bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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 4.0 International license.

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

Variation of nonsynonymous/synonymous rate ratios at *HLA* genes over time and phylogenetic context

Bárbara Domingues Bitarello, Rodrigo dos Santos Francisco,
Diogo Meyer

Abstract Many *HLA* loci show an excess of nonsynonymous (dN) with respect to synonymous (dS) substitutions at codons of the antigen recognition site (ARS), a hallmark of adaptive evolution. However, it remains unclear how these changes are distributed over time and across branches of the *HLA* phylogeny. In particular, although *HLA* alleles can be assigned to functionally and phylogenetically defined groups ("lineages"), a test for differences in ω ($\omega = dN/dS$) within and between lineages is lacking. We analysed variation of ω across divergence times and phylogenetic contexts (placement of branches in the phylogeny).

We found a significant positive correlation between ω at ARS codons and divergence time, and that branches between lineages have higher ω than those within lineages. The excess of nonsynonymous changes between lineages attained significance when we used non-ARS codons to account for the fact that, even under purifying selection, ω is inflated for recently diverged alleles. Although less intensely selected, within-lineage variation at ARS codons bears evidence of selection, in the form of higher ω than those of non-ARS codons.

Our results show that ω ratios of class I *HLA* genes vary over time, and are higher in branches connecting alleles from distinct lineages. These results suggest that although within-lineage variation bears evidence of balancing selection, the between-lineage changes have been more intensely selected. Our findings indicate the importance of considering the effect of timescale when analysing ω values over a wide spectrum of divergences, and the value of using additional markers (in our case the tightly linked non-ARS codons) to account for the temporal dynamics of ω.

Keywords balancing selection, *HLA*, MHC, dN/dS, allelic lineages, antigen recognition site

Address: Departamento de Genética e Biologia Evolutiva, Universidade de 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 August 22, 2014. 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 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*) genes. Several findings suggest these genes 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 or amino-acids that make up the antigen recognition site (ARS), which determines which peptides the molecule can 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 *HLA* genes. The first is heterozygote advantage, which assumes that heterozygotes have higher fitness values because they are able to mount an immune response to a greater array of pathogens - an idea based on the observation that mice which are heterozygous for the MHC have increased immunological surveillance (Doherty and Zinkernagel 1975). Heterozygote advantage has received support from experiments in semi-natural populations of mice (Penn et al 2002), showing 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). A second form of balancing selection at MHC genes is negative frequency dependent 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 biologically compelling, most forms of summarizing genetic observation are incapable of differentiating this mode of selection from heterozygote advantage (Spurgin and Richardson 2010). A third model involves selection that is 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 *HLA* 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 *HLA* 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

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

robust to complications introduced by demographic history (Meyer and Thomson 2001; Hughes and Yeager 1998; Garrigan and Hedrick 2003). However, important questions regarding the properties of balancing selection at *HLA* genes remain unresolved. First, while most tests for selection have provided strong evidence for selection at *HLA* 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.

A second question concerns what type of variation is targeted by selection. 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 within the same lineage often differ by only minor structural variations in the ARS, while alleles in different lineages have functionally different ARS. Classical MHC genes have many alleles, which can be grouped to reflect phylogenetic relatedness and functional attributes (with these groups commonly referred to as "lineages"). Although nucleotide diversity within lineages exceeds genome-wide averages, between lineage diversity is substantially higher than within (Takahata and Satta 1998). This raises the question of whether variation within lineages is under a different mode and intensity of selection with respect to differences among lineages.

We address these questions by analysing the temporal and phylogenetic dynamics of dN/dS (or ω) for ARS codons at the class I (*HLA-A*, *-B* and *-C*) loci. To quantify divergence times among alleles we estimate dS from non-ARS codons, thus avoiding the issue of statistical non-independence between the two measures. Additionally, we use a phylogenetic approach to compare the intensity of selection on branches within and between lineages, and on terminal and internal branches. Throughout, we evaluate the effect of intragenic recombinant alleles on the analyses. Our pairwise comparisons of alleles shows that more divergent pairs show higher ω for ARS codons than closely related pairs of alleles. Our phylogenetic analyses support the hypothesis that selection is stronger in branches connecting different lineages, or which are internal to the phylogeny, provided that a bias toward overestimating ω for recent divergence is taken into account. Although positive selection is weaker within than between lineages, our findings show that there is statistical support for deviation from a regime of neutrality or purifying selection within lineages. We conclude that divergence within lineages is not neutral, and that between lineage divergence bears a stronger signature of selection than within lineage variation.

