A sequence-based approach demonstrates that balancing selection in classical human leukocyte antigen (HLA) loci is asymmetric

Paola G. Bronson1,∗,†, Steven J. Mack2, Henry A. Erlich3 and Montgomery Slatkin1

1Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA, 2Genetics, Children’s Hospital Oakland Research Institute, Oakland, CA 94609, USA and 3Human Genetics, Roche Molecular Systems, Pleasanton, CA 94588, USA

Received July 2, 2012; Revised September 26, 2012; Accepted October 4, 2012

Balancing selection has maintained human leukocyte antigen (HLA) allele diversity, but it is unclear whether this selection is symmetric (all heterozygotes are comparable and all homozygotes are comparable in terms of fitness) or asymmetric (distinct heterozygote genotypes display greater fitness than others). We tested the hypothesis that HLA is under asymmetric balancing selection in populations by estimating allelic branch lengths from genetic sequence data encoding peptide-binding domains. Significant deviations indicated changes in the ratio of terminal to internal branch lengths. Such deviations could arise even if no individual alleles present a strikingly altered branch length (e.g. if there is an overall distortion, with all or many terminal branches being longer than expected). DQ and DP loci were also analyzed as haplotypes. Using allele frequencies for 419 distinct populations in 10 geographical regions, we examined population differentiation in alleles within and between regions, and the relationship between allelic branch length and frequency. The strongest evidence for asymmetrical balancing selection was observed for HLA-DRB1, HLA-B and HLA-DPA1, with significant deviation ($P \leq 1.1 \times 10^{-4}$) in about half of the populations. There were significant results at all loci except HLA-DQB1/DQA1. We observed moderate genetic variation within and between geographic regions, similar to the rest of the genome. Branch length was not correlated with allele frequency. In conclusion, sequence data suggest that balancing selection in HLA is asymmetric (some heterozygotes enjoy greater fitness than others). Because HLA polymorphism is crucial for pathogen resistance, this may manifest as a frequency-dependent selection with fluctuation in the fitness of specific heterozygotes over time.

INTRODUCTION

The human major histocompatibility complex (MHC) is an important gene-dense region spanning over 4 Mb of DNA on chromosome 6p21.3 (1). The MHC exhibits the strongest linkage disequilibrium observed in the human genome and almost half of expressed loci in this region have functions related to immune activation and response (2–5). These include the class I and II human leukocyte antigen (HLA) membrane glycoproteins, which play a critical role in activating the adaptive immune response. HLA classes I and II are expressed on antigen-presenting cells and present intracellular (viral) and extracellular (bacterial) peptides, respectively, for T-cell recognition. HLA loci are highly polymorphic. For example, HLA-B has over 2000 known alleles, encoding over 1500 unique sequences. Evidence suggests that in addition to novel mutations, intergenic and interallelic recombination and gene conversion are responsible for generating novel HLA alleles (6–9).

Symmetric overdominance in HLA

HLA has been shaped both by demographic history and selection, and particularly, balancing selection has been shown to
maintain HLA diversity (10–12). It has long been proposed that balancing selection could be pathogen-driven through a fitness advantage conferred by the ability to present a wider range of antigens to T-cells (13,14). Balancing selection is an evolutionary process that maintains allelic diversity over time. The conventional model of balancing selection in the MHC (symmetric overdominance) assumes that all HLA alleles are equivalent in the sense that every heterozygote has the same fitness and every homozygote has the same, somewhat lower, fitness (15). The underlying assumption is that heterozygotes may have greater fitness than homozygotes because their immune systems can recognize a wider range of pathogens. Hedrick and Thomson (16) concluded that balancing selection inferred from the Ewens–Watterson test must be symmetric. However, this purely symmetrical model is over-simplified and the reality is most likely closer to some combination of heterozygote advantage and selective advantage in specific geographic regions. The equality of fitnesses of all heterozygotes and of all homozygotes in the model of symmetric overdominance is assumed because there is no definitive evidence to the contrary despite the many allele-specific disease associations known for HLA loci. This conventional model is purely for convenience and unlikely to be valid. In reality, the fitness of particular homozygotes or heterozygotes is likely to vary, due to the distribution of pathogens and allele-specific disease associations that pertain to particular environmental and behavioral contexts.

