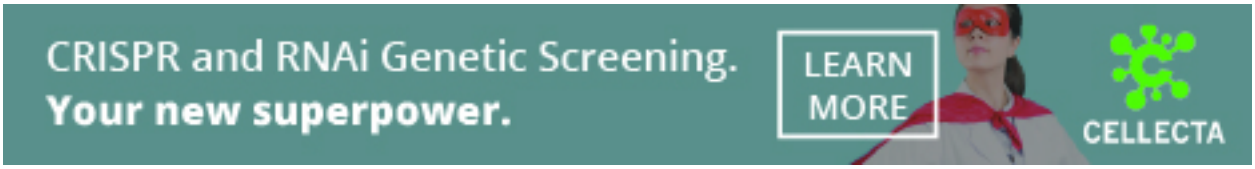
The Distribution of Variation in Regulatory Gene Segments, as Present in MHC Class II Promoters

Lindsay G. Cowell, Thomas B. Kepler, Michael Janitz, et al.

Genome Res. 1998 8: 124-134
Access the most recent version at doi:10.1101/gr.8.2.124



RESEARCH

The Distribution of Variation in Regulatory Gene Segments, as Present in MHC Class II Promoters

Lindsay G. Cowell,[1,2] Thomas B. Kepler,[2,3] Michael Janitz,[1] Roland Lauster,[1] and N. Avrion Mitchison[1]

[1]Deutsches RheumaForschungsZentrum, D-10117 Berlin, Germany; [2]Biomathematics Program, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203 USA

Diversity in the antigen-binding receptors of the immune system has long been a primary interest of biologists. Recently it has been suggested that polymorphism in regulatory (noncoding) gene segments is of substantial importance as well. Here, we survey the level of variation in MHC class II gene promoters in man and mouse using extensive collections of published sequences together with unpublished sequences recently deposited by us in the EMBL gene bank using the Shannon entropy to quantify diversity. For comparison, we also apply our analysis to distantly related MHC class II promoters, as well as to class I promoters and to class II coding regions. We observe a high level of intraspecies variability, which in mouse but not in man is localized to a significant extent near the binding sites of transcription factors—sites that are conserved over longer evolutionary distances. This localization may both indicate and enhance heterozygote advantage, as the presence of two functionally different promoters would be expected to confer flexibility in the immune response.

A surge of interest in the variability of regulatory, noncoding gene segments is under way. Collections have been made of promoter sequences for MHC class I and class II genes in man (Louis et al. 1993; Yao et al. 1995) and class II in mouse (Janitz et al. 1997a), which add substantially to the information previously available for promoter sequences in the hemoglobin (Weatherall 1986; Labie and Elion 1996) and glucose-6-phosphate dehydrogenase (Vulliamy et al. 1992) genes. These are all genes that encode ''extrovert'' proteins that handle diverse foreign structures such as antigens and parasites, in contrast to the ''introvert'' proteins that handle conserved structures internal to the body (Mitchison 1997); hemoglobin and glucose-6-phosphate dehydrogenase have been included in this list as extrovert proteins, as the high level of polymorphism evident in their structural genes results largely from their interaction with malaria parasites. The genes that encode extrovert proteins often vary in their coding sequences, so as to provide a range of binding sites for these foreign structures, and vary also in their noncoding regions (for reference, see Mitchison 1997). Variation in noncoding gene segments is also evident to a lesser extent among genes encoding introvert proteins, particularly among cytokines (and their receptors and natural antagonists) (Daser et al. 1996). The functional consequences of this promoter sequence variation is not generally known, but its importance in determining expression level is becoming clearer for the class II MHC genes (Louis et al. 1994; Woolfrey and Nepom 1995), the ACE I gene (Villard et al. 1996), and cytokines genes (Messer et al. 1991; Pociot et al. 1992). For these MHC genes, valuable collections have been made of very distantly related sequences (Benoist and Mathis 1990), which enable comparisons to be made with the level of intraspecific variation. Thus, the basis for systematic study of this form of variation is now available, in a preliminary way at least.

One of the main functions of regulatory DNA sequences is to provide sequences recognized by transcription factors. These sites in the MHC class II promoter region are grouped in the S, X (X1 and X2), and Y boxes, although there are other regulatory sites farther upstream (Glimcher and Kara 1992; M. Janitz, L. Reiners-Schramm, and R. Lauster, in prep.) and within the 5′-untranslated region (Janitz et al. 1997b). This grouping leads one to ask two simple questions: To what extent are these box sequences conserved during evolution, and how is

[3]Corresponding author.
E-MAIL kepler@stat.ncsu.edu; FAX (919) 515-1909.

 8:124–134 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

variation within a species distributed in relation to the boxes? Over long evolutionary distances the box sequences are known to be conserved (Benoist and Mathis 1990; Sültmann et al. 1993), but the expectation for intraspecific variation is less clear. To the extent that the expression of these genes is under balancing selective pressure, one would expect to find variation at the boxes. It is important to add that the available information does not allow exact comparison to be made between different genes in this form of variation. That is because the sequence collections do not themselves contain information about gene frequencies in the natural population nor are the phenotypic effects of these genes known. Indeed, the role of natural selection cannot be addressed in laboratory mice or other domesticated species.

Class II MHC genes have been chosen for this study not only because their promoters are relatively well understood and their alleles available in quantity, but also because the proteins that they encode have been studied in depth (for thorough discussions, see Germain 1993; Hansen et al. 1993). Each class II molecule is composed of one α chain comprising an outer (membrane–distal) α_1 and an inner (membrane–proximal) α_2 domain, and a β chain likewise divided into β_1 and β_2 domains. The membrane–distal domains contain the residues that make contact with peptide and the T-cell receptor; the only known natural function of these molecules is to present peptide to T cells. These distal domains are encoded in the 5′ half of the structural gene of these type-1 membrane proteins. Class I MHC molecules have a fundamentally similar structure, although the two domains that make up the peptide-binding site are the α_1 and α_2 domains of a single α chain; the analog of the class II β_2 domain is a single-domain molecule that is encoded outside the MHC. It is the contact residues that are most variable in amino acid and base sequence. The crystal structure of both class I and class II MHC molecules has been determined (Madden 1995), thus enabling the contact residues to be defined in detail. For present purposes the important point is that the 5′ half of the class II coding genes are most variable, as is the 5′ one-third of the class I coding gene. This variability is concentrated at the peptide binding residues.