2 Materials and Methods

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

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 estimations and 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 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, non-ARS, consisting of the remaining codons (Table 1). Indels, null alleles, alleles encoding proteins with low cell surface expression (with the mutation inside or outside the coding region), alleles encoding proteins which are expressed as secreted molecules only and alleles with mutations that are putatively related to cell surface expression were removed from all downstream analyses. The non-ARS data sets were used for estimation of dS, used in the pairwise approach as an independent proxy for allelic divergence.

2.2 Recombination detection, clade filter and branch models

The complete alignments for each locus were used to generate NJ (Saitou and Nei 1987) trees in NEIGHBOR (PHYLIP package, Felsenstein 1989) using the F84 method, $k = 2$ and empirical base frequencies for the distance matrices in DNADIST (Felsenstein 1989). Intragenic recombinants were detected by applying RDP3 (Martin et al 2010) to the complete alignments and manual inspection.

2.2.1 RDP3

We chose RDP, Chimaera, Maxcho, GENECONV, BootScan and SiScan for recombination detection. 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. In all other cases we used the default values. Trace evidence of recombination were ignored and we considered as significant $p < 0.05$ in at least 3 of these methods. These specifications were chosen based on a test alignment we provided to the software, in which parental and daughter *HLA-B* sequences were known a priori. Following this initial procedure, we visually inspected the filtered alignments for the detection of additional recombinant sequences. After listing the recombinant sequences, these were removed from the pairwise, site models and branch models data sets, resulting in a “recombinant” (R) and “non-recombinant” (NR) data set for each locus (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

2.2.2 Clade filter

For the branch models, we used t (expected number of nucleotide substitutions per codon) matrices obtained in the pairwise analyses of the NR 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 (this information is known *a priori* from the IMGT/HLA annotation, and is thus independent of our trees) had to group together in a clade and alleles which did not group in such manner were manually "pruned" from trees to adjust to this "clade criterium". Tree branches were then labeled as "within" (w) lineages or "between" (b) lineages (or terminal/internal; figure 3) for the BM analyses. Table 1 summarizes the number of alleles used for each set of analyses.

2.2.3 Branch Models

With these pruned data sets we compared branch models 0 (one ω for all branches) and 2 (two or more categories of branches with independent ω) from CODEML *via* likelihood ratio tests (LRTs, see below). We provided CODEML with a topology based on the non-ARS data set, using branch lengths simply as starting points for ML estimation (fix_blength=1). For all CODEML analyses (SM, BM and pairwise), the Goldman and Yang (1994) model was used for estimation of substitution rates, 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. BM analyses were performed solely for NR (without recombinants) data sets (see tables 2 and 3) .

2.3 Pairwise approach

Pairs with $dN/dS > 5$ were considered NAs, as high values of dN/dS are mainly caused by near zero dS values and, therefore, would bias our results (Wolf et al 2009). Correlations between allelic divergence and omega values were computed by Pearson's correlation index. NA values were treated by casewise deletion. Significance for these correlations was evaluated via a in-house implementation of the Mantel Test. We repeated these Mantel tests using Spearman's and Kendall's correlation indexes (Online Resource 1, Table S13) and treating NA values with overall mean imputation. We obtained quantiles of the $dS_{\mathrm{non-ARS}}$ 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 ω values for "within" and "between" were evaluated via a Wilcoxon rank sum test (Figure 2).

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

2.4 Site models

For the SM approach, the clade filter (see above) 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 site models from CODEML to verify which codons have evidence for $\omega > 1$. M0 (one ratio) assumes the existence of only one ω ratio for all codons (it is the simplest model and used solely for an evaluation of consistency of the parameter estimates of the more complex models). 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 where $\omega_0 < 1$, and$0 < \omega < 1$, while M8 adds an extra category to M7, ω_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 codons classification proposed by Bjorkman et al. (1987) are referred to as BJOR, the peptide binding environments described in (Chelvanayagam 1996) are referred to as CHEV and the list of codons in *HLA* genes with evidence of $\omega > 1$ from Yang and Swanson (2002) is referred to as YANG (Online Resource 1, Table S9 and Figure 1). M1 *vs* M2 and M7 *vs* M8 were compared via likelihood ratio tests (see below).

