Linkage Disequilibrium and Age Estimates of a Deletion Polymorphism (1597ΔC) in HLA-G Suggest Non-Neutral Evolution

Carrie Aldrich, Charles Wambebe, Lillian Odama, Anna Di Rienzo, and Carole Ober

ABSTRACT: A single base deletion in HLA-G (1597ΔC) that is a null allele for the full-length protein is found at frequencies of 6%–11% in populations of African descent. To test the null hypothesis that 1597ΔC "drifted" to this frequency by neutral evolutionary processes, we compared the pattern of variation and linkage disequilibrium (LD) around this allele and around a conservative amino acid replacement polymorphism in HLA-G (T31S) in African-Americans and Nigerians. There was no significant LD between the 31S allele and flanking STRPs at 150–200 kb in either sample, but significant LD was observed between the 1597ΔC allele and the same flanking STRPs (p < 0.001 in both samples). To further characterize the evolutionary history of these variants, age estimates were determined assuming evolutionary neutrality. If these alleles were neutral, their frequencies indicate that they arose approximately 9500 (95% CI = 1557, 17557) generations ago (200,000 years). However, using LD to estimate the allele age, the 1597ΔC allele was estimated to have arisen only 744 (95% CI = 375,2713) generations ago (18,000 years ago), whereas the 31S allele has a much older estimate of 3241 (95% CI = 1680, 20500) generations ago (81,000 years ago). These data suggest that these two polymorphisms in the HLA-G gene have had different evolutionary histories. We propose that natural selection has acted on the 1597ΔC allele.

INTRODUCTION

The classical class I human leukocyte antigen (HLA-A, HLA-B, and HLA-C) genes are among the best examples of diversifying selection in modern molecular biology [1]. However, the nonclassical class I HLA genes, HLA-E, HLA-F, and HLA-G, have very different patterns of nucleotide polymorphism, suggesting that they may have been subject to different evolutionary pressures than the classical class I genes [2–4]. For example, the extensive polymorphism of the classical genes, which has been hypothesized to protect against a wide variety of pathogens, is nearly absent in the nonclassical genes. In particular, HLA-G has only two conservative amino acid substitutions (T31S and L110I). Whereas the classical class I genes are expressed ubiquitously, HLA-G is expressed primarily in the fetal tissues at the maternal–fetal interface. Finally, HLA-G has several isoforms that arise from alternative splicing (Figure 1). These include both transmembrane and soluble forms, and isoforms that exclude exon 3, which encodes the α2 domain. The HLA-G proteins without the α2 domain, HLA-G2 and HLA-G2sol, resemble another class of HLA molecules, the class II HLA [5].

Recently, a deletion polymorphism in exon 3 at nucleotide 1597 (1597ΔC) has been reported [6, 7]. This deletion of a single cytosine eliminates expression of the full-length isoforms (HLA-G1 and G1sol) in homozygous conceptuses [7], and is a striking departure from the otherwise conservative nature of variation in this gene. Surprisingly, this nonconservative polymorphism is common (~8%) in populations of African descent (Table 1). The nonconservative nature of this polymorphism, its relatively high frequency in some populations, and the putative role of HLA-G in placental development suggest that the maintenance of this mutation may have been subject to nonneutral evolutionary forces.

To test the null hypothesis that 1597ΔC reached its current frequencies in African populations by random genetic drift, we studied variation in the genomic regions...
flanking the HLA-G locus. We reasoned that if the 1597ΔC variant "drifted" to high frequencies it should be relatively old and exhibit correspondingly low levels of linkage disequilibrium (LD) with flanking markers. On the other hand, if it was driven to high frequencies by selection, it may be relatively young and exhibit more LD with flanking markers. To differentiate between these two possibilities, we estimated the age of 1597ΔC by two methods: one based on its frequency and assuming neutral evolution [8] and one based on the amount of LD with flanking markers [9]. In addition, we studied a second HLA-G polymorphism that is a conservative Thr → Ser substitution at amino acid 31 in the α1 domain. The 31S allele occurs at similar frequencies to the HLA-G null allele in our study samples and serves as a "control" allele in this study.

MATERIALS AND METHODS
Sample Composition
DNA samples from healthy African-American subjects (n = 95) from the Chicago area (Chicago, Illinois, USA), who had participated in previous studies in our lab [10], and samples from 89 random Nigerian patients undergoing screening for the sickle cell trait in Abuja, Nigeria were used in this study. In addition, the frequency of 1597ΔC was determined in DNA samples from 88 healthy, unrelated Sardinians and 35 healthy, unrelated subjects from the Cameroon. This study was approved by the Institutional Review Board of the University of Chicago.