We are not aware of previous attempts to measure noncoding variation in this context, except by simply counting the positions where bases vary (Guardiola et al. 1996). In immunology, the Wu and Kabat (1970) approach is used frequently to measure protein sequence variability. We suggest, however, that both are inadequate measures of diversity. For the present analysis, therefore, we use the Shannon entropy (Shannon and Weaver 1949), a convenient and natural general measure of diversity (Patil and Taillie 1982).

RESULTS

Mus musculus MHC class II Promoter Sequences

The diversity index H_i (see Methods) was computed for each position in the alignment of *Mus musculus* *Aα*, *Aβ*, and *Eβ* MHC class II haplotypes. Most of the positions in all three data sets are invariant with $H_i = 0$ or are invariant with the exception of a single differing nucleotide and have $H_i = 0.27178$. As can be seen in Figure 1, positions with a larger diversity index are clustered together in two groups, with one loosely clustered group in the 3′ half of the promoter and one dense group centered over the X box. The S box is diverse in comparison to the flanking regions but is relatively invariant when compared to the X and Y boxes. The region of greatest diversity begins 5′ of the X box in the pyrimidine tract where transcription factors are known to bind, peaks over the X box, and extends into the region just 3′ of X.

The average diversity of the boxes, $\overline{H}_b$ (see Methods, below), was computed for each set of sequences by averaging H_i over all positions occurring in any one of the boxes. The average diversity outside the boxes, $\overline{H}_n$ was computed by averaging over the rest of the positions. The combined difference $\overline{H}_b^T - \overline{H}_n^T$, where $\overline{H}^T = \overline{H}^{A\alpha} + \overline{H}^{A\beta} + \overline{H}^{E\beta}$, is 0.0737. It was tested for statistical significance, as described in Methods, below with a resulting $P = 0.0135$. Using the Nei index of diversity, we find a combined difference value of 0.0798, with a corresponding $P = 0.028$. Thus, the variability in the *Mus* promoter proximal region is concentrated to a significant extent in the regions of transcription factor binding, namely the S, X, and Y boxes. The difference $\overline{H}_b - \overline{H}_n$ computed for each locus separately (Table 1), however, is not statistically significantly different from 0 at the traditional 0.05 level; under the null hypothesis, the probability of obtaining the differences seen for each locus independently were between $P = 0.143$ and $P = 0.158$. This is clearly dependent on the number of positions 5′ of the S box in the analysis. We included just five or six positions.

Interspecies Comparison of MHC Class II Promoter Sequences

Six α-chain sequences from *M. musculus* (*A* and *E*),

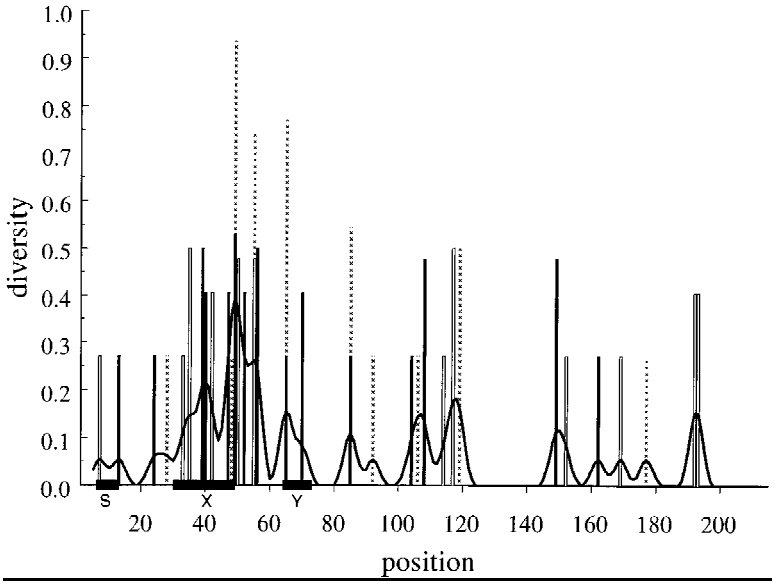


Figure 1 The diversity of the promoter regions measured by the Shannon entropy for the *Mus Eβ* (open bar), *Aα* (solid bar), *Aβ* (hatched bar) loci (shown stacked). Each locus is represented by eight haplotypes. The solid lines are moving averages computed using a Gaussian kernel 2 nucleotides wide. The width of this smoothing kernel is chosen to be the minimum length that produces smoothing sufficient for ease of visual inspection. Small variations in this width do not have an appreciable impact on the appearance of this plot. The locations of the promoter boxes are indicated by heavy lines on the abscissa. Base numbers start 6 nucleotides 5′ of the S box. Sequence lengths are as follows where the last 3 bases form the transcription start site ATG: *Eβ,* 215 bases; *Aα,* 174 bases; *Aβ,* 199 bases.

Homo sapiens (*DQ*, *DR*, and *DP*), and zebrafish and seven β chains from *M. musculus* (*A* and *E*), *H. sapiens* (*DQ*, *DR*, and *DP*), *Spalax ehrenbergi* (*DP*), and chicken (*B–L*) were aligned. H_i, $\overline{H}_b$, and $\overline{H}_n$ were computed as described in Methods, below. Our results, shown in Table 1, are consistent with the expectations and results of Benoist and Mathis (1990) and Sültmann et al. (1993). The boxes, defined by their conservation over long evolutionary distances, vary much less than the non-box regions. The difference $\overline{H}_b - \overline{H}_n$ is statistically highly significant ($P < 10^{-3}$). In the α chains, all but 2 of 15 invariant positions are located in the boxes. In the β chains, all but 8 of 20 invariant positions are located in the boxes and 5 of these 8 positions are immediately 5′ of X in the pyrimidine tract where the binding factors RF-X and NF-X are known to contact the DNA (Benoist and Mathis 1990).