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, only one was outside of the exons 2 and 3 range (codon 305) . Figure 1 shows the overlap between the BJOR and CHEV ARS classifications, the YANG set of codons and our approach.

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.

2.5 Likelihood Ratio Test (LRT)

When comparing two nested models, as in the SM (M1 and M2, M7 and M8) and BM (0 and 2) approaches (see above), the LRT test statistic is given by:

$$2\Delta l = 2(l_2 - l_0),$$

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

where l_2 and l_0 are ML values for the more parameter rich model (M8 or M2 in the SM approach, or model 2 in the BM approach). Degrees of freedom: 2 d.f for SM and 1 d.f for BM. It is expected that the use of a chi-square distribution for significance evaluation is a conservative approach (Yang 2006).

2.6 Breslow-Day test

In order to compare ARS and non-ARS codons with respect to the distribution of synonymous and nonsyonoymous 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 (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 within/between or terminal/internal branches for each locus, and for ARS and non-ARS codons (Table 5).

We defined the odds ratio (OR) as ${}^{N_{\mathtt{within}}}/_{S_{\mathtt{within}}}/{}^{N_{\mathtt{between}}}/_{S_{\mathtt{between}}}$ and used a Breslow-Day test for homogeneity of odds ratios to test the hypothesis that contingency tables from ARS and non-ARS codons have the same odds ratio. We applied the same test to internal/terminal branches.

3 Results

3.1 Evidence for selection and assessment of recombination

Before investigating how ω varies over time and phylogenetic context, we tested (a) whether selection is detectable in our data set with pairwise comparisons and phylogenetic 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.

We quantified the mean pairwise dN/dS ($\overline{\omega}$), and found $\overline{\omega}_{\mathtt{ARS}} > 1$ for all loci (Table 4). We used the non-ARS codons from the same sequences as an internal control, and found that $\overline{\omega}_{\mathtt{ARS}}$ is 3.9 (*HLA-A*), 4.0 (*HLA-B*) and 3.2-fold (*HLA-C*) greater than $\overline{\omega}_{\mathtt{non-ARS}}$ (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 $\overline{\omega}_{\mathtt{ARS}} > \overline{\omega}_{\mathtt{non-ARS}}$ 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,

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

Table 4). Qualitatively similar results were obtained when we computed the ratio of mean substitution rates, $\overline{dN}/\overline{dS}$ (Table 4). These results are robust to the presence of recombinants (Online Resource 1, Table S1).

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; $p < 0.01$, likelihood ratio test). 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 inferred to be under selection (using site models from CODEML, from here on referred to as "SM") and those defined as ARS by structural analyses of *HLA* molecules. We defined a set of ARS codons based on the work of Chelvanayagam (1996), which identified "peptide binding environments", i,.e, the amino acid residues in a fixed neighborhood of the peptide binding residues in known crystal structure complexes, providing a less restrictive description of the antigen binding sites (Chelvanayagam 1996). 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) 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 *HLA* (Figure 1 in main text and Online Resource 1, Table S9). This provides a highly significant association between the ARS and positively selected sites for all loci ($p < 10^{-11}$, chi-square test). There is extensive overlap between the two ARS classifications (Bjorkman et al 1987; Chelvanayagam 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).

Our results show that: (a) the pairwise and phylogenetic site models methods implemented by CODEML strongly support adaptive evolution on the ARS codons of *HLA* loci - as also described by Yang and Swanson (2002) through branch models; (b) there is an enrichment of codons with $\omega > 1$ in the CHEV set of codons (Online Resource 1, Table S9), supporting the use of this classification for our study; (c) although the overall results were robust to the presence of recombinants - as indicated by simulation studies Anisimova et al (2003) -, the estimated values for ω 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.

3.2 The time-dependence of ω at *HLA* 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

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

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 $dS_{\text{non-ARS}}$ (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 ($dS_{\text{non-ARS}}$), the resulting $\omega_{\texttt{ARS}}$ 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 ω values for pairwise comparisons at *HLA-A*, *-B*, and *-C*).