Evidence against symmetric overdominance

HLA class I diversity is higher in populations that have been subjected to a higher diversity of pathogens (17). In addition, individuals heterozygous at HLA loci are reported to exhibit increased survival and resistance to pathogens, but these studies may not always have accounted for allele frequencies (18–22). Furthermore, a recent theoretical model suggests that overdominance could only have maintained HLA diversity under high degenerate pathogen recognition and low intersection advantage (23). Degenerate pathogen recognition refers to the ability of different alleles at a given HLA locus to encode multiple molecules that recognize the same pathogen (23). Intersection advantage is the selective fitness provided by carrying two different alleles that recognize the same pathogen compared with carrying only one allele that recognizes that pathogen (23). Although degenerate pathogen recognition is probably high, evidence suggests that intersection advantage is unlikely to be low (23–28).

Other mechanisms of balancing selection in HLA

Negative frequency-dependent selection has also been proposed as a model of balancing selection that could maintain HLA diversity (15,29). Also referred to as rare-allele advantage or diversifying frequency-dependent selection, it assumes that rare alleles may provide higher fitness through increased resistance to novel pathogens, and that the frequencies of these rare alleles will increase as exposure to the pathogen increases in the population. This cycle is driven by host–pathogen coevolution. Allele-based fitness functions indicate that frequency-dependent selection is the driving force in maintaining HLA diversity (30–32). However, frequency-dependent selection and overdominance are not mutually exclusive. Furthermore, HLA diversity could potentially be maintained by fluctuating selection, where varying spatial and temporal patterns of the frequency and strains of pathogens create a short-term directional selection that is independent of host–pathogen coevolution (33,34).

Our objective was to apply a sequence-based approach to test the hypothesis that HLA class I and class II alleles are under symmetrical balancing selection in single populations (Tables 1 and 2). In addition, we tested the hypotheses that the branch lengths of HLA alleles are correlated with allele frequency within populations, and that the genetic variation of HLA alleles within and between geographical regions is greater than that observed in the rest of the genome.

RESULTS

A genealogy that significantly deviates from the expected shape \( P \leq 1.1 \times 10^{-5} \) indicated changes in the ratio of terminal to internal branch lengths, compared with a neutral genealogy that is consistent with symmetrical balancing selection (Fig. 1), and in these cases indicates asymmetrical balancing selection. Such deviations could plausibly arise even if no individual alleles presented a strikingly altered branch length (e.g. if there is an overall distortion, with all or many terminal branches being longer than expected). Some HLA alleles had a significantly higher ratio of terminal to internal branch lengths than expected under symmetrical balancing selection (Fig. 2 and Table 2). Over all populations, this phenomenon was most pronounced for the HLA-B and HLA-DRB1 loci, with \( P \leq 1.1 \times 10^{-5} \) in over half of populations. At the HLA-DPA1 and HLA-C loci, over a third of the populations exhibited \( P \leq 1.1 \times 10^{-5} \). Conversely, very few of the populations displayed genealogies that deviated from expected topologies for the HLA-DQA1, HLA-DQB1 and HLA-DPB1 loci. At the HLA-A locus, a 10th of the populations demonstrated significant deviation. There were significant deviations from symmetric balancing selection at all loci except HLA-DQB1/DQA1. In particular, consider the HLA-B locus in the Northern Irish population, \( R_{SD} = 6.92 \ (P = 1 \times 10^{-5}) \).

### Table 1. Analysis of distinct populations

<table>
<thead>
<tr>
<th>Geographical regions (N = 10)</th>
<th>Number of populations analyzed (N = 419)</th>
<th>Number of individuals included in populations (N = 50,303)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSA 34</td>
<td>4198</td>
<td></td>
</tr>
<tr>
<td>NAF 18</td>
<td>1489</td>
<td></td>
</tr>
<tr>
<td>EUR 88</td>
<td>14,708</td>
<td></td>
</tr>
<tr>
<td>SWA 37</td>
<td>3038</td>
<td></td>
</tr>
<tr>
<td>SEA 56</td>
<td>6491</td>
<td></td>
</tr>
<tr>
<td>OCE 61</td>
<td>3596</td>
<td></td>
</tr>
<tr>
<td>AUS 4</td>
<td>8695</td>
<td></td>
</tr>
<tr>
<td>NEA 63</td>
<td>10,696</td>
<td></td>
</tr>
<tr>
<td>SAM 38</td>
<td>3382</td>
<td></td>
</tr>
</tbody>
</table>