In Figure 2, one can see that for the α and β chains there are distinct points of low diversity centered over the S, X, and Y boxes. The diversity within the boxes is approximately half that outside the boxes (Table 1). Figure 2 reveals that the distribution of diversity differs little between the α and β chains, especially in the 3′ half of the promoter. When examining the 5′ half, one sees that the diversity in the S box is much lower in the β chains. There are no invariant positions in the S box of the α chains, but the overall level of diversity is clearly lower in this region when compared to non-box regions. In the β chains, there is an additional point of low diversity 5′ of the X box over the pyrimidine tract. The overall diversity in the X box for the α and β chains is greater than that in the Y box. It is interesting to note that the point of low diversity centered over the X box in the α chains occurs at the 5′ end of the box in the X_1 portion of the box, whereas in the β chains, this point occurs at the 3′ end of the box in the region

Table 1. Results of the Promoter Sequence Analysis

		Total	$\overline{H}_b$	$\overline{H}_n$	*p*	Ψ	*p*
Mus	*A*α	0.03298	0.05682	0.02698	0.157	0.07377	0.0135
	*A*β	0.01679	0.03270	0.01307	0.0158	—	—
	Eβ	0.02137	0.04141	0.01712	0.143	—	—
Multiple	α	0.5234	0.3043	0.6424	$<10^{-3}$	—	—
	β	0.51447	0.32114	0.61834	$<10^{-3}$	—	—
Human	class II	0.05651	0.03584	0.06140	>0.2	—	—
	class I	0.02969	—	—	—	—	—

$\overline{H}$ is the average of the diversity index over each position in the region, where *b* denotes positions in one of the boxes and *n* denotes those not in one of the boxes. Ψ is the difference between $\overline{H}_b$ summed over all three *Mus* data sets and $\overline{H}_n$ summed over all three *Mus* data sets. (—) The computation was not relevant in the context of the data set and therefore not done.

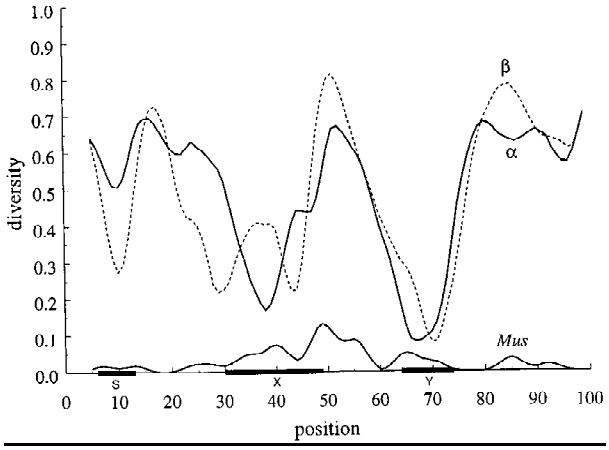


Figure 2 (*Top* two curves) Diversity of promoter sequences in distantly related class II genes over 6 sequences (104 bases in length) from 3 species (α chain) and 7 sequences (103 bases in length) from 4 species (β chain) (see Methods). The diversity is represented by a moving average computed with a Gaussian kernel 2 nucleotides wide. (*Bottom* curve) Moving average over *Mus* promoters (repeated from curve in Fig. 1).

known as X_2. This may be related to the fact that the X_2 binding factors are thought to be members of the Fos/Jun/CREB family, whose other members, TRE and CRE, are known to bind to half-sites (Benoist and Mathis 1990).

When comparing the interspecies and *Mus* intraspecies distributions (Fig. 2), it is evident that the two distributions are mirror images of each other with the interesting exceptions that there is a corresponding peak just 3′ of the X box in all three sets of sequences. The distribution inside the X box of the *Mus* sequences follows the distribution seen in the β chains, namely diversity in the X_1 region and similarity in the X_2 region, whereas the distribution seen in the X box of the α chains mirrors the two.

Human MHC Promoter Sequences

As can be seen in Figure 3, variability in the human MHC class II promoter region is distributed more evenly across the promoter and not concentrated near the X- and Y-box regions. The non-box region may have a higher average diversity value, although this difference, as seen here, is not significant (see Table 1 for exact values). As in the mouse, most of the variable positions are invariant, with the exception of just one or two sequences, and have correspondingly low H_i values, but there are many more variable positions.

No conserved boxes have been located in the MHC class I promoter sequences. But when looking at overall diversity and its general distribution, one sees (Fig. 4) the opposite picture as that seen in the case of human MHC class II promoters. Here there are very few variable positions, but most of those that vary have H_i values near 0.5. This tends to result from the occurrence of two different nucleotides at one position with approximately equal frequencies. The overall diversity is almost half that seen in the human class II promoter sequences (Table 1).

MHC Class II Coding Sequences

An examination of the human class II coding sequences reveals diversity throughout both of the extracellular domains, but at a higher level in the β_1 domain (Fig. 5). As expected, the level of diversity is highest in the areas of peptide binding, especially in amino acid residues 9, 11, 13, 70, and 71. Residues 13, 70, and 71 all correspond to pocket 4, whereas residue 11 is the only β-chain residue in pocket 6 (Travers 1997). In the mouse, the diversity is also concentrated in exon 1, where the peptide binding residues are located (Fig. 6). The overall level of di-

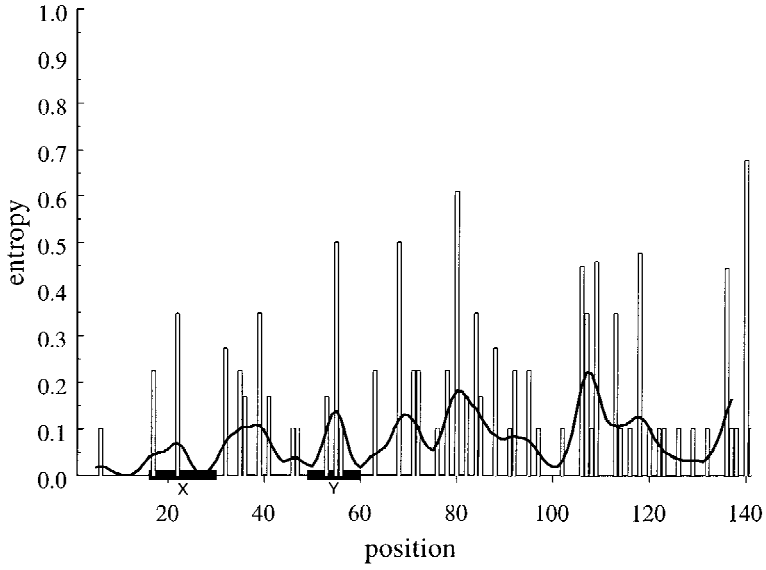


Figure 3 The diversity of the promoter regions for the human DRβ locus (open bars), represented by 32 haplotypes. The solid line indicates the moving average. (Details as in Fig. 1.) Base numbers start 15 nucleotides 5′ of the X box; 277 bases were included in the analysis.