Pairwise estimates show that $\omega_{\texttt{ARS}}$ increases as a function of divergence time (Table 4). Indeed, ω_{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 (Online Resource 1, Table S13; $r_{\texttt{HLA-A}} = 0.25$, $p < 0.001$; $r_{\texttt{HLA-B}} = 0.12$, $p < 0.001$; $r_{\texttt{HLA-C}} = 0.19$, $p < 0.001$) and were robust to different correlation measures (Online Resource 1, Table S13). We also compared the pairwise ω for pairs of alleles within and between lineages (Figure 2). For all loci, the median value of $\omega_{\texttt{ARS}}$ is higher than 1 for the between lineage contrasts, and lower than 1 for the within lineage contrasts, and the distribution of ω is significantly higher for between 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 ω 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 ω for all branches to one that estimates two values of ω (between or within; terminal or internal; see Figure 3 and Methods). For the branch models, the tree topology was obtained from the non-ARS (NR) data set. For all loci we found higher values of $\omega_{\texttt{ARS}}$ for branches which are between lineages, than for branches within lineages, although significance was not attained for these tests (Table 2). For the “internal-terminal” contrast we found higher $\omega_{\texttt{ARS}}$ for internal branches and there was statistical support for higher ω for internal than for terminal branches for *HLA-C* (Table 2).

Our results show that both pairwise comparisons and phylogenetic branch models indicate a heterogeneity of ω throught the diversification of *HLA* alleles, with higher ω values asssociated to contrasts between more divergent alleles (pairwise approach) or to branches connecting different lineages or that are internal to the phylogeny (phylogenetic branch models), although the difference was not significant for the “within-between” contrasts (Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

3.3 Significantly more nonsynonymous changes between lineages at ARS codons

Our study estimates ω for allele pairs or branches sampled within a single species, and over varying timescales (recent to remote divergence). Both these features imply in possible biases to the estimation of ω, which we now discuss.

Kryazhimskiy and Plotkin (2008) used analytical and simulation approaches to show that under positive selection the behavior of ω within a single population is not a monotonic function of the intensity of selection, so that ω within a population can be low, even under positive selection. This occurs because the scenario they explored assumes that recent positive selection fixes a favourable nonsynonymous variant (thus decreasing ω). However, this scenario clearly does not apply to *HLA* genes, were balancing selection maintains multiple nonsyonynous polymorphisms simultaneously segregating within a population, contributing to $\omega > 1$.

Another complexity in the interpretation of ω arises from that fact that many studies have shown that genes under purifying selection show surprisingly high ω (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 lineages could result in inflated ω values, explaining the modest differences in ω within and between lineages seen in the phylogenetic analyses (Tables 2 and 3). To explore this issue further, we took advantage of the availability of non-ARS codons from the same set of sequences, which provide a convenient internal control, given their tight linkage to the ARS codons (and hence similar within *vs* between-lineage phylogenetic structure). We find that non-ARS codons have larger values of ω within than between lineages, and for terminal *versus* internal brances ($p < 0.05$ for *HLA-A* in the within *versus* between lineage contrast, and for *HLA-A* and *HLA-C* in the tips *versus* internal contrast; LRT; Table 3). This distribution of ω 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 ω (Kryazhimskiy and Plotkin 2008).

To formally test whether ARS and non-ARS codons have a different distribution of synonymous and nonsynonymous changes within and between lineages (or for internal and terminal branches) we employed a contingency table approach implemented and described by Templeton (1996). We used the inferred number of synonymous (S) and nonsynonymous (N) changes on each branch 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). We defined the odds ratio (OR) as $OR = {}^{N_{\mathtt{within}}}/_{S_{\mathtt{within}}}/{}^{N_{\mathtt{between}}}/_{S_{\mathtt{between}}}$. For all loci, we find that $OR > 1$ for non-ARS codons (proportionally more nonsynonymous changes within lineages) and $OR < 1$ for ARS codons (proportionally more nonsynonymous changes between lineages), as shown in Table 5. This finding is consistent with the maximum likelihood estimates of ω for branches (Tables 2 and 3), and the increased pairwise ω for alleles between lineages, relative to within

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

(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). Using a Breslow-Day test for homogeneity of odds ratios (Table 5), we rejected the null hypothesis that contingency tables from ARS and non-ARS codons have the same OR ($p - value = 0.0069$). Our analysis with the contrast between terminal and internal branches ($OR = {}^{N_{\texttt{terminal}}}/_{S_{\texttt{terminal}}} / {}^{N_{\texttt{internal}}}/_{S_{\texttt{internal}}}$) showed the same pattern, with proportionally more nonsynonymous changes in internal branches for ARS codons ($p - value = 0.00013$; Table 5). Finally, although there is evidence for an excess of nonsynonymous changes between lineages (or for terminal branches) for ARS codons, there is also an enrichment for nonsynonymous changes within lineages for ARS codons, when compared to non-ARS codons ($P < 0.001$; Fisher's exact test).