This study analyzed 419 distinct populations, from 10 geographical regions worldwide, with publicly available HLA allele frequency information available (11,59).
Table 2. Analysis of HLA alleles and haplotypes

<table>
<thead>
<tr>
<th>Locus</th>
<th>Location on Chr 6p21.3 (base pair, Build 37)</th>
<th>Number of sequences (N = 1112)</th>
<th>Number of base pair analyzed</th>
<th>Number of genealogies (N = 1223)</th>
<th>Percent of genealogies with significant deviation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-A</td>
<td>29 910 247–29 913 661</td>
<td>155</td>
<td>610</td>
<td>129</td>
<td>11.6</td>
</tr>
<tr>
<td>HLA-C</td>
<td>31 236 529–31 239 855</td>
<td>70</td>
<td>578</td>
<td>92</td>
<td>34.8</td>
</tr>
<tr>
<td>HLA-B</td>
<td>31 321 649–31 324 989</td>
<td>286</td>
<td>557</td>
<td>120</td>
<td>61.7</td>
</tr>
<tr>
<td>HLA-DRB1</td>
<td>32 546 546–32 557 562</td>
<td>179</td>
<td>274</td>
<td>224</td>
<td>53.6</td>
</tr>
<tr>
<td>HLA-DQA1</td>
<td>32 605 183–32 611 429</td>
<td>8</td>
<td>249</td>
<td>167</td>
<td>4.2</td>
</tr>
<tr>
<td>HLA-DQB1</td>
<td>32 627 657–32 634 466</td>
<td>30</td>
<td>270</td>
<td>177</td>
<td>2.3</td>
</tr>
<tr>
<td>HLA-DPA1</td>
<td>33 032 346–33 048 555</td>
<td>9</td>
<td>246</td>
<td>34</td>
<td>44.1</td>
</tr>
<tr>
<td>HLA-DPB1</td>
<td>33 043 703–33 057 473</td>
<td>64</td>
<td>264</td>
<td>123</td>
<td>2.4</td>
</tr>
<tr>
<td>HLA-DQA1/DQB1</td>
<td>169 519 130 7</td>
<td>7.4</td>
<td>130</td>
<td>7.4</td>
<td></td>
</tr>
<tr>
<td>HLA-DPA1/DPB1</td>
<td>142 510 27</td>
<td>27</td>
<td>0</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

This study examined 1001 HLA alleles (unique exonic sequences encoding peptide-binding domains) and 311 HLA haplotypes in 419 distinct populations, using publicly available sequence information for each allele (60). Because the data were incomplete, there were less than 419 populations for each locus. The threshold for statistical significance after correcting for multiple testing was $P \leq 1.1 \times 10^{-4}$.

Figure 1. (A) Example of a genealogy that did not deviate from the expected shape: HLA-DRB1 in the Tolai population from Papa New Guinea ($R_{SD} = 2, P = 0.12$). (B) Example of a genealogy that significantly deviated from the expected shape: HLA-DRB1 in the Zuni population from NAM ($R_{SD} = 4.2, P = 2 \times 10^{-7}$). $R_{SD}$ represents the proportion of allelic terminal branch lengths in the population. $P$-values indicate the significance of the deviation from the expected $R_{SD}$ value under the null (1.25). HLA, human leukocyte antigen.
Figure 2. Distribution of $P$-values quantifying the significance of the deviation of observed $R_{SD}$ values from those expected in a model of symmetrical balancing selection: (A) HLA-A, (B) HLA-C, (C) HLA-B, (D) HLA-DRBI, (E) HLA-DQA1, (F) HLA-DQB1, (G) HLA-DPA1, (H) HLA-DPB1, (I) HLA-DQA1/DQB1 haplotype, (J) HLA-DPA1/DPB1 haplotype. The threshold for statistical significance after correcting for multiple testing was $P \leq 1.1 \times 10^{-4}$. HLA, human leukocyte antigen.
Results were similar for the HLA-C ($R_{SD} = 6.71, P = 1 \times 10^{-5}$) and HLA-DRB1 ($R_{SD} = 7.84, P = 1 \times 10^{-5}$) loci in this population. Weaker evidence was observed at the HLA-A locus ($R_{SD} = 3.21, P = 1.5 \times 10^{-2}$) in this population. There were some regional trends in the proportion of $R_{SD}$ values that were significant for certain loci. Populations in Sub-Saharan Africa (SSA), North Africa (NAF), Europe (EUR) and Northeast Asia (NEA) tended to have a larger proportion of significant $P$-values at HLA-DRB1. Populations in SSA, NAF, EUR and NEA tended to have a larger proportion of significant $P$-values at HLA-B. Populations in EUR and SWA tended to have a larger proportion of significant $P$-values at HLA-C. Populations in South America (SAM) tended to have a larger proportion of significant $P$-values at HLA-DPA1, HLA-DPB1, HLA-DQA1 and HLA-DQB1.