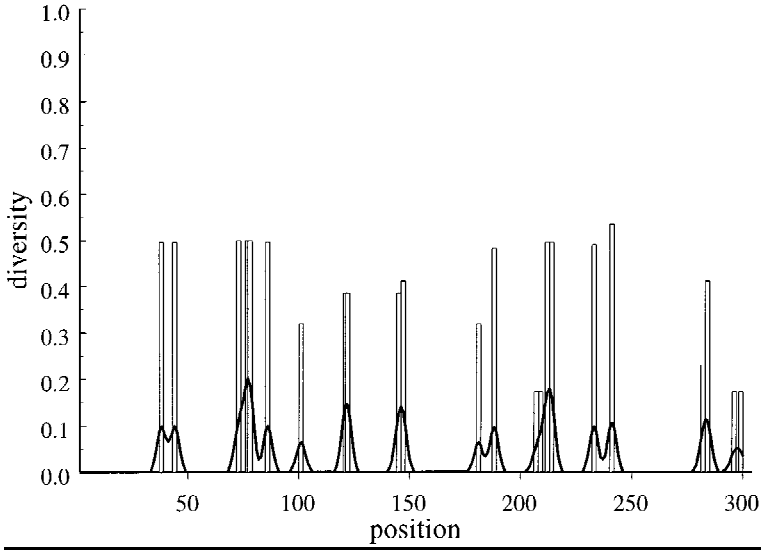


Figure 4 Diversity over MHC HLA-B locus (open bars) (31 haplotypes). The solid line indicates the moving average. Sequences are 303 bases long. (Details as in Fig. 1.)

versity is highest for the Aβ sequences. This is due not only to the presence of more variable positions, but also to the occurrence of more positions with a large index value. The overall level of diversity in the Eβ sequence set is much lower than either of the other two. Caution must be applied when comparing the human and mouse coding sequences. The entropy has been estimated using very different sample sizes, and it can be shown that the sample entropy is a downward-biased estimator for the population entropy, with bias decreasing for decreasing sample size.

DISCUSSION

Measurement of Diversity

The concept of diversity as a property of a group of objects is of interest in many different disciplines including genetics, linguistics, business, economics, and ecology (Patil and Taillie 1982). This broad interest in quantifying diversity and comparing diversity has led to the development of several different diversity indices along with an analytic framework for examining the models behind the various indices and their properties (Patil and Taillie 1982). We have chosen to use the Shannon entropy as a measure of diversity in this analysis. The use of the Shannon entropy as a measure of diversity may be novel for many in our intended readership; therefore, in addition to the brief comments in the discussion to follow, please see the Appendix, which presents a slightly broader background for the mathematical issues involved. Those wishing a more extensive discussion of the quantitation of diversity can find such treatments in Hill (1973), Kempton (1979), Patil and Taillie (1982), and Pielou (1977).

The Shannon entropy was originally devised in the development of information theory (Shannon and Weaver 1949) and can be interpreted as the information gain expected in the performance of a single measurement from the population under discussion. This aspect of entropy has been exploited within molecular biology (Román-Roldán et al. 1996) and, in particular, with regard to DNA regulatory binding sites and the question of how much ''information'' is required to identify such regions (Schneider et al. 1986). These formal aspects of the entropy are of secondary importance for the present analysis. Entropy, and even diversity per se, are not the quantities of ultimate interest to us. The quantity that most unambiguously measures the effect we seek to document, however, will depend on many biological features: the specifics of DNA-protein binding, the mechanisms of evolutionary diversification and selection, the details of differential gene regulation, and the extent of the advantage gained through the increased flexibility of the immune response. In short, this ideal measurement is not yet available. The entropy is very likely to be highly correlated with this unknown index over the

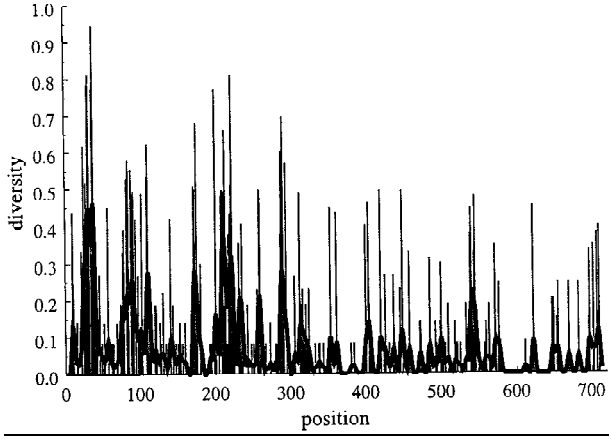


Figure 5 Diversity over human *DRβ* coding regions. Sequences are 718 bases long. Average diversity over all positions at this locus is 0.06825. (Other details as in Fig. 1.)

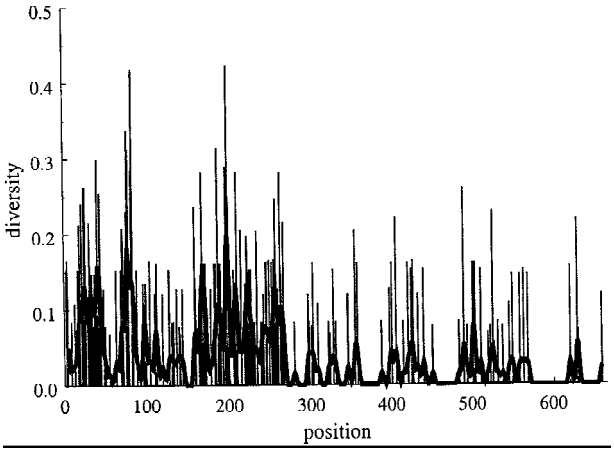


Figure 6 Diversity over *Mus Eβ, Aα,* and *Aβ* coding regions. Sequences are 667 bases long. Bars are averages over the three loci. (Other details as in Fig. 1.) Note that the ordinate is scaled to show a maximum value of 0.5 rather than 1 as in Figs. 1–5. The average diversity over all positions for each locus is as follows: *Eβ,* 0.02995; *Aα,* 0.04143; *Aβ,* 0.05141.

range of its observed variation and will therefore serve as an adequate indicator. More detailed models of the underlying mechanisms will eventually provide more powerful statistical tests.