4 Discussion

Our study documents a positive correlation between ω values and the degree of divergence between allele pairs. This result is supported by phylogenetic analyses, which show higher ω values for branches connecting lineages, or branches which are internal to the phylogeny. A heterogenous nonsynonymous substitution rate was also reported in a recent study (Yasukochi and Satta 2014), which found that dN for ARS codons is not linearly correlated with divergence time in classical *HLA* loci. 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 lineages, possibly as a consequence of continuous selective pressure by a specific pathogen. In the present study our goal is to explicity test for heterogeneity in the ω ratios over *a priori* defined groups of alleles or timescales. 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. Our findings indicate that long-term balancing selection has resulted in an enrichment for adaptive changes between *HLA* lineages, with recent and within lineage changes showing proportionally weaker signatures of molecular adaptation.

The finding that $\omega_{\texttt{ARS}}$ between lineage is greater than within is consistent with the divergent allele advantage model, according to which heterozygotes for more divergent alleles have higher fitnes than those carrying similar alleles. Under this model, excess of nonsynonymous changes between *HLA* allelic lineages would be expected, which is a result we have shown for the ARS data set.

Although our results show that ω within lineages is lower than between, our phylogenetic results show that within lineage variation does not behave neutrally, and ω for ARS codons within lineages is still higher than ω for non-ARS codons within lineages - a scenario which is expected to inflate ω Rocha et al (2006). This result suggests that within lineage variation cannot be viewed as neutral.

Recently several papers have drawn attention to the effects of divergence times on dN/dS estimation, 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 between lineages) is conservative in the light of these findings, which predict decreased ω 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 (thanks

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

to the 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 ω, with increased evidence of selection between lineages, exactly the opposite to that seen for non-ARS codons.

An important caveat to this interpretation is that the temporal dynamics of dN/dS appear to be sensitive to the selective regime which is assumed to be operating. Thus, while several authors have shown that under purifying selection we expect increased dN/dS at low divergence, 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. Thus, simulation and analytical treatment of the temporal dynamics of dN/dS under balancing selection, which currently has not been explored, will be an additional source of information to intepret the dynamics we have described for *HLA* class I genes.

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 out all data manipulation and statistical analyses and drafted the manuscript. RDF participated in the detection of recombination in the data sets and participated in the interpretation of the site models analyses. DM conceived of the study, participated in its design, coordination and in the statistical analyses and helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements The authors thank Aida M. Andrés and Richard Single for helpful discussions. This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (2009/09127-8; 08/56502-6 and 11/12500-2 to B.D.B) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (152676/2011-2 to BDB, 142130/2009-5 to RSF and 308960/2009-2 to DM).

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

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

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

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

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

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

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):430–41

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. Genetics 144(3):1263–1270

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):1107–18

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 August 22, 2014. 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 4.0 International license.

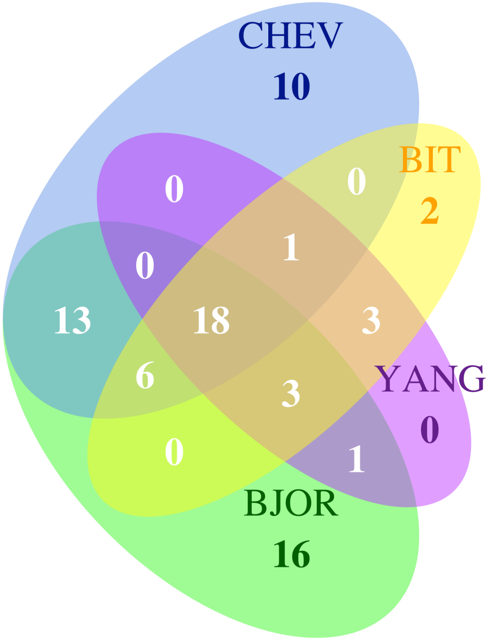

Fig. 1 Overlap between two ARS classifications and two site models studies. BJOR and CHEV are ARS classifications (Bjorkman et al 1987; Chelvanayagam 1996); YANG is a list of codons with significant in *HLA* genes; BIT is the set of codons with from our SM (site models) approach (see Materials and Methods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

