

# Evidence of Selection on Silent Site Base Composition in Mammals: Potential Implications for the Evolution of Isochores and Junk DNA

Adam Eyre-Walker

Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton, BN1 9QG, United Kingdom

Manuscript received May 8, 1998

Accepted for publication February 22, 1999

## ABSTRACT

It has been suggested that mutation bias is the major determinant of base composition bias at synonymous, intron, and flanking DNA sites in mammals. Here I test this hypothesis using population genetic data from the major histocompatibility genes of several mammalian species. The results of two tests are inconsistent with the mutation hypothesis in coding, noncoding, CpG-island, and non-CpG-island DNA, but are consistent with selection or biased gene conversion. It is argued that biased gene conversion is unlikely to affect silent site base composition in mammals. The results therefore suggest that selection is acting upon silent site G + C content. This may have broad implications, since silent site base composition reflects large-scale variation in G + C content along mammalian chromosomes. The results therefore suggest that selection may be acting upon the base composition of isochores and large sections of junk DNA.

**B**ASE composition varies along mammalian chromosomes over scales of hundreds of kilobases to megabases, with G + C contents fluctuating from ~35% up to ~55% (Bernardi 1995). Large blocks of DNA of different compositions are often referred to as "isochores." The variation in composition appears to reflect a level of genome organization since the density of genes, recombination events, and short interspersed elements (SINES) are higher in the G + C rich parts of the genome, whereas long interspersed elements (LINEs) are almost exclusively restricted to the G + C poor regions (Bernardi 1995). The base composition of coding sequences also appears to reflect the variation in G + C content along chromosomes; genes in the G + C rich parts of the genome tend to have higher G + C contents than genes in the A + T rich parts of the genome. This correlation is manifest at all three codon positions, in introns and in flanking DNA sequences (Bernardi 1995; Clay *et al.* 1996).

The reason for the compositional variation along mammalian chromosomes remains unknown. It has been variously suggested that it is due to mutation bias (Filipski 1987; Sueoka 1988; Wolfe *et al.* 1989), selection (Bernardi and Bernardi 1986), and gene conversion (Holmquist 1992; Eyre-Walker 1993). Unfortunately there has been little decisive evidence published that resolves these alternative hypotheses.

Here I consider whether mutation bias is responsible for variation in the G + C content of synonymous sites, and intron and flanking DNA sequences, using two pop-

ulation genetic tests. The first of these is a derivative of a test suggested by Ballard and Kreitman (1994) and Akashi (1995) [see Eyre-Walker (1997)]; the second is a test suggested by Sawyer *et al.* (1987), as used by Akashi and Schaeffer (1997). If the base composition of a sequence is stationary (*i.e.*, not changing systematically) and maintained solely by mutation bias, the number of A or T mutations segregating as polymorphisms at sites that were ancestrally G or C is expected to be equal to the number of G or C mutations segregating as polymorphisms at sites that were ancestrally A or T, under the infinite-sites model (Eyre-Walker 1997). The reason is as follows. Under neutrality each new mutation has the same (average) probability of fixation. When the base composition of a sequence is stationary, the number of GC → AT substitutions must equal the number of AT → GC substitutions (by definition), and hence there must be equal numbers of GC → AT and AT → GC mutations (Eyre-Walker 1997). For brevity, I will henceforth refer to GC → AT and AT → GC mutations as AT and GC mutations, respectively. The frequency distributions of AT and GC mutations segregating in a population are also expected to be the same under an infinite-sites model, since all neutral mutations have the same (average) frequency distribution (Sawyer *et al.* 1987; Akashi and Schaeffer 1997); *i.e.*, considering only sites that have a polymorphism in our sample, we expect the number of sites segregating a GC mutation at frequency  $f$  to be equal to the number of sites segregating an AT mutation at frequency  $f$ .

There are therefore two independent tests of whether the composition of a sequence is determined by mutation bias: a test of whether the number of GC mutations

Author e-mail: a.c.eyre-walker@sussex.ac.uk

equals the number of AT mutations and a test of whether the frequency distributions of AT and GC mutations are the same. I have applied these two tests to polymorphisms segregating at synonymous, intron, and flanking DNA sites, in the MHC genes of several mammals. I follow Li (1997) and collectively refer to variation segregating at synonymous, intron, and flanking DNA sites as "silent."

MHC genes were used because they are highly polymorphic and have been extensively sequenced in mammals. However, MHC genes pose a number of potential problems for an analysis of this sort. First, the alleles are not a random sample; generally alleles have been sequenced because they have a unique amino acid sequence; and second, there is balancing selection upon some of the amino acid variants (Parham and Ohta 1996). It turns out that neither of these departures from an "ideal" random sample biases either population genetic test. The reason is that A, T, G, and C nucleotides are interspersed along the sequence, so on average A, T, C, and G sites share a common sampling scheme, population history, and genetic history. This property of the MacDonald-Kreitman family of tests (*e.g.*, MacDonald and Kreitman 1991; Akashi 1995; Akashi and Schaeffer 1997; Eyre-Walker 1997) makes them extremely powerful. There is, however, one problem with using MHC genes that is not so easily addressed. Some of the genetic variation segregating in MHC genes is old; *e.g.*, some of the variation in humans predates the origins of the human and old-world monkey lineages (Parham and Ohta 1996). As a consequence, the infinite-sites assumption may not hold, since some polymorphisms may be the product of more than one mutation.