To the best of our knowledge, investigators addressing the question of promoter sequence variability in the past have used the frequency of variable bases per region as a measure of region variability. This is clearly inadequate in that all information included in the knowledge of the number of different nucleotides appearing at a position is lost as well as all information contained in knowledge of their relative abundances. The Wu and Kabat (1970) approach to measuring sequence variation has been applied extensively in immunology, namely to amino acid sequences in antibodies (Kabat et al. 1991) and MHC molecules (Parham et al. 1989). It is thus worth mentioning that we also explicitly prefer the Shannon entropy over the Wu–Kabat index. The Wu–Kabat index played a valuable role in early qualitative discussions of diversity, but in contrast to the Shannon entropy, has no obvious natural interpretation and is not well suited to quantitation. [For a more detailed discussion of the mathematical deficiencies of the Wu–Kabat index, and a comparison of its properties with those of indices based on the Shannon entropy, see Shenkin et al. (1991).] The nucleotide diversity index of Nei (1987) is also widely used. We have repeated our primary analyses using the Nei index in place of the Shannon index for the sake of historical continuity; the findings are consistent between the two indices. A brief discussion of the Nei and Shannon indices is included in the appendix.

Intra- and Interspecies Variation

The S, X, and Y boxes of the MHC class II promoters were initially identified by the level of their conservation across species. Our analyses are consistent with these observations: in interspecies comparisons, the noncoding DNA within the S, X, and Y boxes show significantly lower diversity than the noncoding DNA outside these boxes (Fig. 2). The contrast between intra- and interspecific variation is shown in Figure 2, where the *Mus* average diversity curve is plotted together with that for interspecific variation. [*Mus castaneus* does not figure as distant, as its promoter fits within the same group as the laboratory strains examined here (Janitz et al. 1997a; M. Janitz, L. Reiners-Schramm, and R. Lauster, in prep.), as was shown by tree analysis (result not shown).] Within *Mus* MHC class II promoters, we find *greater* diversity within the S, X, and Y boxes than outside them (Fig. 1 and accompanying statistical analysis). We take this as evidence for diversifying selection within the boxes.

The human class II promoters examined did not show the same pattern of localized variation but rather were consistently variable over the entire promoter region. It is difficult to arrive at any specific conclusions about this latter fact. Although the boxes are not overly diverse, they are also not overly conserved. The extent of variation within the boxes in this series has excited comment previously (Louis et al. 1993). We suggest two alternative perspectives. On the one hand, the level of variation in the promoter box regions may simply be tolerated rather than actively selected for. On the other hand, it may be that diversity in the boxes actually was selected for in the past but has reached its maximum tolerable level, whereas the non-box regions have continued to accumulate changes.

Selective Pressure

MHC promoter polymorphism is likely to be maintained by balancing selection in favor of heterozygotes. The presumed advantage of heterozygosity lies in the flexibility that it confers on the immune response: The protein under the control of the promoter in one allele is better able to mediate resistance to one type of pathogen, whereas the other allele is better suited to defense against another pathogen. The means by which this is brought

about is likely to be in selective expression in particular cell types (Daser et al. 1996; Mitchison 1997). It is thus interesting that the one *Mus* MHC class II locus with an invariant promoter, *Eα* is the only locus with a variant upstream enhancer (M. Janitz, L. Reiners-Schramm, and R. Lauster, in prep.); this enhancer variation has also been shown to have functional consequences (M. Janitz, L. Reiners-Schramm, and R. Lauster, in prep.). Whether the lack of localized variation in the human promoter sequences implies a difference in the selective pressures on the two species cannot be inferred from this analysis. That is, however, an important question, and a similiar study including additional human MHC class II loci as well as intraspecific comparisons with a variety of other species is needed. There are other aspects of variation in human MHC class II promoters, including allele frequencies and the consequences of haplotype selection, that are important but necessitate extensive review, beyond the scope of the present work.

Further information about the functional significance of this promoter polymorphism is likely to be acquired through expression studies on naturally occurring variants (for review, see Guardiola et al. 1996; Müller and Mitchison 1997). More critical testing is likely to come from site-specific mutagenesis and promoter/exon reshuffling in cell lines and eventually in mice.

The relative conservation of MHC class I promoters in comparison with those of class II is not surprising. Class II MHC proteins, expressed primarily on specialized antigen-presenting cells, mediate a number of subtly different immunoregulatory cell–cell interactions that require precise control of expression level (Guardiola et al. 1996; Constant and Bottomly 1997) as well as tissue specificity of expression. In contrast, class I MHC proteins are expressed by nearly all cells; their expression is upregulated drastically at sites of inflammation, thus enabling virus-infected cells to function simply as antigen-presenting machines (Germain 1993).

In their MHC class II promoters, *Mus* and *Spalax* show very little similarity outside the conserved S, X, and Y boxes, in contrast to the close similarity of laboratory mice and *M. castaneus*. *Mus* and *Spalax* have separate ancestry tracing back into the Oligocene, soon after the origin of the rodents, so it would be of interest to compare species related more closely to the mouse, such as *Rattus*, *Mastacomys* (both of which have laboratory strains), and *Apodemus*, all of which diverged post-Miocene (Thenius 1980). The EMBL database does not contain promoter sequences for these species.

Within the promoter boxes, the diversity is comparable to that in the coding sequences. Coding sequences within the MHC have been shown to be subject to diversifying selection (Hughes and Nei 1988). We suggest that the advantage conferred by this exon diversity is further enhanced by diversity in the associated promoters (Mitchison 1997). Once functional diversity of the protein has been established, the potential exists for differential expression of the proteins to effect immunoregulation. Variability in the promoters accomplishes this and provides additional flexibility to the immune response. Thus, we expect that the extent of diversity in the exons and promoters may be correlated.