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) prunned data set is the NR data set after prunning for alleles which do not cluster within their respective allelic lineages (see Methods)

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

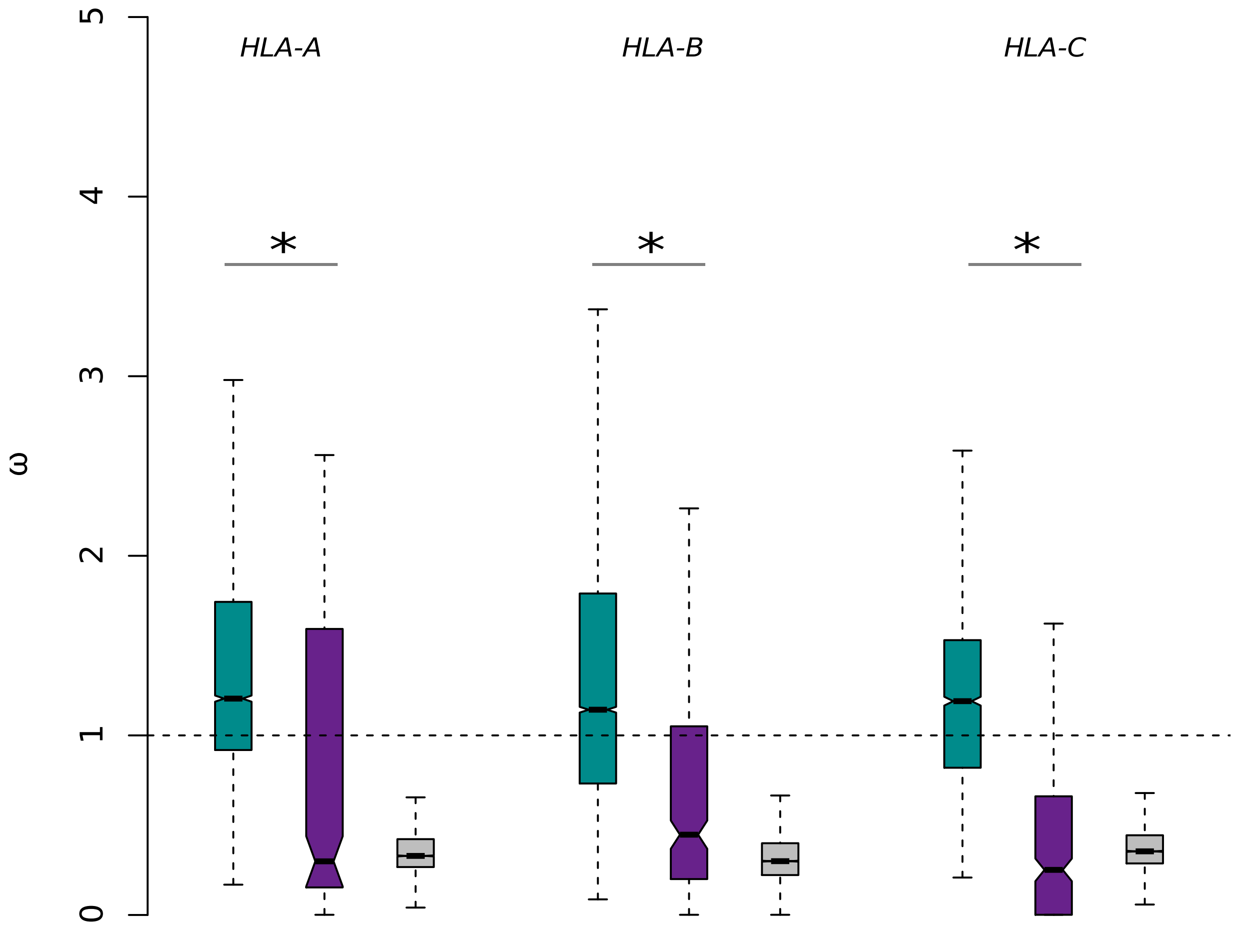

Fig. 2 Comparisons between within and between lineages for ARS codons. These results refer to data sets prior to the removal of recombinants; Green, between; purple, within; gray, non-ARS (all comparisons); * significant difference between $\bar{\omega}$ (within) and $\bar{\omega}$ (between) ($p < 0.001$, Wilcoxon rank sum test)