## MATERIALS AND METHODS

**Data:** MHC genes for which multiple alleles have been sequenced within a mammalian species were extracted from alignments held on the EBI server (<ftp://ftp.ebi.ac.uk/pub/databases/hla/>) in the case of the human HLA-A/B/C, DPB1, DQA1, DQB1, and DRB1 genes, or from GenBank. Care was taken to ensure that sequences were genuinely allelic, and not alleles from different paralogous loci. The HLA-CW\*0301 allele was removed from the EBI alignment because it is incorrect according to the associated "readme" file. The transmembrane encoding exon of PERB11.1 was excluded since there is a frameshift in some sequences (Gaudieri *et al.* 1997). Sequences were aligned by hand. For only one gene was there any problem with alignment; in the 5' flank of the human DQA1 gene there is a region of ~50 bp in which there are a number of short repeats. This region was excluded. Alignments, other than those held on the EBI server, can be obtained from the author.

**Stationarity tests:** The two population genetic tests detailed above are only valid when the base composition of the sites under consideration is stationary (*i.e.*, not changing systematically). To test for stationary base composition, I used the test suggested by Eyre-Walker (1994); if the base composition of two homologous sequences is stationary (and sufficiently

diverged so the level of substitution exceeds the level of polymorphism) then the number of sites that are G or C in one sequence and A or T in another equals the number of sites that are A or T in the first sequence and G or C in the second (Eyre-Walker 1994). This is true even if there are multiple substitutions, or the lineages evolve at different rates (Eyre-Walker 1994). For each gene (for which there were multiple alleles) a consensus was constructed across alleles by taking the most common codon, or nucleotide in the case of noncoding DNA, at each site. Polymorphic sites were not removed from the analysis since this biases the analysis. Partial sequences were completed with data from a randomly picked allele from the appropriate locus. A complete sheep DQA2 sequence was not available. Each consensus was compared against a number of homologous sequences, either paralogous genes from the same species (*e.g.*, human HLA-A and HLA-B) or a homologous gene from a different species (*e.g.*, human DQB1 and cow DQB1). Paralogy and orthology are not always easy to determine for genes from different species since MHC genes duplicate frequently. This is unlikely to bias the results, since it is more likely that two paralogues will diverge in composition, than two orthologues. Since there were no significant differences in composition among the three human class I genes and between the two mouse class I genes, a consensus for each species, across loci, was constructed for tests involving class I genes from different species. No suitable outgroup sequences could be found for the human DPB1 gene and most of the noncoding sequences. The sequences were aligned by hand.

**Infinite sites:** Both population genetic tests assume that each polymorphism we observe arose from a single mutation. To investigate whether the data are consistent with the infinite-sites model, two statistics were calculated for each coding, exon, intron, and flanking DNA sequence: the average pairwise silent site divergence, otherwise known as the nucleotide diversity ( $\pi$ ), and the maximum silent site divergence between alleles. The maximum silent site divergence was calculated to ensure there were not large differences between allelic classes. In an ideal population  $\pi = 4N_e u / (1 + 8N_e u)$  (Crow and Kimura 1970), this reduces to  $4N_e u$  under the infinite-sites model. Hence if  $\pi < 0.1$  the model conforms fairly closely to the infinite-sites model. Silent site divergences were calculated as the proportion of silent sites that had a silent difference. The analysis was restricted to those codons with no nonsynonymous variation segregating in the sample and to codons for which there were at least 10 sequences.

**Polymorphism analysis:** If a sequence is subject solely to mutation bias then under the infinite-sites model the number of GC mutations equals the number of AT mutations, and their frequency distributions are the same. The direction of the mutation that gave rise to each polymorphism was inferred from the frequency of the alleles segregating at each site; the more common nucleotide was inferred to be the ancestral nucleotide. This method is unbiased under the null hypothesis and was chosen because parsimony is unreliable when base compositions are biased (Collins *et al.* 1994; Perna and Kocher 1995; Eyre-Walker 1998). The method is unbiased because each neutral mutation has the same expected frequency distribution; *i.e.*, the probability that an AT mutation will reach a frequency  $f$  is equal to the probability that a GC mutation reaches the same frequency, under the infinite-sites model. If more than two nucleotides were segregating at a site, the two most frequent alleles were used in the analysis; out of 514 polymorphisms at threefold, fourfold, and noncoding sites, only 15 had more than two nucleotides segregating (which suggests the data conform to the infinite-sites model). The analysis of polymorphisms segregating in MHC sequences was restricted to data sets in which there were at least 10

sequences. Sites polymorphic for A and T or G and C were ignored. This does not bias the analysis.

Whether the number of AT mutations segregating was the same as the number of GC mutations in a gene was tested using a two-tailed binomial test. The frequency distributions of AT and GC mutations were compared for each gene using a two-tailed Mann-Whitney test. The difference in the average frequency of AT and GC mutations was also calculated for each gene. Note the average frequency in each case was calculated using only those sites at which a polymorphism was segregating. A two-tailed single sample *t*-test was performed on these frequency differences to test whether GC mutations are on average segregating at higher or lower frequencies than AT mutations across genes.

**Shared polymorphisms:** The three human class I genes and two mouse class I genes are fairly closely related, so they were checked for shared polymorphisms. HLA-A, -B, and -C coding sequences share 8 out of 95 scored synonymous AT and GC polymorphisms; HLA-A, -B, and -C intron 1 sequences share 2 out of 30 polymorphisms; H-2K and H-2D share 4 out of 50 polymorphisms; and H-2D and H-2K 3' flanking sequences share 3 out of 36 polymorphisms. The results from these genes are therefore largely independent.