METHODS

To address the question of how diversity is distributed within regulatory sequences, we have made several comparisons focusing on diversity of MHC class II promoter sequences. The first comparison examines the level of diversity seen within the conserved boxes versus that seen outside of these boxes. For the intraspecies comparisons, 8 haplotypes from laboratory strains of *M. musculus* were used for the *Aα*, *Aβ*, and *Eβ* loci, as well as 32 haplotypes for the human *DRβ* genes. The alignment and sequences for the *Mus b*, *d*, *k*, and *q* haplotypes (all three loci) were taken from Janitz et al. (1997a). Haplotypes *z*, *p*, *j1*, and *j2* (*Eβ* locus) and haplotypes *z*, *p1*, *p2*, and *j* (both *A* loci) can be found in GenBank under accession numbers Y13072–Y13083. They were aligned to the original alignment using the program GeneWorks 2.1 (Oxford Molecular Group). The 32 human sequences and their alignment were taken from Louis et al. (1993). The *Mus Eα* locus was not included in the analysis because it is known to be invariant. For the interspecies comparison, the alignment and sequences for six α chains and seven β chains were taken from Benoist and Mathis (1990). An additional sequence, a zebrafish α chain sequence, was taken from Sültmann et al. (1993) and was aligned to the Benoist and Mathis sequence using GeneWorks 2.1 (Oxford Molecular Group). For all species, the sixth base 5′ of S was used as the 5′ boundary; the 3′ boundary corresponds to that used in Benoist and Mathis (1990). These MHC class II promoter sequence comparisons are supplemented with a comparison between human MHC class I and class II promoter sequences, as well as an examination of the diversity seen in human and *Mus* MHC class II coding sequences. Thirty-one haplotypes of the HLA-B locus were used for the class I analysis. All sequences and their alignment are taken from Yao et al. (1995). For each set of sequences, the 5′ and 3′ boundaries, as well as the location of the boxes in promoters, is consistent with that in the cited references unless otherwise noted.

To examine the class II coding sequence diversity, 41 alleles of the *DRβ* genes in humans were used along with 9, 10, and 12 haplotypes for the *Mus Aα*, *Aβ*, and *Eβ* loci, respectively. For the human *DRβ* sequence set, the alignment and sequences for the following alleles were taken from Figueroa et al. (1991): *DRB1*1603*, *DRB1*0411*, *DRB1*0801*, *DRB1*08022*, *DRB1*08031*, *DRB1*09011*, *DRB3*0101*, *DRB3*0301*, and *DRB4*0103*. The allele *DRB1*1302* (accession no. U83584) was taken from GenBank; all of the *DRB1*,

DRB3, *DRB4*, and *DRB5* alleles available in the Graphical Interface to MHC Sequence Database (Histo) were also included (Travers 1997). Sequences taken from the Histo database are accessed by their allele designation. These sequences were aligned to the alignment of Figueroa et al. (1991) using the program GeneWorks 2.1 (Oxford Molecular Group). For the *Mus* sequence sets, all sequences and their alignment were taken from the Histo database. The following haplotypes were used: *I-A K-A′ CL, I-A B-A′ CL, I-A F-A′ CL, I-A U-A′ CL, I-A D-A′ CL, I-A S-A′ CL, I-A R-A′ CL, I-A Q-A′ CL, NON Aa′, I-A K-B′ CL, I-A K-B′ ′CL, I-A U-B′ CL, I-A b NOD′ CL, I-A F-B′ CL, I-A S-B′ ′CL, H-2 Ab/p′ CL, I-A D-B′ CL, I-A B-B′ CL, I-A Q-B′ CL, I-E (I-E) D-B EB24-1′ CL, I-E U-B′ CL, I-E B EBB24′ CL, I-E S-B′ CL, I-E K-B′ CL, MOUSE H-2 (H-2) Z′ CL, H-2 Eb W17′ CL, NON Eb′ CL, B10.A(3R)-E-β′ CL*, and *NOD Eb′ CL* (Travers 1997). Only the exons for the β1 and β2 domains were considered.

In each case, the Shannon entropy (Patil and Taillie 1982) was computed for each position in the alignment. For purposes of this analysis, gaps are ignored, that is, at each position where gaps occur, those sequences with gaps are omitted. The Shannon entropy is defined by

$$H_i = -\sum_j p_{i,j} \log_4 (p_{i,j}) \quad (1)$$

(Shannon and Weaver 1949). The position in the alignment is indexed by *i*, whereas *j* indexes the four nucleotides. $p_{i,j}$ is the frequency of nucleotide *j* at position *i*. The average H_i, $\overline{H}$, was computed over position within the regions of interest and then used to compare the diversity of the regions. Positions occurring in any one of the boxes were grouped into one region with average diversity $\overline{H}_b$ and the remaining positions were grouped into a second region with average diversity $\overline{H}_n$. To detect statistically significant differences in the diversity of the two regions, a *t* statistic was computed for the difference $\overline{H}_b - \overline{H}_n$. Let this value of *t* be denoted t_0. H_i is not normally distributed, hence a comparison of t_0 to Student's *t* is not appropriate. Therefore, the probability of obtaining *t* such that $|t| > t_0$ under the null hypothesis (that diversity between the two regions does not differ) was estimated by randomizing the positions in the alignment 1000 times, calculating *t* for each permutation and then counting the frequency of permutations yielding *t* such that $|t| > t_0$.

Additionally, we wanted to estimate the probability under the null hypothesis of observing a combined difference as large as or larger than that observed in our three *Mus* data sets. We summed $\overline{H}_n$ over all three data sets, $\overline{H}_n^{A\alpha} + \overline{H}_n^{A\beta} + \overline{H}_n^{E\beta}$, and subtracted that from the sum of $\overline{H}_b$ over all three data sets. Call this difference Ψ_0. We then randomized the positions in the alignments of each data set independently two thousand times and computed the difference Ψ for each permutation. The probability of observing a difference greater than or equal to that present in our data sets was then estimated by the frequency of Ψs such that $|\Psi| > \Psi_0$.

This analysis was repeated using the Nei index of nucleotide diversity (Nei 1987) in place of the Shannon entropy. A brief discussion of the measurement of nucleotide diversity is included in the Appendix.

We developed Fortran 90 programs (available on request, from lgcowell@unity.ncsu.edu) using the Microsoft Developer Studio with Fortran Powerstation 4.0 (Copyright 1994–1995 Microsoft Corporation) to compute sitewise diversity indices and perform permutations. Remaining computations were done using the commercially available software package Splus version 3.2 (Statsci, a division of Mathsoft).

ACKNOWLEDGMENTS

This work was supported by a Fulbright grant (to L.G.C.), by National Science Foundation award MCB–9357637 (to T.B.K.), and by the Deutsche Forschungsgemeinschaft and the Senate Administration for Research and Education of the City of Berlin. We are grateful to T. Shiroishi and Sonoko Habu for genomic DNA from the MSM strain and P.-A. Cazenave for genomic DNA from *M. castaneus*.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aczél, J. and Z. Daroczy. 1975. *On measures of information and their characterizations.* Academic Press, New York, NY.

Benoist, C. and D. Mathis. 1990. Regulation of major histocompatibility complex class II genes: X and Y and other letters of the alphabet. *Annu. Rev. Immunol.* 8: 681–715.