Locus	ω^{a}	$\omega_{\mathtt{b}}{}^{b}$	$\omega_{\mathtt{w}}{}^{c}$	$2\Delta l^{d}$	$\omega_{\mathtt{i}}{}^{e}$	$\omega_{\mathtt{t}}{}^{f}$	$2\Delta l'$
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, ω estimate under model 0 (one for all branches); b, ω between lineages; c, ω within lineages d, negative log-likelihood difference between two nested models; e, ω for internal branches; f, ω for terminal branches

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

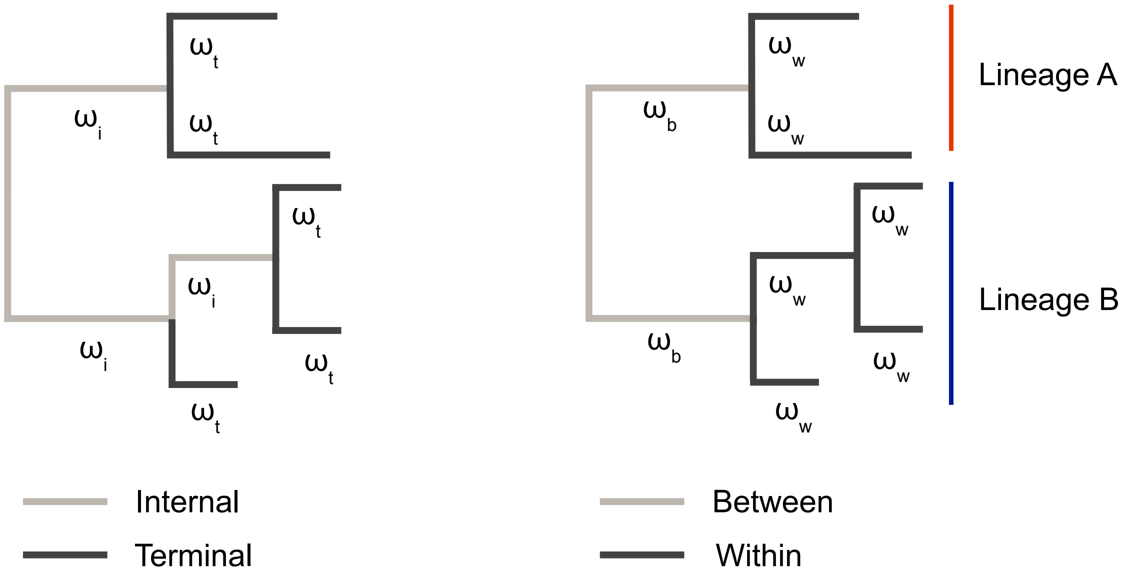

Fig. 3 Schematic representation of the allelic phylogenies used in the branch models approach. Left: terminal *vs* internal branches; right: within *vs* between lineages; For the branch models approach, we labeled branches of each tree (*HLA-A*, *-B* and *-C*) as "within/between" or "terminal/internal"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

Locus	ω^{a}	$\omega_{\mathtt{b}}{}^{\mathrm{b}}$	$\omega_{\mathtt{w}}{}^{\mathrm{c}}$	$2\Delta l^{\mathrm{d}}$	$\omega_{\mathtt{i}}{}^{\mathrm{e}}$	$\omega_{\mathtt{t}}{}^{\mathrm{f}}$	$2\Delta l'$
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, ω estimate under model 0 (one for all branches); b, ω between lineages; c, ω within lineages; d, negative log-likelihood difference between two nested models; e, ω for internal branches; f, ω for terminal branches

bioRxiv preprint doi: https://doi.org/10.1101/008342; this version posted August 22, 2014. 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 4.0 International license.

		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

Data set	Substitution	Branch category			
		w	b	t	i
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; w, within; b, between; t, terminal; i, internal