**CpG islands:** CpG islands were identified using two techniques. First, the position of CpG and GpC dinucleotides was plotted along the genomic sequence or, when the genomic sequence was not available, along the mRNA sequence. CpG islands are evident on this basis as regions in which the frequency of CpG is similar to the frequency of GpC, whereas outside the island, the frequency of CpG is considerably lower [see examples in Bird (1987)]. Second, the observed frequency of CpG dinucleotides was calculated for every exon, intron, and flanking sequence in the analysis, along with the frequency expected from the base composition. An intron or exon was deemed to be in a CpG island if the observed over expected frequency was above 0.6 (Gardiner-Garden and Frommer 1987). There was almost perfect agreement between the two methods. The one exception was exon 1 of some of the human class I alleles that have observed-over-expected CpG values just below 0.6; however, the exon is flanked by DNA with ratios well in excess of 0.6: 200 bp upstream has an observed-over-expected CpG ratio of 0.93, and intron 1 has a value of 0.81. The exon was therefore taken to be part of the CpG island. The following sequences were classified as being in a CpG island: exons 1-3 of all class I genes (*i.e.*, human HLA-A, HLA-B, and HLA-C; mouse H-2D and H-2K); exon 2 of all class II  $\beta$  chain genes except H-2M- $\beta$ 2 (*i.e.*, human DPB1, DQB1, and DRB1; mouse H2-IA- $\beta$  and H2-IE- $\beta$ ); and intron 1 of all class I genes (*i.e.*, human HLA-A, HLA-B, and HLA-C intron 1). All other sequences were classified as non-CpG island.

## RESULTS

Before we can test whether the base composition of silent sites is solely due to mutation bias we need to ensure that the base composition of the sequences is stationary and that the infinite-sites model holds. If both of these conditions are met, and the G + C content of silent sites is due to mutation bias alone, then we expect the number of silent AT mutations segregating in a population to be equal to the number of GC mutations and their frequency distributions to be same.

**Stationary base composition:** If the base composition of a sequence is changing in a systematic fashion, it is

**TABLE 1**  
**Testing for stationary G + C content**

Sequence 1	Sequence 2	GC <sub>1</sub> :AT <sub>2</sub>	AT <sub>1</sub> :GC <sub>2</sub>
Human HLA-A	Human HLA-B	18	13
Human HLA-A	Human HLA-C	18	13
Human HLA-B	Human HLA-C	8	10
Mouse H-2D	Mouse H-2K	11	7
Human class I	Mouse class I	25	18
Human class I	<i>Cow class I</i>	24	20
Mouse class I	<i>Cow class I</i>	17	22
Human DQA1	Human DQA2	6	9
Human DQA1	Sheep DQA	7	6
Human DQA1	Mouse H2-IA- $\alpha$	22	20
Sheep DQA	Mouse H2-IA- $\alpha$	11	6
Human DQB1	Human DQB2	9	4
Human DQB1	Mouse H2-IA- $\beta$	15	24
Human DQB1	Cow DQB1	16	20
Mouse H2-IA- $\beta$	Cow DQB1	24	22
Cow DQB1	<i>Cow DQB2</i>	7	6
Human DRB1	<i>Human DRB2</i>	4	3
Human DRB1	Mouse H2-IE- $\beta$	21	27
Human DRB1	Cow DRB3	13	19
Mouse H2-IE- $\beta$	Cow DRB3	27	26
Mouse H-2M- $\alpha$	<i>Human DMA</i>	42	19**
Mouse H-2M- $\alpha$	<i>Cow DMA</i>	34	25
Mouse H-2M- $\beta$ 2	<i>Human DMB</i>	34	28
HLA-A intron 1	HLA-B intron 1	4	4
HLA-A intron 1	HLA-C intron 1	4	4
HLA-B intron 1	HLA-C intron 1	4	4
H-2D 3' flank	H-2K 3' flank	10	7

The table gives the number of silent sites that are G or C in one sequence and A or T in the other. Genes for which multiple alleles are available are in roman typeface, genes used to test for stationary base composition are shown in italics. \*\*  $P < 0.01$ .

possible to explain any pattern of polymorphism in terms of changes in the pattern of mutation (Eyre-Walker 1997). To test for stationarity, I used a test suggested by Eyre-Walker (1994): if the base composition of two homologous sequences is stationary, then the number of sites that are G or C in one sequence and A or T in another equals the number of sites that are A or T in the first and G or C in the second (Eyre-Walker 1994). Table 1 shows that for only one sequence in the data set is there any evidence that silent site base composition is not stationary; the mouse H2-M- $\alpha$  gene is significantly higher in G + C content than the human DMA gene.

**Infinite-sites model:** Under the infinite-sites model each mutation occurs at a unique site, so each polymorphism is the product of a single mutation; *i.e.*, there have been no multiple hits. To investigate whether the infinite-sites model is reasonable for silent site polymorphisms in MHC genes, the average and the maximum pairwise silent site divergences between alleles were calculated for each sequence (Table 2). The values are

**TABLE 2**  
**Investigating the infinite-sites assumption**

Species	Gene	Average pairwise silent site divergence	Maximum pairwise silent site divergence
Synonymous			
Human	DPB1	0.004	0.048
	DQA1	0.041	0.100
	DQB1	0.045	0.139
	DRB1 <sup>a</sup>	0.029	0.102
	HLA-A	0.026	0.057
	HLA-B	0.026	0.085
	HLA-C	0.017	0.045
Mouse	PERB 11.1	0.011	0.032
	H2IA-β	0.017	0.063
	H2-IE-β	0.018	0.060
	H-2D	0.023	0.055
	H-2K	0.029	0.053
	H-2M-α	0.009	0.024
	H-2M-β2	0.007	0.017
Others	Cow DQB1	0.024	0.085
	Cow DRB3	0.015	0.058
	Sheep DQA2	0.060	0.152
Noncoding			
Human	DQA1 5' flank	0.041	0.083
	DQA1 intron 1	0.049	0.109
	HLA-A intron 1	0.045	0.085
	HLA-A intron 3	0.013	0.030
	HLA-B intron 1	0.025	0.055
	HLA-C intron 1	0.017	0.038
	PERB intron W	0.008	0.023
	PERB intron X	0.007	0.018
	PERB intron Y	0.008	0.039
	PERB intron Z	0.003	0.020
Mouse	H2-IE-β intron 2	0.005	0.013
	H-2D 3' flank	0.015	0.036
	H-2K 3' flank	0.044	0.105