Bowman, K.O. 1971. Comments on the distribution of indices of diversity. In *Statistical ecology* (ed. G.P. Patil, E.C. Pielou, and W.E. Walters), vol. 3, pp. 315–366. Penn State University Press, University Park, PA.

Constant, S.L. and K. Bottomly. 1997. Induction of Th1 and Th2 $CD4^+$ T cell responses: The alternative approaches. *Annu. Rev. Immunol.* 15: 297–322.

Daser, A., H. Mitchison, A. Mitchison, and B. Müller. 1996. Non-classical-MHC genetics of immunological disease in man and mouse. The key role of proinflammatory cytokine genes. *Cytokine* 8: 593–597.

Figueroa, F., C. O'hUigin, H. Inoki, and J. Klein. 1991. Primate *DRB6* pseudogenes: Clue to the evolutionary origin of the *HLA–DR2* haplotype. *Immunogenetics* 34: 324–337.

Germain, R.N. 1993. Antigen processing and presentation. In *Fundamental immunology* (ed. W.E. Paul), pp. 629–676. Raven, New York, NY.

Glimcher, L.H. and C.J. Kara. 1992. Sequences and factors: A guide to MHC class-II transcription. *Annu. Rev. Immunol.* 10: 13–49.

Guardiola, J., A. Maffei, R. Lauster, N.A. Mitchison, R.S. Accolla, and S. Sartoris. 1996. Functional significance of polymorphism among MHC class II promoters. *Tissue Antigens* 48: 615–625.

Hansen, T.H., B.M. Carreno, and D.H. Sachs. 1993. The major histocompatibility complex. In *Fundamental immunology* (ed. W.E. Paul), pp. 577–628. Raven, New York, NY.

Hill, M.O. 1973. Diversity and evenness: A unifying notation and its consequences. *Ecology* 54: 427–432.

Hughes, A.L. and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335: 167–170.

Janitz, M., N.A. Mitchison, L. Reiners-Schramm, and R. Lauster. 1997a. Polymorphic MHC class II promoters exhibit distinct patterns in various antigen presenting cell lines. *Tissue Antigens* 49: 99–106.

Janitz, M., L. Reiners-Schramm, and R. Lauster. 1997b. Enhancer activity in the 5′ untranslated region of the *H2-Eb* gene. *Immunogenetics* 45: 432–435.

Kabat, E.A., T.T. Wu, H.M. Perry, K.S. Gottesman, and C. Foeller. 1991. *Sequences of proteins of immunological interest.* U.S. Department of Health and Human Services, Washington, DC.

Kempton, R.A. 1979. The structure of species abundance and measurement of diversity. *Biometrics* 35: 307–321.

Khinchin, A.I. 1957. *Mathematical foundations of information theory*. Dover Publications, New York, NY.

Labie, D. and J. Elion. 1996. Sequence polymorphisms of potential functional relevance in the β-globin gene locus. *Hemoglobin* 20: 85–101.

Louis, P., J.F. Eliaou, C.S. Kerlan, V. Pinet, R. Vincent, and J. Clot. 1993. Polymorphism in the regulatory region of HLA-DRB genes correlating with haplotype evolution. *Immunogenetics* 38: 21–26.

Louis, P., V. Pinet, P. Cavadore, S. Kerlan Candon, J. Clot, J.F. Eliaou, P. Louis, and R. Vincent. 1994. Differential expression of HLA-DRB genes according to the polymorphism of their regulatory region. *C.R. Acad. Sci. Ser. III Sci. Vie* 317: 161–166.

Madden, D.R. 1995. The three-dimensional structure of peptide–MHC complexes. *Annu. Rev. Immunol.* 13: 587–622.

Messer, G., U. Spengler, M.C. Jung, G. Honold, K. Blomer, G.R. Paper, G. Riethmuller, and E.H. Weiss. 1991. Polymorphic structure of the tumor necrosis factor (TNF) locus: An *Nco*I polymorphism in the first intron of the human TNF-β gene correlates with a variant amino acid in position 26 and a reduced level of TNF-β production. *J. Exp. Med.* 173: 209–219.

Mitchison, N.A. 1997. Partitioning of genetic variation between regulatory and coding gene segments: The predominance of software variation. *Immunogenetics* 46: 46–52.

Müller, B. and N.A. Mitchison. 1997. The importance of the back-signal from T cells into antigen-presenting cells in determining susceptibility to parasites. *Philos. Trans. R. Soc. Lond. B* 352: 1327–1330.

Nei, M. 1987. In *Molecular evolutionary genetics*, p. 256. Columbia University Press, New York, NY.

Parham, P., D.A. Lawlor, R.D. Salter, C.E. Lomen, P.J. Bjorkman, and P.D. Ennis. 1989. HLA-A, -B, -C: Patterns of polymorphism in peptide binding proteins. In *Immunobiology of HLA* (ed. B. Dupont et al.), p. 10. Springer-Verlag, New York, NY.

Patil, G.P. and C. Taillie. 1982. Diversity as a concept and its measurement. *J. Am. Statist. Assoc.* 77: 548–563.

Pielou, E.C. 1977. Ecological diversity and its measurement. In *Mathematical ecology*, p. 291. John Wiley and Sons, New York, NY.

Pociot, F., J. Molvig, L. Wogensen, H. Worsaae, and J. Nerup. 1992. A TaqI polymorphism in the human interleukin–1 β (IL–1 β) gene correlates with IL–1 β secretion in vitro. *Eur. J. Clin. Invest.* 22: 396–402.

Román-Roldán, R., P. Bernaola–Galván, and J.L. Oliver. 1996. Application of information theory to DNA sequence analysis: A review. *Pattern Recogn.* 29: 1187–1194.

Schneider, T.D., G.D. Stormo, L. Gold, and A. Ehrenfeucht. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188: 415–431.

Shannon, C.E. and W. Weaver. 1949. *The mathematical theory of communication*. University of Illinois Press, Urbana, IL.

Shenkin, P.S., B. Erman, and L.D. Mastrandrea. 1991. Information-theoretical Entropy as a measure of sequence variability. *Proteins Struct. Funct. Genet.* 11: 297–313.

Sültmann, H., W.E. Mayer, F. Figueroa, C. O'hUigin, and J. Klein. 1993. Zebrafish MHC class II α chain encoding genes: Polymorphism, expression, and function. *Immunogenetics* 38: 408–420.

Thenius, E. 1980. *Grundzüge der Faunen- und Verbreitungsgesichte der Säugetiere*, 2nd ed. Fischer Verlag, Stuttgart, Germany.