DNA Extraction
DNA was extracted from whole blood, placental samples or dried blood spots, depending on sample availability. For the whole blood samples, DNA was either extracted using a phenol/chloroform protocol [11] or was "salted out" using a commercially available kit (Gentra Systems, Minneapolis, MN, USA). Placental samples were extracted as described previously [10, 12]. DNA was extracted from dried blood spots in the Nigerian sample as follows. Two 7-mm punches from dried blood spots were added to 500-μl PBS + 50-μg proteinase K. The samples were incubated overnight at 55°C, then cooled on ice for 10 minutes, followed by a standard phenol-chloroform extraction. The dried DNA pellets were dissolved
TABLE 1  Frequency of 1597ΔC in diverse population samples

<table>
<thead>
<tr>
<th>African or African descent</th>
<th>Reference</th>
<th>Number of chromosomes</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zimbabwe–Shona</td>
<td>[24]</td>
<td>216</td>
<td>0.111</td>
</tr>
<tr>
<td>African American</td>
<td>[23]</td>
<td>84</td>
<td>0.083</td>
</tr>
<tr>
<td>African American</td>
<td>[7]</td>
<td>272</td>
<td>0.074</td>
</tr>
<tr>
<td>Nigerian</td>
<td>This study</td>
<td>178</td>
<td>0.073</td>
</tr>
<tr>
<td>Cameroon–Ewondo</td>
<td>This study</td>
<td>32</td>
<td>0.063</td>
</tr>
<tr>
<td>Cameroon–Bamileke</td>
<td>This study</td>
<td>38</td>
<td>0.053</td>
</tr>
<tr>
<td>Ghanaian</td>
<td>[23]</td>
<td>84</td>
<td>0.048</td>
</tr>
</tbody>
</table>

Non-Africans

| Spanish                    | [5]       | 228                   | 0.031     |
| Mexican Americans          | [7]       | 102                   | 0.029     |
| Sardinians                 | This study | 162                   | 0.025     |
| German/Croat               | [28]      | 264                   | 0.023     |
| Danish                     | [25]      | 154                   | 0.006     |
| Portuguese                 | [26]      | 117                   | 0         |
| European (U.K.)            | [27]      | 38                    | 0         |
| European (U.S.)            | [7]       | 134                   | 0         |
| Japanese                   | [23]      | 82                    | 0         |

in 10- to 20-μl TE buffer, pH 8.0, and 5 μl was used in a 100-μl polymerase chain reaction (PCR).

HLA-G, HLA-A*30, and Microsatellite Genotyping

Genotyping for the HLA-G polymorphisms 31S and 1597ΔC and for the HLA-A30 allele was performed using sequence-specific oligonucleotide probes on immobilized PCR-amplified DNA. The methods for genotyping HLA-G have been described previously [7, 13]. For HLA-A*30 typing, exons 2 through 3 of HLA-A were amplified from genomic DNA as described [14]. The resulting PCR product was immobilized on nylon membranes as for HLA-G and probed with three sequence-specific, biotin-labeled oligonucleotide probes, one specific for HLA-A30 (probe A30: GGTATGAACACGCACCTAC) and two others specific for non-HLA-A30 alleles (A02/23: GTAGGCGTACTGTGGTACC and Anon30/02/32: GGGTACCCGGACCAGCTCTAC). Positive and negative control samples were included on each membrane to ensure specificity of hybridization. Positive hybridization was detected via conjugation of the biotin-labeled probes with streptavidin-horseradish peroxidase, and color development of tetramethyl benzidine substrate.

Short tandem repeat polymorphisms (STRP; D6S1260, D6S1281, D6S265, D6S258, MOGc, TNFa) were genotyped by polyacrylamide gel electrophoresis of radiolabeled PCR products according to standard protocol [15].

Statistical Analysis

Haplotype frequency estimation and pairwise linkage disequilibrium calculations were performed with the HAPLO and LinkD programs, respectively [16, 17]. The HAPLO program estimates the haplotype frequencies of a population sample from genotype data using a Baum/EM algorithm that maximizes the likelihood of the haplotype distribution given the genotype data. LinkD performs standard pairwise linkage disequilibrium calculations including D', a measurement of linkage disequilibrium that is normalized for the maximum value possible under the current allele distribution [18]. Statistical significance was determined by Fisher's exact test [19].

