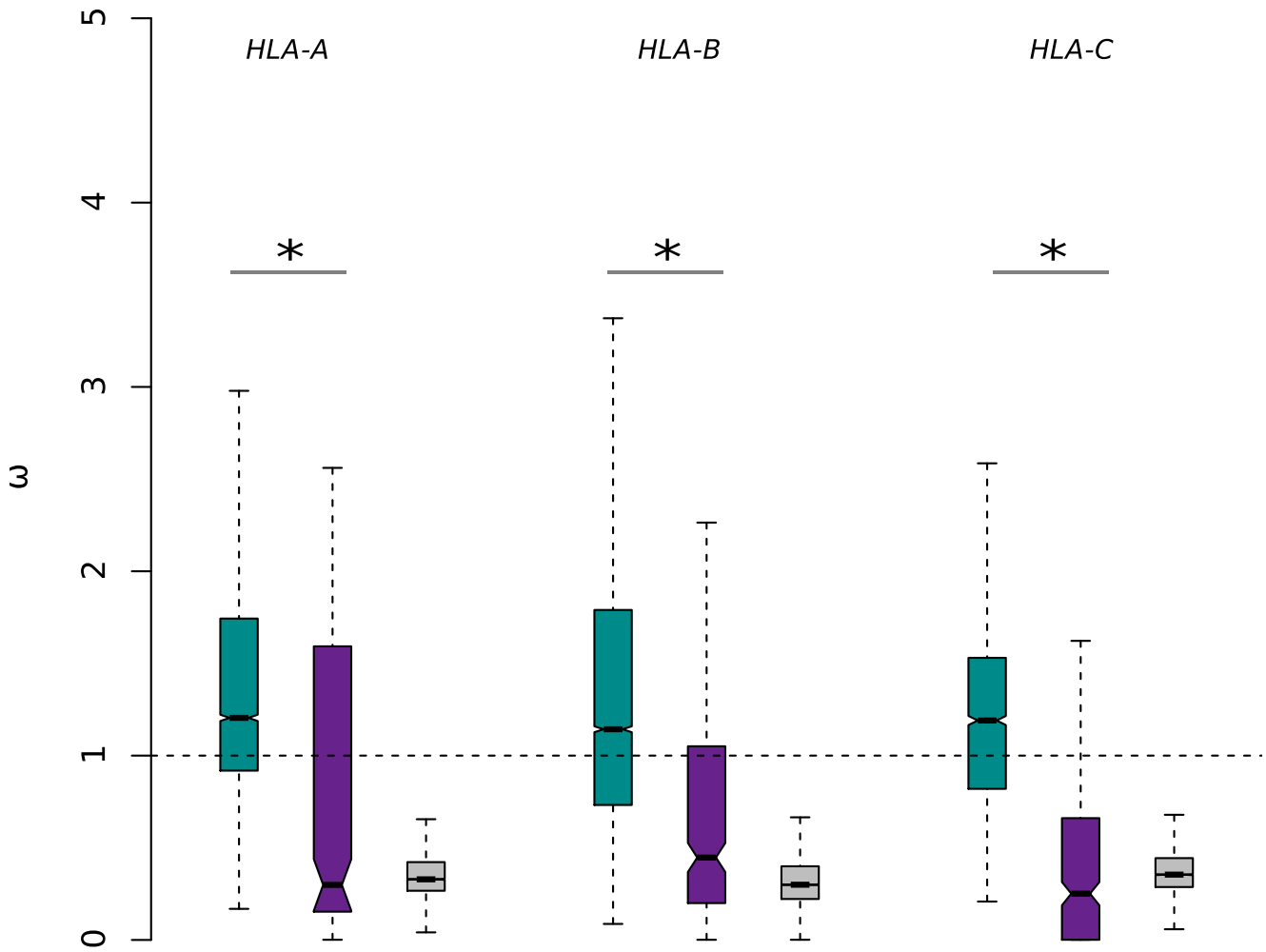


**Fig. 2** Pairwise estimates for intra-lineage and inter-lineage pairs of alleles. These results refer to ARS data sets prior to the removal of recombinants, for pairwise analyses; Green, inter-lineage; purple, intra-lineage; gray, non-ARS ; * significant difference between $\bar{\omega}$ (intra) and $\bar{\omega}$ (inter) ($p < 0.001$, Wilcoxon rank sum test)

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted October 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



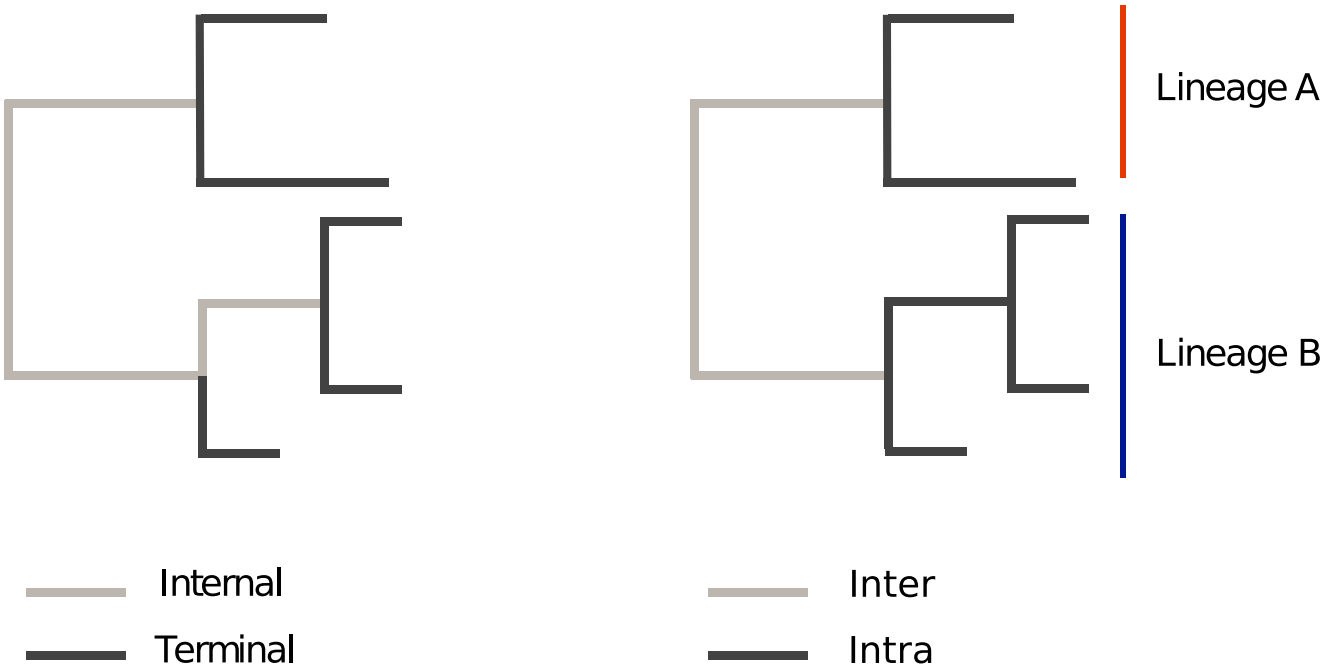


**Fig. 3** Schematic representation of the allelic phylogenies used in the branch models approach. Left: terminal *vs* internal branches; right: intra-lineage *vs* inter-lineage; For the branch models approach, we labeled branches of each tree (*HLA-A*, *-B* and *-C*) as "intra/inter" or "terminal/internal" and ran model 2 (CODEML), which allows for two independent $\omega$ values to be estimated, according to these labels