<sup>a</sup> Values given for DRB1 exon 2 sequences.

low; the maximum silent site divergence between any pair of alleles is 0.15, and most of the nucleotide diversities are below 0.05. The average and maximum pairwise divergences for individual exons are also low and fairly consistent along each gene; *e.g.*, for HLA-A the nucleotide diversities for the eight exons are 0.039, 0.031, 0.016, 0.033, 0.024, 0.027, 0.003, and 0.000; for the three DQB1 exons for which we have data, they are 0.049, 0.022, and 0.045. In an ideal population this level of nucleotide diversity would be consistent with the infinite-sites model; *i.e.*, when  $\pi = 0.05$ ,  $4N_e u / (1 + 8N_e u) \approx 4N_e u$ . The data therefore suggest that the infinite-sites assumption is reasonable, a conjecture supported by the observation that out of 514 polymorphic sites that are able to have more than two nucleotides segregating, only 15 sites actually have more than two.

**Polymorphism analysis:** Since the sequences appear to be stationary in base composition and conform to the infinite-sites model, we expect the number of AT mutations segregating to equal the number of GC mutations, if mutation is responsible for the bias in silent site G + C content. For coding sequences, however, the number of synonymous AT mutations exceeds the number of GC mutations with only one exception in 15 MHC genes that show a difference ( $P < 0.001$ ; Table 3). The difference is significant for many genes individually, for all three groups of organisms considered (Table 3), and overall (Table 4).

The frequency distributions of synonymous GC and AT mutations are also expected to be the same if mutation bias is responsible for synonymous codon bias. For 11 out of 12 genes for which there are both AT and GC mutations segregating, the average frequency of GC mutations is greater than the average frequency of AT mutations ( $P < 0.01$ ). The average frequency of GC mutations is also significantly greater than the average frequency of AT mutations across human genes and over the whole data set (Table 4). The average difference in the frequency of GC and AT mutations is also nearly significant for the mouse genes ( $P < 0.08$ ).

The patterns are similar for noncoding sequences. In all but two sequences the number of AT mutations is greater than the number of GC mutations ( $P < 0.05$ ). The number of AT mutations is significantly greater than the number of GC mutations for a number of genes individually, in both humans and mice, and over the whole data set (Table 4). However, the average frequency of GC mutations is only significantly greater than the average frequency of AT mutations for the mouse sequences. This may be due to the fact that the frequency distribution test is weaker than the numbers of polymorphisms test.

**Sequencing errors:** These patterns of polymorphism are not due to sequencing errors. If we repeat the analyses removing all singletons (*i.e.*, sites at which one of the alleles is present as a single copy), the results remain qualitatively unchanged (Table 4). There is a large excess of AT mutations segregating in both coding and noncoding sequences, and for synonymous mutations the average frequency of GC mutations is significantly greater than the average frequency of AT mutations.

**CpG islands:** Many of the MHC genes contain a CpG island (see materials and methods for a list). These are short (~1 kb), G + C biased sequences, rich in the dinucleotide CpG, which have been implicated in the control of gene expression (Lewis and Bird 1991; Cross and Bird 1995). The results might therefore reflect processes that only affect the composition of CpG islands. This does not appear to be the case; in the non-CpG-island DNA there is a significant excess of AT mutations in both coding and noncoding DNA (Table 4). However, in neither is the average frequency of GC mutations significantly greater than the average

**TABLE 3**  
**The pattern of silent polymorphism in MHC genes**

Species	Gene	No. of alleles	No. of sites	G + C (%)	AT:GC	Frequency difference
Synonymous						
Human	DPB1 <sup>a</sup>	65	103	83.2	5:0	—
	DQA1 <sup>a</sup>	15	226	59.2	11:7	0.070
	DQB1 <sup>a</sup>	25	222	77.1	19:3***	0.037
	DRB1 <sup>a</sup>	135/32 <sup>b</sup>	231	69.9	21:10	0.123*
	HLA-A	58	347	77.8	28:6***	-0.004
	HLA-B	125	350	77.7	20:12	0.014
	HLA-C	34	351	78.0	19:10	0.045
	PERB11.1 <sup>a</sup>	15	174	61.3	1:1	0.133
	Overall				124:49***	0.060*
Mouse	H2-IA-β <sup>a</sup>	24	187	83.3	10:5	0.013
	H2-IE-β <sup>a</sup>	18	89	77.1	9:0**	—
	H-2D	13	317	70.8	16:6	0.054
	H-2K	20	351	70.2	14:14	0.155**
	H-2M-α	18	251	66.9	6:0*	—
	H-2M-β2	22	248	63.1	2:5	0.137
		Overall				57:30**
Others	Cow DQB1 <sup>a</sup>	13	79	91.7	5:0	—
	Cow DRB3 <sup>a</sup>	21	81	82.0	6:0*	—
	Sheep DQA2 <sup>a</sup>	13	73	49.2	8:5	0.032
		Overall				19:5**
Noncoding						
Human	DQA1 5' flank	19	650	51.0	35:32	0.019
	DQA1 intron 1	15	437	38.4	33:34	-0.026
	HLA-A intron 1	14	129	76.6	12:0***	—
	HLA-A intron 3	16	291	52.4	7:3	0.071
	HLA-B intron 1	23	128	76.6	9:1*	0.125
	HLA-C intron 1	11	130	77.1	7:1	-0.039
	PERB intron W	15	281	51.2	9:1*	-0.057
	PERB intron X	15	164	61.8	4:1	-0.033
	PERB intron Y	15	51	61.1	2:0	—
	PERB intron Z	14	82	57.3	1:0	—
		Overall				119:73**
Mouse	H2-IE-β intron 2	11	1490	49.5	16:3**	0.044
	H-2D 3' flank	10	304	53.5	16:10	0.036
	H-2K 3' flank	16	336	50.5	4:6	0.034
		Overall				36:19*