Branch length was not correlated with allele frequencies in populations ($r^2 < 0.10$, see Supplementary Material, Table S1 and Fig. S1). Population differentiation between regions and within regions was generally low to moderate ($F_{ST} < 0.15$, Fig. 3 and Supplementary Material, Table S2).

**DISCUSSION**

Results from this study provide strong evidence supporting different levels of fitness conferred by different HLA alleles in most populations, with the exception of HLA-DQB1/DQA1. This evidence is strongest for the HLA-B locus, which is the most polymorphic HLA locus, both worldwide (with 2123 distinct alleles identified) and in this study (286 alleles reported). Among HLA loci, the reported within-population variation is also highest at HLA-B (with values of 92.5 and 92.8%) (10,35), and we observed a mean Total $F_{ST}$ value of up to 0.217 (at the HLA-B*07:04 allele) for this locus. This extensive molecular diversification may in part be due to asymmetric balancing selection favoring genotypic novelty over the course of the expansion of the human population.

It is important to note that HLA allele frequency data show a strong influence of population history rather than selection (36). Results from the current empirical study provide evidence for asymmetrical balancing selection, but identifying when this asymmetry arose is beyond the scope of this study. The asymmetry reflects the whole history of the alleles. This paper does not conclude that the different degrees of asymmetry found between populations are independent of one another. They partly reflect the ancestral history of humans and even primate ancestors.

$R_{SD}$ estimates have been shown to be significantly biased upward in distance-based genealogies when the per-base-pair recombination rate is $>0.008$ (37). Several recent genome-wide estimates of recombination rates present values of HLA in ranges compatible with applying the $R_{SD}$ statistic to test deviation from expectations under a model of symmetric overdominance. For example, there are no recombination hotspots located in the classical HLA loci, and the recombination rates in the classical HLA loci are very low (38). It has been reported that there is no recombination in HLA-C, HLA-DQA1 and HLA-DQB1, and that the recombination in HLA-DRB1 is very small (38). The MHC region contains short recombination hotspots separated by large recombination coldspots: 80% of the recombination in the MHC region occurs in 10% of the sequence, in contrast to the rest of the genome, where 50% of the recombination occurs in <10%
of the sequence (39). However, a previous study of multiallelic balancing selection with recombination conducted simulations that indicated that even very low levels of recombination may have the potential to bias genealogy shape (40). Intragenic recombination and gene conversion have been previously reported in HLA (41,42). Therefore, it is plausible that the asymmetry in the trees observed in results from the current study may be partly due to recombination. Although recombination plays a role in the results of this study, selection may also have been important.

Overall, our results are in concordance with recent studies of selection at the HLA loci. A large meta-analysis of HLA allele frequencies in 70,000 individuals recently confirmed the role of balancing selection in maintaining HLA variation in the classical HLA loci (11). HLA-DQA1 and HLA-C showed the strongest evidence of balancing selection, whereas HLA-DPB1 was compatible with a neutral genealogy (11). A recent study of sequence divergence in 23,500 individuals supported these findings, and also suggested that asymmetrical balancing selection may be present (12). Balancing selection has been inferred for the HLA-A and HLA-B loci in previous studies as well (38).

There is limited evidence from studies of humans to support the role of asymmetrical balancing selection in HLA evolution. The intense evolutionary arms race between the human immunodeficiency virus (HIV) and humans is an interesting case of HLA evolution (43). The HLA A2/6802 supertype (HLA-A*02:02, HLA-A*02:05, HLA-A*02:14 and HLA-A*68:02) has been shown to be associated with negative recombination in HIV-exposed East African female commercial sex workers and infants born to HIV-positive mothers (44,45).

