Ancient Interlocus Exon Exchange in the History of the HLA-A Locus

Austin L. Hughes and Masatoshi Nei

Center for Demographic and Population Genetics, The University of Texas Health Science Center, Houston, Texas 77225

Manuscript received November 22, 1988
Accepted for publication March 31, 1989

ABSTRACT

The major histocompatibility complex (MHC) in humans and chimpanzees includes three classical class I loci, A, B, and C, which encode glycoproteins expressed on the surface of all nucleated cells. There are also several nonclassical class I loci including E, which have more limited expression. By analyzing published sequences, we have shown that in exons 4 and 5, A locus alleles from both humans and chimpanzees are much more similar to E than to B or C alleles, whereas in exons 2 and 3 alleles from all three classical class I loci are much more similar to each other than any one is to E. We propose that some 20 million years ago, interlocus recombination led to the formation of a hybrid gene in which exons 2 and 3 were derived from the original A locus and exons 4 and 5 were derived from the E locus. The fact that such an ancient event can still be detected suggests that interlocus recombination is rare in the MHC and does not significantly contribute to MHC polymorphism, which is known to be extremely high. The present finding, however, supports Gilbert's idea that exons in a gene may occasionally be replaced by those from another gene in the evolutionary process.

In a recent study (Hughes and Nei 1988), we compared rates of synonymous and nonsynonymous (amino acid altering) nucleotide substitution in various regions of mouse and human class I major histocompatibility complex (MHC) genes. We found that the rate of nonsynonymous substitution substantially exceeds that of synonymous substitution in the 57 codons encoding the antigen recognition site (Bjorkman et al. 1987a, b) of the class I molecule, whereas in other regions of the same genes the synonymous rate generally exceeds the nonsynonymous rate. We interpreted these results as evidence of overdominant selection acting on the antigen recognition site (Hughes and Nei 1988). However, our study revealed an anomalous feature that could not be explained by the overdominance hypothesis alone. That is, in exon 4, which encodes the α3 domain of the class I molecule, the HLA-A gene in humans was found to be highly divergent from the HLA-B and -C genes.

In order to explain this observation, we proposed that in the past interlocus exon exchange had occurred. We hypothesized that the ancestral A locus received exon 4 from a locus not closely related to it or to B and C. At that time we had no idea about the donor locus. Here we show that the divergent nature of exon 4 of HLA-A can be explained as a result of past recombination with an HLA-E-like gene.

DNA SEQUENCES USED

The human classical class I loci, HLA-A, -B, and -C, encode MHC molecules, which are expressed on all nucleated cells. Their main function is to present intracellularly processed foreign peptides to cytotoxic T cells (Klein 1986). The chimpanzee genome also includes three classical class I loci, designated as ChLA-A, -B, and -C (Mayer et al. 1988). Typically in mammals, there are several nonclassical class I genes in addition to the classical genes; nonclassical genes have a more limited tissue expression and their function is not known (Howard 1987). It has been suggested that most or all nonclassical genes have no function and, even though they are sometimes expressed, they are effectively pseudogenes (Howard 1987; Klein and Figueroa 1986). A few nonclassical human genes have been sequenced. One such locus has been designated HLA-E (Koller et al. 1988).

Classical and nonclassical class I genes have a similar structural organization. Exon 1 encodes the leader peptide, and exons 2, 3 and 4 encode the three extracellular domains, α1, α2 and α3, respectively. The 57 residues of the antigen recognition site are located in the α1 and α2 domains. Exon 5 encodes the transmembrane portion of the molecule, and exons 6, 7 and 8 encode the cytoplasmic tail.

The number of amino acid residues in the cytoplasmic tail varies considerably among different loci. In this paper we compared published nucleotide sequences for exons 2–5 of eight human and two chimpanzee classical I alleles. Exons 1 and 6–8 were excluded from our analysis because alignment among distantly related genes was uncertain in these regions. Intron sequences were not used since several published sequences were for cDNA. The sequences analyzed included two HLA-A alleles (Koller and Orr 1985; Nguyen et al. 1985), two HLA-B alleles (Kottmann et al. 1986;
Zemmour et al. 1988), two HLA-C alleles (Gussow et al. 1987), one ChLA-A allele, one ChLA-B allele (Maye et al. 1988), and HLA-E gene.