The numbers of AT and GC silent mutations and the average frequency of GC mutations, minus the average frequency of AT mutations, segregating in mammalian MHC genes, along with the number of sites and alleles analyzed are given. G + C (%) is the G + C content of the sites analyzed. The overall frequency difference figure is the mean across genes. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ .

<sup>a</sup>Partial sequences. Most of exon 1 is missing for human DQA1, DQB1, and DRB1, and mouse H2-IA-β; only exon 2 is available for human DPB1, mouse H2-IE-β, cow DQB1 and DRB3, and sheep DQA2. Only two exons of PERB11.1 are available.

<sup>b</sup>Of the DRB1 alleles 32 are largely complete; however, there are an additional 103 exon 2 alleles that have been sequenced.

frequency of AT mutations. In the CpG-island DNA there is an extreme bias toward AT mutations in both coding and noncoding DNA, and GC mutations segregate at significantly higher frequencies than AT mutations.

The CpG-island DNA may give more significant results because on average CpG-island DNA is more G + C rich than non-CpG-island DNA (0.856 vs. 0.625 for coding sequences; 0.768 vs. 0.500 for noncoding). Under the alternative hypotheses of selection and biased

**TABLE 4**  
**Summary of polymorphism results**

	Coding (synonymous)		Noncoding (intron/flanking)		Overall	
	AT:GC	Frequency difference	AT:GC	Frequency difference	AT:GC	Frequency difference
Overall	200:84***	0.067 ± 0.016**	155:92***	0.017 ± 0.018	355:176***	0.045 ± 0.013**
No singletons	139:65***	0.059 ± 0.021*	109:77*	0.004 ± 0.024	248:142***	0.037 ± 0.017*
Non-CpG island	87:54**	0.055 ± 0.032	127:90*	0.011 ± 0.016	214:144***	0.037 ± 0.020
CpG island	113:30***	0.119 ± 0.018***	28:2***	0.043 ± 0.082	141:32***	0.104 ± 0.022***

The frequency difference figure is the mean difference across genes between the mean frequency of GC and the mean frequency of AT mutations. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ .

gene conversion, the difference between the number of AT and GC mutations, and the difference between their average frequencies, are expected to be greater in sequences of high G + C content (see below).

#### DISCUSSION

The pattern of silent site polymorphism in mammalian MHC sequences is inconsistent with the action of mutation bias: we expect equal numbers of AT and GC mutations segregating in the population, but we observe significantly more AT mutations than GC mutations; furthermore, we expect GC and AT mutations to be segregating at similar frequencies, but silent GC mutations are segregating at significantly higher frequencies than AT mutations.

**Stationarity:** If the sequences are not stationary, however, then it is always possible to explain the pattern of polymorphism in terms of changes in the pattern of mutation (Eyre-Walker 1997). For example, if the mutation pattern had changed in favor of AT sometime in the past, we would expect an excess of AT mutations segregating (Eyre-Walker 1997); furthermore, if the change was very recent we would expect AT mutations to be segregating at lower frequencies than GC mutations, because the change in the mutation pattern would be reflected in the youngest mutations that segregate at the lowest frequencies (Akashi and Schaeffer 1997; Eyre-Walker 1997).

Three observations suggest that the pattern of silent site polymorphism is not due to a change in the pattern of mutation. First, there is no evidence that the sequences are undergoing systematic changes in composition (Table 1). Second, there would have to have been at least two independent changes in the mutation pattern to explain the synonymous site data. Since none of the species used in this study share polymorphisms, we would need one recent change in the mutation pattern to explain the significant synonymous site frequency distribution result in humans. We would then need another change to explain the significant excess of AT mutations in the other organisms. If we note that

the frequency distribution result is also approaching a significant level in mice, we might need as many as three independent changes in the mutation pattern, in the same direction, to explain the data. This seems improbable. Third, the change in the mutation pattern would have to be extreme. We can estimate the change in the mutation pattern that would explain the polymorphism data and express it as the change in the G + C content that would eventually result [see Equation 9 in Eyre-Walker (1997)]. On average, synonymous, intron, and flanking DNA sites used in this analysis would be expected to decline in G + C content by 31.3% ( $\pm 6.0\%$ ) [*i.e.*, the G + C content is expected to decline on average from  $x\%$  to  $(x - 31.3)\%$ ]. This is a dramatic shift in composition. It therefore seems unlikely that the aberrant pattern of silent site polymorphism in MHC sequences is caused by changes in the pattern of mutation bias.

**Infinite sites:** If the infinite-sites assumption of the two tests is violated we expect an excess of AT mutations segregating, with GC mutations segregating at higher frequencies on average, in G + C biased sequences. However, two lines of evidence suggest that the infinite-sites assumption is reasonable. First, only 3% of the polymorphic sites have more than two nucleotides segregating (excluding twofold degenerate sites); the number of multiple hits must therefore be low. Second, the silent site nucleotide diversities are low; for most genes the nucleotide diversity is  $< 0.05$ , and this is also true for synonymous variation in most individual exon sequences. This level of nucleotide diversity would be consistent with the infinite-sites assumption in an ideal population since  $4N_e u / (1 + 8N_e u) \approx 4N_e u$  when  $\pi < 0.1$ .