Within Region F

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele</th>
<th>Regional Total $F_{ST}$</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-A</td>
<td>68:02</td>
<td>0.35</td>
<td>SAM</td>
</tr>
<tr>
<td></td>
<td>24:02</td>
<td>0.29</td>
<td>SEA</td>
</tr>
<tr>
<td></td>
<td>57:01</td>
<td>0.27</td>
<td>SEA</td>
</tr>
<tr>
<td>HLA-B</td>
<td>07:04</td>
<td>0.40</td>
<td>NEA</td>
</tr>
<tr>
<td></td>
<td>37:01</td>
<td>0.27</td>
<td>SEA</td>
</tr>
<tr>
<td></td>
<td>39:02</td>
<td>0.26</td>
<td>SAM</td>
</tr>
<tr>
<td>HLA-DRB1</td>
<td>04:11</td>
<td>0.43</td>
<td>NAM</td>
</tr>
<tr>
<td></td>
<td>12:02</td>
<td>0.29</td>
<td>SAM</td>
</tr>
<tr>
<td></td>
<td>04:04</td>
<td>0.25</td>
<td>SEA</td>
</tr>
</tbody>
</table>

Table 3. HLA alleles with regional Total Wright’s fixation index ($F_{ST}$) > 0.25

Upon careful examination, it appears that the population structure in HLA is more localized than that which could be captured by our global Within Region $F_{ST}$ and Between Region $F_{ST}$ estimates. Some HLA alleles had much smaller Within Region $F_{ST}$ and Between Region $F_{ST}$ estimates compared with regional Total $F_{ST}$ values (Table 3). For example, HLA-DRB1*04:11 had regional Total $F_{ST}$ values of 0.43 and 0.39 in North America (NAM) and SAM, respectively, whereas the Within Region $F_{ST}$ and Between Region $F_{ST}$ values were 0.087 and 0.168, respectively. There were also several alleles with relatively large Between Region $F_{ST}$ or Within Region $F_{ST}$ values ($F_{ST}$ > 0.15, Fig. 3 and Table 4).

We found that the classical HLA alleles have low-to-moderate population differentiation within and between regions, similar to the rest of the genome. An $F_{ST}$ value between 0.15 and 0.25 is considered to be large and an $F_{ST}$ value >0.5 is considered to be extremely large. The genome-wide level of population differentiation has previously been reported to be low, with $F_{ST}$ values of 0.071 between Europeans and Africans, 0.083 between Africans and Chinese/Japanese and 0.052 between Chinese/Japanese and Europeans (57). The frequency of HLA alleles is known to vary widely between populations, allowing them to be used as ancestry informative markers to control for population structure in genetic studies of association. Therefore, it is somewhat unexpected that results from this study demonstrate low-to-moderate population differentiation between regions and within regions ($F_{ST}$ < 0.15, Fig. 3). Nevertheless, there were some larger regional $F_{ST}$ values ($F_{ST}$ > 0.25, Table 3).

We found that the classical HLA alleles have low-to-moderate population differentiation within and between regions, similar to the rest of the genome. An $F_{ST}$ value between 0.15 and 0.25 is considered to be large and an $F_{ST}$ value >0.5 is considered to be extremely large. The genome-wide level of population differentiation has previously been reported to be low, with $F_{ST}$ values of 0.071 between Europeans and Africans, 0.083 between Africans and Chinese/Japanese and 0.052 between Chinese/Japanese and Europeans (57). The frequency of HLA alleles is known to vary widely between populations, allowing them to be used as ancestry informative markers to control for population structure in genetic studies of association. Therefore, it is somewhat unexpected that results from this study demonstrate low-to-moderate population differentiation between regions and within regions ($F_{ST}$ < 0.15, Fig. 3). Nevertheless, there were some larger regional $F_{ST}$ values ($F_{ST}$ > 0.25, Table 3).