RESULTS AND DISCUSSION

Figure 1 shows DNA sequences for exons 2–5 from the HLA-E and from selected human and chimpanzee classical class I genes. Inspection of these sequences suggests that the relationship between the E gene and the classical genes for exons 2 and 3 is different from that for exons 4 and 5. In exons 2 and 3, one can see many unique substitutions which set apart from the classical sequences. In fact, E shows unique amino-acid residues at 31 positions; these residues are not shared by any of the classical genes illustrated in Figure 1 or by any of the other classical genes in our sample. Of these 31 residues unique to E, 18 are in the antigen recognition site of the MHC molecule, which includes many highly polymorphic residues. Exon 4 shows fewer nonsynonymous differences among sequences, as is expected in a highly conserved domain. However, there are many synonymous substitutions that are shared by most A and E sequences but not by B and C. For example, at codons 225, 257, 259, and 267, the third-position nucleotide is identical for E and all A sequences (including those not illustrated in Figure 1), whereas all B and C sequences
TABLE 1
Nucleotide substitutions per site (d) (above diagonal) and synonymous substitutions per synonymous site (ds) (below diagonal) in exons 2 and 3 in comparisons among human and chimpanzee class 1 alleles

<table>
<thead>
<tr>
<th></th>
<th>A3</th>
<th>A244</th>
<th>Ch-A108</th>
<th>B13</th>
<th>B44</th>
<th>Ch-B2</th>
<th>Cu1</th>
<th>Cu2</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3</td>
<td>.015</td>
<td>.030</td>
<td>.128</td>
<td>.121</td>
<td>.104</td>
<td>.108</td>
<td>.121</td>
<td>.179</td>
<td></td>
</tr>
<tr>
<td>A244</td>
<td>.050</td>
<td>.075</td>
<td>.115</td>
<td>.119</td>
<td>.119</td>
<td>.119</td>
<td>.141</td>
<td>.186</td>
<td></td>
</tr>
<tr>
<td>Ch-A108</td>
<td>.030</td>
<td>.055</td>
<td>.128</td>
<td>.121</td>
<td>.104</td>
<td>.108</td>
<td>.121</td>
<td>.179</td>
<td></td>
</tr>
<tr>
<td>B13</td>
<td>.129</td>
<td>.101</td>
<td>.125</td>
<td>.119</td>
<td>.119</td>
<td>.119</td>
<td>.141</td>
<td>.186</td>
<td></td>
</tr>
<tr>
<td>B44</td>
<td>.118</td>
<td>.108</td>
<td>.117</td>
<td>.055</td>
<td>.077</td>
<td>.121</td>
<td>.117</td>
<td>.175</td>
<td></td>
</tr>
<tr>
<td>Ch-B2</td>
<td>.104</td>
<td>.091</td>
<td>.104</td>
<td>.066</td>
<td>.079</td>
<td>.119</td>
<td>.115</td>
<td>.191</td>
<td></td>
</tr>
<tr>
<td>Cu1</td>
<td>.109</td>
<td>.110</td>
<td>.134</td>
<td>.091</td>
<td>.079</td>
<td>.119</td>
<td>.115</td>
<td>.191</td>
<td></td>
</tr>
<tr>
<td>Cu2</td>
<td>.150</td>
<td>.155</td>
<td>.165</td>
<td>.155</td>
<td>.088</td>
<td>.062</td>
<td>.189</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.208</td>
<td>.161</td>
<td>.233</td>
<td>.168</td>
<td>.184</td>
<td>.177</td>
<td>.217</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The prefix "Ch-" denotes chimpanzee alleles; others are human genes.

(including those not illustrated) share a different third-position nucleotide. Nonsynonymous differences are more numerous in exon 5 than in exon 4. In these regions A and E share many second and third position substitutions. At amino acid positions 288, 292, and 293, all A and E sequences share an amino acid residue differing from that of all B and C (including those not illustrated).

In order to quantify the pattern of nucleotide substitution among the genes in our sample, we estimated the number of nucleotide substitutions per site (d) and the number of synonymous substitutions per synonymous site (ds) (Nei 1987) in pairwise comparisons among the 10 genes. ds's were estimated by method I of Nei and Gojobori (1986). We computed these values separately for exons 2–3 and exons 4–5. Table 1 shows the results for exons 2–3. In exons 2–3, E is quite divergent from classical genes of the human and chimpanzee in terms of both total substitutions (d) and synonymous substitutions (ds). In the case of exons 4–5, however, A locus alleles from both the human and chimpanzee are in general more similar to E than they are to B or C locus alleles (Table 2). The divergence of A and E from B and C is particularly marked in the case of synonymous substitutions. The amino acid sequence is conserved at most positions in the transmembrane region (encoded by exon 5) and is highly conserved in the α3 domain (encoded by exon 4). Thus the rate of synonymous substitution exceeds the nonsynonymous rate in exons 4 and 5 (Hughes and Nei 1988, and unpublished data). For this reason, synonymous substitutions may give a clearer picture of evolutionary relationships among the genes than total substitutions.