It is possible to calculate the proportion of mutations that would be classified as AT in an ideal population if mutation bias is the cause of compositional bias (see appendix). The proportion of mutations classified as AT increases with  $\pi$  and, to some extent, with sample size (Figure 1). For all but one gene, sheep DQA2, the nucleotide diversity is below 0.05, and the average across genes is 0.023; with  $\pi = 0.05$  we expect at most 56% of

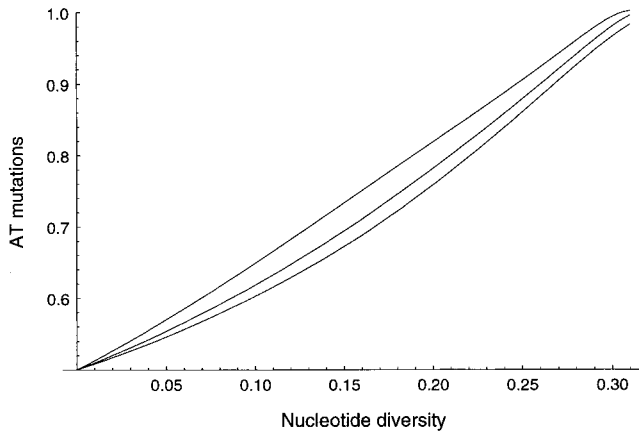


Figure 1.—The proportion of neutral mutations segregating in a population that would be classified as an AT mutation, when the G + C content is 80%, as a function of the nucleotide diversity (see appendix, Equations A3 and A4). The lines, from bottom to top, are for sample sizes of 10, 50, and 100 sequences, respectively.

the mutations to be classified as AT in sequences that are 80% G + C. In contrast, 80% ( $\pm 4\%$ ) of the polymorphisms in sequences (individual exons, introns, and flanking sequences) with silent site G + C contents between 75 and 85% are AT in the MHC sequences examined. To explain these values  $\pi$  would have to be  $\sim 0.2$ , which is  $\sim 10$  times greater than the average nucleotide diversity observed.

It is ultimately difficult to conclusively prove that the results are not due to departures from an infinite-sites model. However, the available evidence suggests the infinite-sites assumption is reasonable. It therefore seems unlikely that the patterns of silent site polymorphism in MHC genes are due to mutation biases.

**Alternative explanations:** The pattern of silent site polymorphism is, however, consistent with either selection or biased gene conversion (BGC). Under both of these hypotheses the pattern of polymorphism is expected to be biased in one direction, with the less common type of mutation segregating at higher frequencies on average. This can be demonstrated formally for directional selection and biased gene conversion, but it is also evident from the following intuitive argument, which also applies to stabilizing selection. Let us assume for simplicity that the rate at which a G or C site mutates to an A or T is equal to the rate of the reverse process (*i.e.*, there is no mutation bias). Consider a sequence in which selection or biased gene conversion has elevated the G + C content from 50 to 80%; it is evident that 80% of the new mutations in each generation must be GC  $\rightarrow$  AT changes, and 20% AT  $\rightarrow$  GC changes (ignoring A  $\leftrightarrow$  T and G  $\leftrightarrow$  C mutations). Furthermore, if we assume the sequence is stationary in composition, the number of GC  $\rightarrow$  AT substitutions must equal the number of AT  $\rightarrow$  GC substitutions; therefore the fixation probability of each GC mutation must be higher

than each AT mutation, since GC mutations are rarer; hence we expect GC mutations to segregate at higher frequencies. The results from the MHC genes are therefore consistent with selection or BGC acting in most sequences to elevate the G + C content above that expected under mutation alone.

**Gene conversion:** Gene conversion is thought to have played a major role in generating haplotype diversity in MHC genes (Parham and Ohta 1996). However, this does not imply that gene conversion is either biased or common. Gene conversion can be unbiased (Szostak *et al.* 1983), but to affect composition, gene conversion must be biased and frequent: *i.e.*,  $N_e w \approx 1$  to affect base composition (Eyre-Walker 1992), where  $(w + 1)/2$  is the proportion of gametes with one of the alleles produced by a heterozygote, and  $N_e$  is the effective population size. To generate significant haplotype diversity the rate of gene conversion can be similar to the rate of mutation, or even below it, if gene conversion produces selectively advantageous haplotypes more frequently than mutation. Since  $N_e u \sim 0.001$  in humans (Li and Stadler 1991) the rate of gene conversion can be several orders of magnitude lower than the level needed to affect base composition, while still producing haplotype diversity (note that BGC of the level required to affect G + C content does not reduce nucleotide diversity significantly).

Two observations suggest silent site G + C content of the MHC genes is not affected by biased gene conversion. First, the process of recombination has been extensively studied in the mouse MHC, but BGC has almost never been observed either in recombinant or nonrecombinant chromosomes (Khambata 1996). Biased gene conversion is thought to occur by the biased repair of mismatches, formed in heteroduplex DNA during recombination (Szostak *et al.* 1983). If heteroduplex DNA spans several markers we expect to see a mosaic pattern of segregation (Sant'Angelo *et al.* 1992); *i.e.*, in a cross of two haplotypes, A and B, we expect to see chromosomes of the type AABABB. This pattern has been observed only once out of more than 80 crosses (Uematsu *et al.* 1986). Instead, chromosomes are almost always AAABBB (Kobori *et al.* 1986; Padgett *et al.* 1991; Shiroishi *et al.* 1991; Sant'Angelo *et al.* 1992; Khambata *et al.* 1996). Second, synonymous site (*i.e.*, third codon position) G + C content is generally greater than intron G + C content for most genes (Clay *et al.* 1996). To explain this, BGC would have to be more frequent in exons, yet recombinational breakpoints, in the MHC at least, are concentrated in hotspots in introns and intergenic DNA (Khambata *et al.* 1996; Cullen *et al.* 1997). Finally, there is a problem of parameter sensitivity (Eyre-Walker 1992). BGC has little effect on base composition if  $N_e w \ll 1$ , but leads to extreme base composition bias if  $N_e w \gg 1$ . There is only a one order of magnitude window in which BGC will give intermediate levels of base composition bias.