Allele age estimates from haplotype data were generated as the parameter τ by the decay of haplotype sharing (DHS) method [9], which models linkage disequilibrium as the average amount of the ancestral chromosome that is still intact (or "shared") for any given polymorphism. The DHS method generates estimates and confidence intervals for the time to the most recent common ancestor of a group of haplotypes (TM RCA) by modeling the decay of the ancestral haplotype as a series of breakpoints that are distributed as independent exponential random variables. Strictly speaking, this TM RCA is not the time when a new allele arose by mutation but represents the lower boundary for that time; that is, the age of the variant will generally be somewhat higher than that of the most recent common ancestor in a population. It will be referred to as "allele age." Importantly, the DHS method allows for different demographic scenarios. That is, confidence intervals for the TM RCA can be generated assuming independent recombinational histories as would be expected if the population underwent a rapid expansion from a small initial size ("star-shaped phylogeny"). Alternatively, the recombinational histories can be modeled as dependent (i.e., some portion of sampled chromosomes share a recombinational breakpoint that is identical-by-descent), as expected for a population of constant size. Although the confidence intervals are broadened by assuming recombination dependence, the estimate of TM RCA is unchanged [9]. This method also assumes that there is no selection at marker loci but can allow for mutation and for multiple origins of a variant. The genetic distances for the STRP markers were generated from physical distance using the measurement of recombination from Malfroy et al. [20].

Allele age estimates under neutral evolution were calculated from the allele frequencies as described in Kimura and Ohta [8]. This estimate was generated under the infinite alleles model using a diffusion equation approximation of the dynamics of a neutral mutation. An effective population size of 10,000 [21] was assumed. Confidence intervals were generated by integration of the diffusion equation [22].
RESULTS

The frequency of 1597ΔC in 16 population samples is presented in Table 1. The frequency in our sample of African-Americans is 0.074, which is similar to the 0.083 frequency in another independent sample of African-Americans and to the 0.073 frequency in Nigerians. A recent report of HLA-G allele frequencies in the Shona ethnic group of Zimbabwe described the highest frequency (0.11) of 1597ΔC yet reported. Another sample from the Cameroon in West Africa had a slightly lower frequency of this variant (0.057). Allele frequencies in non-African samples were yet lower, ranging from zero or nearly zero in four Caucasian and one Japanese sample [7, 23, 25–27] to 0.02–0.03 in three eastern and southern European and one admixed (Mexican-American) sample [6, 7, 28].

Whereas the high frequency of 1597ΔC in African populations and its absence in Northern Europeans is intriguing and reminiscent of several well-known examples of positive natural selection [29], this evidence alone is not sufficient to invoke non-neutral evolution. To begin to explore the hypothesis that 1597ΔC has been subject to positive natural selection, we evaluated pairwise LD. The $D' = 0.2$ with $p = 0.05$ and $D' = 0.6$ with $p < 0.001$ for the 154 bp allele of the MOGe locus that is 150 kb telomeric of HLA-G, in African-Americans and Nigerians, respectively. Likewise, 1597ΔC is in strong and significant LD with the 140 bp allele of the D6S265 marker ($D' = 0.9, p < 0.001$ in African-Americans and $D' = 0.6, p < 0.001$ in Nigerians), 200 kb centromeric of HLA-G. However, 31S exhibited no significant LD with either D6S265 or MOGe ($p > 0.05$ for all alleles).

Linkage disequilibrium between 1597ΔC and an allele at the next nearest marker, D6S1683 ($D' = 0.5, p = 0.05$), 350 kb telomeric of HLA-G, was lower and less significant than for MOGe and D6S265. This decay of LD is not unexpected because the chance of recombination interrupting the ancestral haplotype is a function of distance. Markers >500 kb from HLA-G revealed little evidence for LD with any allele of this gene, although an occasional weak association with borderline significance ($p = 0.05$) was noted for these markers (Figure 2).