Upon careful examination, it appears that the population structure in HLA is more localized than that which could be captured by our global Within Region $F_{ST}$ and Between Region $F_{ST}$ estimates. Some HLA alleles had much smaller Within Region $F_{ST}$ and Between Region $F_{ST}$ estimates compared with regional Total $F_{ST}$ values (Table 3). For example, HLA-DRB1*04:11 had regional Total $F_{ST}$ values of 0.43 and 0.39 in North America (NAM) and SAM, respectively, whereas the Within Region $F_{ST}$ and Between Region $F_{ST}$ values were 0.087 and 0.168, respectively. There were also several alleles with relatively large Between Region $F_{ST}$ or Within Region $F_{ST}$ values ($F_{ST}$ > 0.15, Fig. 3 and Table 4).
For example, the largest $F_{ST}$ value was for the HLA-A*34:01 allele (Between Region $F_{ST} = 0.309$).

In conclusion, our study is the largest protein sequence-based study of balancing selection in HLA to date and provides strong evidence that balancing selection in HLA is not symmetric. Further research is required to identify which type(s) of asymmetrical balancing selection may be playing a role in maintaining the diversity in HLA loci.

**MATERIALS AND METHODS**

**HLA frequency data**

This analysis included allele and haplotype frequencies for 419 population samples from around the world with at least 20 individuals each (total $N = 50303$ individuals) (Table 1). The data were available from the dBMHC Anthropology database (www.ncbi.nlm.nih.gov/gv/mhc) and Supplementary Material compiled by Solberg et al. (11) from a systematic review of literature data sets, the 12th and 13th International Histocompatibility Workshops and the AlleleFrequencies.net database (www.pyppop.org/podata) (58,59). Five duplicated data sets and 70 admixed and migrant population samples were excluded from analysis. Population samples from the same region were included in this study, as densely sampled populations from a specific area can be quite different from each other (11).

**HLA sequence data**

The HLA Informatics Group and European Bioinformatics Institute, in collaboration with the international ImMunoGeneTics project (IMGT), have established the IMGT/HLA database (60). Because clinicians regularly test HLA genes to match organ transplants, all HLA sequences are maintained in a global repository and novel alleles must conform to nomenclature regulations (6,61). The exonic sequences of 5000 classical HLA alleles were publicly available from the IMGT/HLA database as part of release version 3.4.0, of which 801 were analyzed in this study (Table 2). An additional 311 HLA haplotypes were also analyzed. The total number of allele and haplotype sequences analyzed was 1112. Official G groups were used for ambiguous alleles.

**Statistical analysis**

Allelic genealogies for each locus in single populations were generated from DNA sequences that encode the peptide-binding domains for the classical HLA loci: exons 2 and 3 for class I loci (HLA-A, HLA-B and HLA-C) and exon 2 for class II loci (HLA-DRB1, HLA-DBP1, HLA-DPA1, HLA-DQB1 and HLA-DQA1). Each genealogy represented one locus and one population and provided terminal branch lengths for alleles that were present in that single population. Because HLA-DQB1/DQA1 and HLA-DBP1/DPA1 gene products form heterodimers, DQ and DP loci were also analyzed as haplotypes. In total, 1223 genealogies were generated (Table 1). The neighbor-joining distance method in PAUP was used to make the genealogies (v 4.0b10 for Unix; http://paup.csit.fsu.edu). The neighbor-joining method iteratively tests all pairs of neighbors (alleles or cluster of alleles) while minimizing the sum of the branch lengths to compute pair-wise distances that are then used to create the tree (62,63). Trees were rooted at the midpoint between the two most distant alleles in the tree, based on branch lengths. The $R_{SD}$ statistic compares the ratio of terminal to internal branches. Thus, it is sensitive to the location of the root, and the choice of rooting can introduce a bias. Rooting the tree at the midpoint would break the longest branch of the tree and minimize its contribution to deviation from the neutral expectation for $R_{SD}$. Therefore, rooting the tree at the midpoint is a conservative choice.

An appropriate model of nucleotide selection was identified based on the analysis of sequence data for exon 2 in HLA-DRB1 and exons 2 and 3 in HLA-B for the Irish population (jModelTest v 0.1.1) (64–66). Three models were considered: the Jukes–Cantor correction, the Kimura two-parameter and the general time-reversible models. The Jukes–Cantor correction assumes that mutations occur at a constant rate and that each nucleotide is equally likely to mutate to any other (62,67). The Kimura two-parameter model extends the Jukes–Cantor correction by accounting for different rates of transitions and transversions (68). Generally, transitions (changes within purines [A, G] or pyrimidines [C, T]) are more common than transversions (changes between purines and pyrimidines). The general time-reversible model accounts for different frequencies of nucleotides and different rates for each pair of nucleotide substitutions (symmetric substitution matrix) (69). A likelihood ratio test was used to compare the Jukes–Cantor and Kimura two-parameter models, and Akaike and Bayesian information criteria were examined to select the best-fitting model. The likelihood ratio test comparing the Jukes–Cantor and Kimura two-parameter models of nucleotide substitution indicated that the Kimura two-parameter model, with unequal base frequencies, fit better than the Jukes–Cantor model ($P < 1 \times 10^{-6}$). The general time-reversible model was selected because it fit even better than the Kimura two-parameter model, based on smaller estimates of Akaike and Bayesian information criteria (data not shown).