**Selection:** Directional and stabilizing selection are both consistent with the population genetic results reported here. Furthermore, stabilizing selection can explain why third position G + C content is greater than intron G + C content. If there is an optimal G + C content for a sequence, third position G + C content would be expected to compensate for the constraint upon the first two codon positions imposed by protein function and, hence, be greater than intron G + C content (Hughes and Yeager 1997). In support of this model, exon and intron G + C contents for individual genes are very similar in humans (Clay *et al.* 1996); on average the exons from a gene are ~6% more G + C rich than the introns in humans (41 genes from Aissani *et al.* 1991) and only 3% in rats and mice [42 genes from Hughes and Yeager (1997)].

**CpG islands:** CpG islands have high G + C content. This is perhaps not surprising given that CpG dinucleotides occur at very low frequencies outside CpG islands. However, the difference in G + C content between island and nonisland DNA appears to be greater than one would expect from the suppression of CpGs in nonisland DNA. For example, the G + C content of synonymous sites in our sample is 86% in the island, compared to 63% in the nonisland DNA; it is 77 vs. 50% in the noncoding sequences. This difference appears too great to be attributed to the removal of one dinucleotide. The results presented here suggest that there might be selection to increase the G + C content of the CpG-island DNA. This may not be surprising given that CpG islands have been implicated in the control of gene expression (Bird 1987; Cross and Bird 1995).

**Isochores:** Third codon position and intron G + C contents are highly correlated to the G + C content of a large block (>100 kb) of DNA, or isochore, in which a gene resides (Bernardi *et al.* 1985; Clay *et al.* 1996). This would seem to imply that whatever causes compositional variation at synonymous, intron, and flanking DNA sites also causes the large-scale variation in composition through intergenic "junk" DNA. One might suggest that the G + C contents of the third codon position, intron, and flanking DNA are affected by several factors, only one of which affects intergenic DNA. However, this seems unlikely, for although third position G + C content is considerably greater than either intron or isochore G + C content, intron composition is quite similar to isochore G + C content (Clay *et al.* 1996). Furthermore, the regression slope between third position and isochore G + C content is significantly greater than one [ $b = 2.46 \pm 0.33$  for the 34 genes from Clay *et al.* (1996)]; this would seem to rule out models in which two factors affecting synonymous codon use are uncorrelated; if they were uncorrelated one might expect the slope to be approximately one. The present results therefore imply that compositional variation in noncoding DNA is not (solely) due to mutation bias,

and that it may well be caused by selection. What the selection might be remains unclear.

I am very grateful to Monty Slatkin who suggested using MHC genes in this analysis and to Hiroshi Akashi, Nick Barton, Adrian Bird, Brian Charlesworth, Brandon Gaut, Shirin Khambata, Richard Kliman, Monty Slatkin, John Maynard Smith, Colm O'Heuigin, Joel Peck, three anonymous referees, and Jody Hey for many helpful discussions and comments on this manuscript. The author is supported by the Royal Society.

#### LITERATURE CITED

- Aissani, B., G. D'onofrio, D. Mouchiroud, K. Gardiner, C. Gautier *et al.*, 1991 The compositional properties of human genes. *J. Mol. Evol.* **32**: 493-503.
- Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139**: 1067-1076.
- Akashi, H., and S. W. Schaeffer, 1997 Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. *Genetics* **146**: 295-307.
- Ballard, J. W., and M. Kreitman, 1994 Unraveling selection in the mitochondrial genome of *Drosophila*. *Genetics* **138**: 757-772.
- Bernardi, G., 1995 The human genome: organization and evolutionary history. *Annu. Rev. Genet.* **29**: 445-476.
- Bernardi, G., and G. Bernardi, 1986 Compositional constraints and genome evolution. *J. Mol. Evol.* **24**: 1-11.
- Bernardi, G., B. Olafsson, J. Filipski, M. Zerial, J. Salinas *et al.*, 1985 The mosaic genome of warm blooded vertebrates. *Science* **228**: 953-958.
- Bird, A. P., 1987 CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* **3**: 342-347.
- Clay, O., S. Caccio, Z. Zoubak, D. Mouchiroud and G. Bernardi, 1996 Human coding and noncoding DNA: compositional correlations. *Mol. Phylogenet. Evol.* **5**: 2-12.
- Collins, T. M., P. H. Wimberger and G. J. P. Naylor, 1994 Compositional bias, character state bias and character state reconstruction using parsimony. *Syst. Biol.* **43**: 482-496.
- Cross, S. H., and A. P. Bird, 1995 CpG islands and genes. *Curr. Opin. Genet. Dev.* **5**: 309-314.
- Crow, J., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.
- Cullen, M., J. Noble, H. Erlich, K. Thorpe, S. Beck *et al.*, 1997 Characterization of recombination in the HLA class II region. *Am. J. Hum. Genet.* **60**: 397-407.
- Eyre-Walker, A., 1992 Studies of synonymous codon evolution in mammals. Ph.D. Thesis. University of Edinburgh, Edinburgh, Scotland.
- Eyre-Walker, A., 1993 Recombination and mammalian genome evolution. *Proc. R. Soc. Lond. Ser. B* **252**: 237-243.
- Eyre-Walker, A., 1994 DNA mismatch repair and synonymous codon evolution in mammals. *Mol. Biol. Evol.* **11**: 88-98.
- Eyre-Walker, A., 1997 Differentiating selection and mutation bias. *Genetics* **147**: 1983-1987.
- Eyre-Walker, A., 1998 Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* **47**: 686-690.
- Filipski, J., 1987 Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett.* **217**: 184-186.
- Gardiner-Garden, M., and M. Frommer, 1987 CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**: 261-282.
- Gaudieri, S., C. Leelayuwat, D. C. Townend, J. Mullberg, D. Cosman *et al.*, 1997 Allelic and interlocus comparison of the PERB11 multigene family in the MHC. *Immunogenetics* **45**: 209-216.
- Holmquist, G. P., 1992 Chromosome bands, their chromatin flavors and their functional features. *Am. J. Hum. Genet.* **51**: 17-37.
- Hughes, A. L., and M. Yeager, 1997 Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* **45**: 125-130.
- Khambata, S., 1996 Analysis of the meiotic recombinational hotspot