Because the LD around 1597ΔC was high and significant, suggesting a relatively recent origin for this allele, we used two age estimates for 1597ΔC and for 31S to quantify whether this large amount of LD was inconsistent with expectations for a neutral allele. The allele age (TMRCA) for 1597ΔC ($\tau = 744$ generations or 18,600 years) estimated by the DHS method [9] is younger than that for 31S ($\tau = 3241$ generations or 81,025 years; Table 2), and younger than the estimate based on the
Three observations support this conclusion. First, the same markers (D6S265 and MOGc) that demonstrate strong LD with 1597ΔC are in equilibrium with the conservative 31S allele in both populations. This is intriguing because 31S occurs at a similar frequency to 1597ΔC and is subject to the same amount of recombination. Furthermore, age estimates of the 1597ΔC and 31S alleles, based on the extent of LD, suggest that the 31S allele may be a significantly older allele than 1597ΔC (31S: 3241 generations, 95% CI = 1680, 20500; 1597ΔC: 744 generations, 95% CI = 325, 2713), despite their similar frequencies. Thus, the 1597ΔC allele appears to have risen to its current frequency in a shorter time compared with the 31S mutation, retaining significant LD with flanking markers. This also indicates that the high LD on the 1597ΔC haplotype is not a result of the low recombination rate that has been noted for the HLA region [20], although we cannot rule out the possibility that the deletion mutation arose on a haplotype with an unusually low recombination rate.

Second, the allele age estimate (TMRCA) for 1597ΔC generated under neutral expectations (frequency-based) is an order of magnitude larger than the DHS estimate (recombination based), with only slight overlap of 95% confidence intervals. Although age estimates for both 31S and 1597ΔC based on DHS are younger than estimates based on the allele frequency (Table 2), the confidence intervals for the age estimates of 31S based on the two methods overlap almost completely, again suggesting that the 31S and 1597ΔC alleles had different evolutionary histories. While the age estimates of 31S are concordant with neutral evolution, those for 1597ΔC are not. Whereas the DHS point estimate for the age of 1597ΔC differs more than tenfold from that generated under neutrality, the confidence intervals for this estimate are dependent on assumptions about population dynamics. When the confidence intervals for the DHS (recombination based) estimate are generated under the assumption of independent histories for the sampled chromosomes (“star-shaped genealogy”), they do not overlap with those for the frequency-based estimate (Table 2). A similar approach to generating confidence intervals based on the assumption of independence was used in calculating the age of the CCR5Δ32 mutation by Stephens and colleagues [34]. However, the assumption that the sampled chromosomes have independent histories is based, in part, on data that may be inappropriate for population inferences for nuclear genes [35]. When the intervals are corrected for the possibility that the lineages of sampled chromosomes are not independent (coalescent genealogy) [9], as in a population that has not undergone recent expansion, there is some overlap of the confidence intervals.

### DISCUSSION

The relatively high frequency of this null mutation was surprising in light of the critical role that HLA-G likely plays in reproduction [30–33]. Although the frequency of 1597ΔC is highest in African samples or samples of African descent, it is also present, albeit at lower frequencies, in several European populations, such as Sardinians and Spanish. This suggests that the mutation may have arisen in Africa and spread to Europe. In fact, Suárez et al. [6] propose that 1597ΔC may have been introduced into Spain on an HLA-A30 haplotype by Moorish conquerors during the 8th century AD. Our data are consistent with this hypothesis. Fifteen of 16 African-American subjects in our study who were positive for 1597ΔC were also positive for HLA-A*30, supporting the hypothesis that the deletion arose on an HLA-A30 haplotype in Africa.

The reason for the unusually high frequency for this null mutation is unknown, but the pattern of high LD found for 1597ΔC suggests that it was driven to its current frequency in African populations by selection.
Third, the LD pattern for $1597\Delta C$ differs from genome-wide data. It demonstrates strong, significant linkage disequilibrium ($D' = 0.6–1.0$) with alleles at flanking markers, D6S265 and MOGc, in both Nigerians and African-Americans. These markers are 200 and 150 kb from HLA-G, respectively. Although significant pairwise LD has been found at this distance in genome-wide surveys, the strength of this association ($D'$) averaged only 0.05 at distances of 200 kb or more [36]. Thus, $1597\Delta C$ has patterns of LD that are remarkable even compared with the rest of the genome. Examination of recent surveys of LD can also provide an empirical perspective on the significance of the association of $1597\Delta C$ with flanking polymorphism. Significant LD beyond 100 kb is unusual but not unprecedented [36, 37]. The confidence intervals placed on age estimates rely on assumptions about population history and the nature of mutation processes. However, most of the currently available data sets do not include populations of African descent [36–39] and examine LD between single nucleotide polymorphisms (SNPs) rather than microsatellites. Finally, not all of the previous studies provided measures of LD that are comparable to those in this report, either because they do not report the strength of the association [37] or because they report only pairwise LD values as opposed to multi-locus estimates, such as those implemented in the DHS program [9]. However, examination of these studies suggests that the significant and strong association of $1597\Delta C$ ($D'$ of 0.9–1.0) with markers $>150$ kb from HLA-G is unusual [36, 40].