Using the distance values shown in Tables 1 and 2, we constructed evolutionary trees using the unweighted pair group method (UPGMA; Sneath and Sokal 1973), to examine the relationships among the genes in our sample. We constructed trees separately for exons 2–3 and exons 4–5. Figure 2, A and B, shows the trees based on total nucleotide differences. In the case of exons 2–3 (Figure 2A), the B and C loci appear to be more closely related to each other than either is to the A locus of either the human or chimpanzee, but E forms a separate cluster, apart from all the classical loci. By the method of Nei, Stephens and Saitou (1985), the branch point connecting E with the cluster of classical loci is significantly different from that connecting the A cluster to the B–C cluster (t = 6.31; P < 0.001). The UPGMA tree for exons 4–5 presents a very different topology (Figure 2B). Here, although B and C alleles again cluster together, A alleles cluster with E rather than with B and C alleles. In this case, the branch point connecting the B–C cluster to the A–E cluster is significantly different from that connecting A and E (t = 4.42; P < 0.001). Note that, while each of the A, B and C loci forms a separate cluster in Figure 2B as in Figure 2A, the details of topology within these clusters differ between the two figures. However, these differences are not statistically significant because the distances between the branch points within clusters are very small.

Figure 2, C and D, shows the trees based on synonymous substitutions only. In the case of exons 2–3, E again forms a separate cluster apart from the classical loci (Figure 2C). Here also, the branch point between the E cluster and the classical gene cluster is significantly different from that between the two major clusters of classical genes (t = 3.29; P < 0.01). Note that B alleles are grouped with A alleles rather than with C alleles in this case. However, the difference between the two branch points involved is not statistically significant. In the case of exons 4–5 (Figure 2D), E again clusters with A alleles, the difference between the branch point connecting A and E and that connecting the A–E cluster to the B–C cluster again being significant (t = 5.19; P < 0.001).

In order to confirm the evolutionary relationships revealed by these UPGMA trees, we constructed trees using the neighbor-joining method (Saitou and Nei
TABLE 2

Nucleotide substitutions per site ($d$) (above diagonal) and synonymous substitutions per synonymous site ($d_s$) (below diagonal) in exons 4 and 5 in comparisons among human and chimpanzee class I alleles

<table>
<thead>
<tr>
<th></th>
<th>A3</th>
<th>A3w24</th>
<th>Ch-A108</th>
<th>B13</th>
<th>N44</th>
<th>Ch-B2</th>
<th>Cw-1</th>
<th>Cw-2</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3</td>
<td></td>
<td>.026</td>
<td>.013</td>
<td>.092</td>
<td>.089</td>
<td>.089</td>
<td>.121</td>
<td>.121</td>
<td>.064</td>
</tr>
<tr>
<td>A3w24</td>
<td>.075</td>
<td></td>
<td>.029</td>
<td>.086</td>
<td>.089</td>
<td>.089</td>
<td>.127</td>
<td>.121</td>
<td>.075</td>
</tr>
<tr>
<td>Ch-A108</td>
<td>.081</td>
<td>.086</td>
<td></td>
<td>.098</td>
<td>.095</td>
<td>.095</td>
<td>.127</td>
<td>.121</td>
<td>.072</td>
</tr>
<tr>
<td>B13</td>
<td>.248</td>
<td>.234</td>
<td>.276</td>
<td></td>
<td>.010</td>
<td>.010</td>
<td>.078</td>
<td>.067</td>
<td>.109</td>
</tr>
<tr>
<td>B44</td>
<td>.221</td>
<td>.234</td>
<td>.248</td>
<td>.031</td>
<td></td>
<td>.020</td>
<td>.078</td>
<td>.067</td>
<td>.112</td>
</tr>
<tr>
<td>Ch-B2</td>
<td>.249</td>
<td>.355</td>
<td>.277</td>
<td>.031</td>
<td>.020</td>
<td></td>
<td>.078</td>
<td>.067</td>
<td>.112</td>
</tr>
<tr>
<td>Cw-1</td>
<td>.280</td>
<td>.295</td>
<td>.310</td>
<td>.168</td>
<td>.156</td>
<td>.182</td>
<td></td>
<td>.015</td>
<td>.121</td>
</tr>
<tr>
<td>Cw-2</td>
<td>.306</td>
<td>.291</td>
<td>.337</td>
<td>.142</td>
<td>.154</td>
<td>.155</td>
<td>.020</td>
<td></td>
<td>.121</td>
</tr>
<tr>
<td>E</td>
<td>.097</td>
<td>.192</td>
<td>.132</td>
<td>.206</td>
<td>.206</td>
<td>.234</td>
<td>.222</td>
<td>.247</td>
<td></td>
</tr>
</tbody>
</table>

The prefix "Ch-" denotes chimpanzee alleles; others are human genes.