- associated with the *Ea* gene of the mouse major histocompatibility complex. Ph.D. Thesis. Rutgers University, New Brunswick, NJ.
- Khambata, S., J. Mody, A. Modzelewski, D. Heine and H. C. Passmore, 1996 *Ea* recombinational hot spot in the mouse major histocompatibility complex maps to the fourth intron of the *Ea* gene. *Genome Res.* **6**: 195–201.
- Kobori, J. A., E. Strauss, K. Minard and L. Hood, 1986 Molecular analysis of the hotspot of recombination in the murine major histocompatibility. *Science* **234**: 173–179.
- Lewis, J., and A. P. Bird, 1991 DNA methylation and chromatin structure. *FEBS Lett.* **205**: 155–159.
- Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- Li, W.-H., and L. A. Stadler, 1991 Low nucleotide diversity in man. *Genetics* **129**: 513–523.
- MacDonald, J. H., and M. Kreitman, 1991 Adaptive evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–654.
- Padgett, K. A., D. C. Shreffler and B. K. Sata, 1991 Molecular mapping of murine *I* region recombinants. *J. Immunol.* **147**: 2764–2770.
- Parham, P., and T. Ohta, 1996 Population biology of antigen presentation by MHC class I molecules. *Science* **272**: 67–74.
- Perna, N. T., and T. D. Kocher, 1995 Unequal base frequencies and the estimation of substitution rates. *Mol. Biol. Evol.* **12**: 359–361.
- Sant'Angelo, D. B., W. P. Lafuse and H. C. Passmore, 1992 Evidence that nucleotide sequence identity is a requirement for meiotic crossing over within the mouse *Eb* recombinational hot spot. *Genomics* **13**: 1334–1336.
- Sawyer, S., D. E. Dykhuizen and D. L. Hartl, 1987 Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Natl. Acad. Sci. USA* **84**: 6225–6228.
- Shiroishi, T., T. Sagai, N. Hanzawa, H. Gotoh and K. Moriwaki, 1991 Genetic control of sex-dependent meiotic recombination in the major histocompatibility complex of the mouse. *EMBO J.* **10**: 681–686.
- Suoeka, N., 1988 Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**: 2653–2657.
- Szostak, J. W., T. L. Orr-Weaver, R. J. Rothstein and F. W. Stahl, 1983 The double-strand-break model for recombination. *Cell* **33**: 25–35.
- Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- Uematsu, Y., H. Kiefer, R. Schulze, K. Fischer-Lindahl and M. Steinmetz, 1986 Molecular characterization of a meiotic recombinational hotspot enhancing homologous equal crossing-over. *EMBO J.* **5**: 2123–2129.
- Wolfe, K. H., P. M. Sharp and W.-H. Li, 1989 Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283–285.

Communicating editor: J. Hey

## APPENDIX

Let us consider a two-allele system in which the mutation rate from GC to AT is  $u$  and the mutation rate from AT to GC is  $v$ . In a diploid the equilibrium distribution of the gene frequency of GC mutations,  $x$ , is

$$F(x) \propto x^{v-1}(1-x)^{u-1}, \quad (\text{A1})$$

where

$$U = 4N_e u$$

$$V = 4N_e v,$$

where  $N_e$  is the effective population size (Crow and Kimura 1970). The probability of observing  $i$  GC and  $n-i$  AT nucleotides at a site in a sample of  $n$  sequences is therefore

$$G(i) = \frac{n!}{i!(n-i)!} \int_0^1 F(x) x^i (1-x)^{n-i} dx \quad (\text{A2})$$

(Tajima 1989), and the proportion of polymorphisms categorized as AT mutations

$$M_{AT} = \frac{\sum_{i=n/2+1}^{n-1} G(i)}{\sum_{i=1, x \neq n/2}^{n-1} G(i)} \quad n \text{ even}$$

$$M_{AT} = \frac{\sum_{i=n/2+1/2}^{n-1} G(i)}{\sum_{i=1}^{n-1} G(i)} \quad n \text{ odd.} \quad (\text{A3})$$

The expected nucleotide diversity is

$$\hat{\pi} = \frac{n}{n-1} \frac{\sum_{i=1}^{n-1} G(i) 2(i/n)(1-i/n)}{\sum_{i=0}^n G(i)}. \quad (\text{A4})$$

The factor  $n/(n-1)$  corrects for bias caused by sampling error.