We compared the observed shape of the trees with the expected shape of the tree under the neutral model of coalescent expectation. The neutral model of coalescent expectation can be used to represent symmetrical balancing selection because a genealogy under symmetric balancing selection has the same properties as a neutral genealogy, but with a different effective population size (70). In this study, a neutral genealogy refers to all homozygotes having equal fitness, and lower fitness than heterozygotes, and all heterozygotes having equal fitness. Deviation was quantified with the population genetics statistic $R_{SD}$, which was based on the number of alleles ($n$), the sum of the terminal branch lengths ($L$) and tree depth ($D$) (71). $R_{SD}$ was equal to $L \times (1–1/n)/D$. The expectation of $R_{SD}$ was 1.25 for genealogy topologies under the neutral model of coalescence. The assumption is that if all classes of heterozygotes are equally fit, then selection should not favor maintaining certain alleles in the population over longer spans of time than other alleles; alleles should have similar branch lengths, and trees should conform to the neutral model of coalescence. However, if specific classes of
heterozygotes are more fit than others, then alleles which were favored by selection and therefore maintained in the population longer spans of time ended up accumulating more mutations and hence are associated with longer branch lengths. The significance of the deviation between observed and expected genealogy was assessed by calculating two-tailed P-values empirically based on at least 10,000 permutations under the neutral model of coalescent expectation [Analyses of Phylogenetics and Evolution (APE v 2.6–2) package in R (R v2.12.1; http://www.r-project.org)] (72). To address multiple testing concerns for the genealogies (N = 1223), we determined significance (P ≤ 1.1 × 10−4) by controlling the false discovery rate, using the Benjamini–Hochberg method at a conservative level of 0.05% (73). Setting a significance threshold with the Bonferroni correction for the number of distinct populations (N = 419) provides a similar cutoff (P ≤ 1.19 × 10−4).

Population differentiation, or genetic variation, of HLA alleles within and between regions was examined with Wright’s fixation index (FST) (74). The Total FST for each allele was calculated as v/m × [1 − m], where v is the variance of allele frequencies and m is the mean allele frequency. Regional Total FST values were calculated for each allele by geographical region [SSA, NAF, EUR, SWA, Southeast Asia (SEA), Oceania (OCE), Australia (AUS), NEA, NAM, SAM] as v/m × [1 − m], where v is the variance of allele frequencies in the specified region and m is the mean allele frequency in the specified region. The Within Region FST was the mean of the 10 regional Total FST values. The Between Region FST was calculated by calculating the Total FST using only mean allele frequencies from each of the 10 geographical regions. An FST value of zero represents freely interbreeding populations and the maximum value of 1 represents fixed differences between populations.

Correlation between HLA allelic branch lengths and population frequencies was tested with Spearman’s correlation test.

Hierarchical analysis of molecular variance (AMOVA) tests was used to partition measures of total variability into variability among (i) geographic regions, (ii) populations within geographic regions and (iii) individuals within populations [Analysis of Ecological Data (ADE4 v 1.5-0) package in R] (75). AMOVA tests are essentially nested analysis of variance tests based on differences between molecular sequences, represented by Euclidean distances. Estimates were calculated for the sums of squares and mean squares, as well as components of covariance and their contribution to the total covariance. Estimated standard errors for our AMOVA results were based on 1000 permutations.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

We would like to thank N. Matzke, T. Heath and E. Durand for their help.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the NIH/NIGMS (grant number R01-GM040282 to M.S.) and the NIH/NIAID (grant number U01-A1067068 to S.J.M.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences, the National Institute of Allergy and Infectious Diseases or the National Institutes of Health.

REFERENCES