Travers, P. 1997. Graphical interface to MHC sequence database. http://histo.cryst.bbk.ac.uk.

Villard, E., L. Tiret, S. Visvikis, R. Rakotovao, F. Cambien, and F. Soubrier. 1996. Identification of new polymorphisms of the angiotensin I-convertin enzyme (ACE) gene, and study of their relationship to plasma ACE levels by two-QTL segregation-linkage analysis. *Am. J. Hum. Genet.* 58: 1268–1278.

Vulliamy, T., P. Mason, and L. Luzzatto. 1992. The molecular basis of glucose-6-phosphate dehydrogenase deficiency. *Trends Genet.* 8: 138–143.

Weatherall, D.J. 1986. The regulation of the differential expression of the human globin genes during development. *J. Cell. Sci. Suppl.* 4: 319–336.

Woolfrey, A.E. and G.T. Nepom. 1995. Differential transcription elements direct expression of HLA-DQ genes. *Clin. Immunol. Immunopathol.* 74: 119–126.

Wu, T.T. and E.A. Kabat. 1970. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* 132: 211–250.

Yao, Z., A. Volgger, S. Scholz, and E.D. Albert. 1995. Sequence polymorphism in the HLA-B promoter region. *Immunogenetics* 41: 343–353.

Received July 28, 1997; accepted in revised form January 5, 1998.

APPENDIX: QUANTITATION OF DIVERSITY

Necessary criteria for a measure of diversity are that (1) for a given number of types of objects, the index takes its greatest value when all types are present in equal proportions, which means that maximum diversity is assigned to the completely even group; (2) given two completely even groups, the group made up of more types should have the larger diversity index; and (3) if two classification systems are used, A and B, the diversity of the doubly classified community, $H(AB)$, should be equal to the sum of the diversity measures under each of the classification systems. That is, $H(AB) = H(A) + H_A(B)$, where $H_A(B)$ is the expectation of $H(B)$ under the classification A. If the classifications A and B are independent, criterion 3 becomes additivity of the indices: $H(AB) = H(A) + H(B)$. Khinchin (1957) has shown that the only function meeting all three of these criteria is the Shannon entropy. In its most general form, the index is $H(p_1, p_2,...p_s) = -C\Sigma_j p_j \log(p_j)$, where C is any positive constant, and in the context of diversity, s is the total number of object types in the group and p_j is the frequency of type j in the group. C is generally taken as 1 and the base for the logarithm as 2, e, or 10. For a detailed characterization of the Shannon entropy, see Aczel and Daroczy (1975).

Following the treatment of Patil and Taillie (1982), we consider diversity to be a measure of the average rarity of the types of objects making up the group. Let p_j be the proportion of the community made up by type j and $R(p_j)$ be the rarity of type j. Then the average rarity can be expressed as $\Delta = \Sigma_j p_j R(p_j)$. Choosing the Shannon index H with base 4 to measure average rarity, we have $R(p_j) = -\log_4(p_j)$, and our rarity function has the following desirable properties: (1) The rarity of type j decreases as j's proportion in the group increases, that is, R is a decreasing function defined on the interval (0,1]; (2) $R(1) = 0$; thus, a group consisting of just a single type has diversity index 0. (3) The diversity increases when the share of the group is distributed more evenly among the types, with criterion 1 above being the most even case. The maximum value is $\Delta_{max} = -\log_4(N)$, where N is the number of types (Patil and Taillie 1982). Note that from property 1, it follows that introducing a new type increases the diversity measure, and properties 1 and 2 imply that Δ is non-negative (Patil and Taillie 1982).

For our purposes, a group is clearly a position in the alignment, and each of the 4 nucleotides is a type. When considering the value of the index at position i, H_i, the value reflects not only the number of different nucleotides appearing at that position, but it also captures the evenness, or relative abundances, of the nucleotides. The maximum value is assigned to the case with all 4 nucleotides appearing, each with $p_{i,j} = 1/4$, and is $H_i = -\log_4(4) = 1$. Given two completely even positions, e.g., one with 2 bases, both occurring with frequency $p_{i,j} = 1/2$, and one with 3 bases and $p_{i,j} = 1/3$, the index values are $H_i = -\log_4(1/2) = 0.50$ and $H_i = -\log_4(1/3) = 0.79$, respectively, and so the position with more nucleotide types has the larger H_i.

All of the characteristics discussed above are important here. Criterion 3, however, is especially useful for our problem and is thus worth discussing in detail. This criterion says that given two random variables X and Y that are jointly distributed with $p(x,y)$, their joint index of diversity can be defined as $H_{x,y} = \Sigma_{x,y} p(x,y) \log[p(x,y)]$. When X and Y are independent, $p(x,y) = p(x)p(y)$, and $H_{x,y} = \Sigma_x p(x) \log[p(x)] + \Sigma_y p(y) \log[p(y)]$. This feature allows us to compute the arithmetic average of the diversity index over a region by considering the identity of a nucleotide and its location in the alignment simultaneously. As long as nucleotide substitutions are introduced independently, the measure of diversity that we are using accurately reflects the true diversity of the promoter regions themselves. If the substitutions are not independent, a more sophisticated measure of diversity that accounts for correlations would be required. There are no indications within this data set that such considerations are necessary.

The statistic we use to estimate the Shannon index is the maximum likelihood estimator, given by replacing the population proportion p_i of the ith nucleotide at a given position, by its empirical frequency, n_i/n in the above formula. This estimator has a negative bias bounded below by $-3/2n$ (finite samples tend to underestimate the true entropy) (Bowman 1971) but has the usual advantages of a maximum likelihood estimator. Although one needs to be aware of this bias when comparing data

sets of greatly differing size, the comparisons made in this paper are between samples with equal size or for which differences in bias are negligible.

Another index of diversity used for DNA sequences is that attributable to Nei (1987). This index measures the average pairwise number of nucleotide differences per nucleotide position in a collection of sequences. The Nei index is equivalent to another index familiar from statistical ecology, the Simpson index, applied site-wise and then averaged over sites, as can be shown in a straightforward manner. A simulation-based comparison of several diversity indices has been conducted and published (Bowman 1971). We have chosen to use the Shannon index because it meets criteria 1–3 above (the Nei/Simpson index does not meet criterion 3) and because it provides a useful connection to other information-based methods for the analysis of DNA sequences. This unity of perspective may have valuable consequences for future research.