FIGURE 2.—Evolutionary trees showing relationships among exons 2–3 (A and C) and exons 4–5 (B and D) of human and chimpanzee class I MHC alleles. Trees A and B are based on total nucleotide difference ($d$), while trees C and D are based on synonymous differences only ($d_s$). The prefix "Ch-" indicates chimpanzee alleles. Abbreviated allele designations: Ch-A = Ch-A108, A24 = A3w24, Ch-B = Ch-B2, Cw-1 = Cw-1, and Cw-2 = Cw-2.

1987) from the same set of data. This method is based on the principle of minimum evolution and is insensitive to differences in the rate of evolution among different lineages. The trees constructed by this method showed the same relationships among loci as the UPGMA trees (the trees not shown).

The extremely long third intron (between exons 3 and 4) of class I MHC genes contains numerous repeats (Ronne et al. 1985) and has been proposed as a recombinational hotspot in the mouse (Watts et al. 1987). Although other introns could be aligned, Koller and Orr (1985) considered alignment of this intron between A and C alleles impossible, suggesting past recombination in this intron. We hypothesize that interlocus recombination occurred in this intron, joining exons 2 and 3 of an ancestral A gene with exons 4 and 5 of an E-like gene (Figure 3). All known A locus alleles in both humans and chimpanzees are apparently descended from this hybrid allele. Because the hybrid sequence is observed in both human and chimpanzee A alleles, the A–E recombinational event apparently occurred prior to separation of the human and chimpanzee lineages (5–7 million years ago). However, it must have occurred after the separation of HLA-A from the HLA-B and -C loci (about 40 million years ago; Klein and Figueroa 1986), since the alleles from the latter loci do not share the hybrid sequence. This suggests that the recombination event occurred some time between 5 and 40 million years ago. If we assume that the E gene in our sample is descended from a gene closely related to the one involved in recombination with A and that synonymous substitution occurs at a rate proportional to time, we can estimate the age of the recombination event by the extent of synonymous substitution between exons 4–5 of A and E in comparison with that between A and B plus C (Figure 3). This method yields an estimate of about 20 million years for the age of the recombination event.

As an alternative hypothesis to explain the data presented here, one might argue that the divergence of exons 2–3 of E from those of classical class I genes
is due to selection on the recognition site of exons 2 and 3. On this hypothesis, A and E would actually be closely related, but their close relationship would be obscured by selection for amino acid substitutions in the recognition site. This hypothesis, however, does not explain why the exons 2–3 of E are markedly different from those of classical genes at synonymous sites (Figure 2C), which are presumably not affected by selection. Furthermore, in a tree based on exons 2 and 3 excluding the 57 recognition site codons, the E again clusters apart from the classical class I sequences (the tree not shown).

When introns were first discovered, it was suggested that new genes might evolve by exon shuffling (Gilbert 1978). By now, a number of examples of this process have been reported (Doolittle 1985; Sühoff et al. 1985a, b), but most of them are based on homologies between exons of very distantly related genes. The event of exon shuffling reported here differs in that the genes involved are closely related and that the donor gene can be identified. In the present case, however, it is not clear whether any change in function accompanied the A-E recombination. A case of exon shuffling was also reported by Holmes and Parham (1985), but this represents an event of intragenic recombination.

The present data support the occurrence of an ancient exon exchange in the history of human and chimpanzee A alleles, but it should be noted that interlocus exon exchange is an extremely rare event. This is obvious because we could identify the trace of recombination only once in the entire history of the HLA-A, B, and C loci that spans tens of millions of years. A number of authors (e.g., Weiss et al. 1983; Mello 1987) proposed that interlocus recombination or gene conversion is an important mechanism of maintenance of MHC polymorphism. However, the exon exchange detected here has little to do with MHC polymorphism, because a large portion of allelic variation in MHC loci is caused by variation in the antigen recognition site of exons 2 and 3. Furthermore, recombination or gene conversion cannot explain the high rate of nonsynonymous nucleotide substitution in comparison with the synonymous rate. As suggested by Hughes and Nei (1988, 1989), the extremely high level of polymorphism at MHC loci (80–90% heterozygosity) seems to be mainly due to over-dominant selection.

This study was supported by grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED


WEISS, E. H., A. L. MELLOR, L. GOLDEN, K. FAHRNER, E. SIMPSON,
Communicating editor: J. AVISE