Two different scenarios could account for these observations: hitchhiking of $1597\Delta C$ with another locus, or selection acting directly on the $1597\Delta C$ allele. The alteration of patterns of variation at one locus due to selective pressures at a nearby locus, known as "hitchhiking," has been demonstrated for the class I pseudogene, HLA-H, which is located within the 250-kb region between the HLA-A and HLA-G loci [41]. If another allele on the chromosome upon which $1597\Delta C$ arose had or has a selective advantage in these populations, the $1597\Delta C$ allele could have risen in frequency due to hitchhiking. This could result in the higher disequilibrium values because its "rapid" rise in frequency would not have allowed time for recombination and mutation to reduce the association of $1597\Delta C$ with the selected polymorphism on the chromosome on which it arose. For example, if $1597\Delta C$ arose on an HLA-A30 haplotype, as the data suggest, and the A30 allele was selected, then the $1597\Delta C$ allele could have hitchhiked along with A30 as it rose in frequency. Although there is no evidence for selection acting on the A30 allele, and no associated diseases [42], including recurrent miscarriage [43], this possibility cannot be ruled out. On the other hand, another allele at a different locus on the A30 haplotype may have been selected, but at present there are no known candidates.

Alternatively, the deletion mutation may have risen in frequency faster than expected because the $1597\Delta C$ allele per se was selected in African populations. The function of HLA-G at the maternal–fetal interface is not completely understood and the identification of a few individuals who are homozygous for the null mutation indicates that the full-length isoform (HLA-G1) is not essential for fetal survival in uncomplicated pregnancies [7, 44]. In these instances it is possible that other HLA-G isoforms substitute for HLA-G1 and contribute toward the survival of the allogeneic fetus. However, the HLA-G*0105N allele, which is defined by the $1597\Delta C$ mutation, was associated with recurrent miscarriage in two independent studies [45, 46], and suggests that fetuses who inherited this allele were less likely to survive to term, despite the fact that they should produce normal levels of the other HLA-G isoforms. Thus, the relatively high frequency of this presumably deleterious mutation in populations of African descent is paradoxical and suggests that there may be some benefit that counteracts this cost. In particular, the selective pressures on the $1597\Delta C$ mutation may vary in different environments, with positive selection counterbalancing the deleterious effects of the deletion allele under certain conditions of pregnancy. If this were the case, then reduced expression of HLA-G1 protein in carriers of the $1597\Delta C$ mutation may actually enhance fetal survival in some environments.

What could this selective advantage be? The fact that the highest frequencies of $1597\Delta C$ are in populations from areas with an historically high pathogen load suggests that intrauterine pathogens may be the selective agents, with increased survival of fetuses carrying the $1597\Delta C$ allele in infected pregnancies. Consistent with this hypothesis is the recent demonstration that soluble HLA-G1 induces apoptosis in activated T cells [47], suggesting that one possible role for G1 protein is to "rid" the uterus of alloreactive maternal T cells in pregnancy. In this case, the reduced HLA-G1 expression in $1597\Delta C$ heterozygous placentas may result in an overall increase in the number of T cells available in the uterus to combat intrauterine infections. In the same manner, the inhibition of natural killer cells via the interaction with HLA-G specific natural killer inhibitory receptors [48] could reduce fitness in the presence of intrauterine infection. A deletion mutation that reduces the amount of the most abundant transcript, HLA-G1, could potentially reverse some of this inhibition, leaving natural killer and T cells free for defense against infections. This survival advantage to fetuses carrying the deletion mutation in regions with high pathogen loads may be considerably greater than the risk of being miscarried,
Evolution of an HLA-G Null Allele

thus potentially maintaining allele frequencies as high as 11% in sub-Saharan Africa [24].

The findings described in this article are consistent with the notion that HLA-G plays an important role in the biologic processes that occur at the maternal–fetal interface and that the HLA-G1 isoforms are involved in these critical interactions. These data further suggest that natural selection may have patterned variation in this gene and be responsible for the high frequency of a null allele in population of African origins. Future studies of the natural history and function of the HLA-G1 isoform in normal and infected pregnancies are required to further elucidate the selective mechanism that influenced the evolution of the 1597ΔC allele.

ACKNOWLEDGMENTS
The authors express their appreciation to Andrew Strahs and Mary Sara McPeek for assistance with the decay of haplotype sharing analysis, and to G. Destro-Bisol and G. Spedini for providing samples from the Cameroon. This work was supported by HD21244 and HD27686.

REFERENCES


