Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents

Peter D. Keightley* and Daniel J. Gaffney

Ashworth Laboratories, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom

Edited by James F. Crow, University of Wisconsin, Madison, WI, and approved September 12, 2003 (received for review May 29, 2003)

Selection against deleterious mutations imposes a mutation load on populations because individuals die or fail to reproduce. In vertebrates, estimates of genomic rates of deleterious mutations in protein-coding genes imply the existence of a substantial mutation load, but many functionally important regions of the genome are thought to reside in noncoding DNA, and the contribution of noncoding DNA to the mutation load has been unresolved. Here, we infer the frequency of deleterious mutations in noncoding DNA of rodents by comparing rates of substitution at noncoding nucleotides with rates of substitution at the fastest evolving intronic sites of adjacent genes sampled from the whole genome sequences of mouse and rat. We show that the major elements of selectively constrained noncoding DNA are within 2,500 bp upstream and downstream of coding sequences and in first introns. Our estimate of the genomic deleterious point mutation rate for noncoding DNA (0.22 per diploid per generation) is similar to that for coding DNA. Mammalian populations therefore experience a substantial genetic load associated with selection against deleterious mutations in noncoding DNA. Deleterious mutations in noncoding DNA have predominantly quantitative effects and could be an important source of the burden of complex genetic disease variation in human populations.

selective constraints | intron | intergenic DNA

Selection against deleterious mutations leads to a mutation load at the population level, because individuals die prematurely or have reduced fertility (1, 2). The mutation load can be defined as the proportion of individuals that are selectively eliminated (individuals that undergo "genetic death"; ref. 3) and depends critically on the genomic deleterious mutation rate, U. For example, under a multiplicative model the load is $1 - e^{-U}$ (where U is the mutation rate per diploid; ref. 4). The mutation load also depends on the manner in which mutations interact with one another between and within loci (4), and on population structure and system of mating (5), and can be reduced, for example, if mutations interact synergistically (4). The genomewide rate for mutations in coding DNA has been estimated on the basis of the fraction of conserved nucleotides at amino acid sites of protein-coding genes (6–8). There is a strong, positive correlation between generation time of a species and U (8), and U in long-lived taxa such as hominids is likely to exceed one event per generation (6, 7). However, the contribution of mutations in noncoding DNA to the genomewide deleterious mutation rate is an unresolved issue, because it has been difficult to relate function with DNA sequence, and, until recently, relevant data have not been available.

Protein-coding gene sequences comprise only a very small proportion of the total genomic content in mammals, most other vertebrates, many invertebrates, and most plants (9). For example, protein-coding sequences are thought to account for only ≈1.5% of the genomes of both humans and mice (10, 11). As much as 45% of the euchromatic portion of mammalian genomes consists of the remnants of transposable element insertions (10) that are only occasionally coopted for use by the host organism. Much of the remainder of the genome consists of unique intergenic and intronic DNA sequences, and motifs that are critical for regulating gene expression reside in these regions. Quantification of the degree of between-species evolutionary conservation is one way of searching for such regulatory regions (12). Over evolutionary time scales, directional selection is expected to drive the efficiency of a functional stretch of the genome toward an adaptive optimum, and most non-neutral mutations within it are expected to be deleterious. The between-species evolutionary divergence of functionally important regions is therefore expected to be lower than the divergence of neutral segments having similar mutation rates; those mutations in functional regions with selection coefficients higher than the reciprocal of the effective population size almost never become fixed between species (13).

A general approach to identify functionally important regions in the genome and to quantify the fraction of deleterious mutations is to search for segments of the genome having lower between-species levels of divergence than the average for the genome or than a linked putatively neutral sequence (14). Previous attempts to quantify the fraction of conserved nucleotides have relied on searching for blocks of DNA sequences that are conserved between distantly related taxa (15–18). However, there are at least two difficulties with this approach. First, estimation of noncoding DNA sequence alignment by heuristic methods can be biased if the true pattern of insertion/deletion (indel) events is unknown (19), and second, variability across the genome in the mutation rate can generate variation in conservation that is unrelated to functional constraint (12).

Here, we attempt to quantify the functional constraints on noncoding DNA in rodents by using the recently released genome sequences of mouse and rat by comparing rates of substitution in segments of noncoding DNA with rates of substitution at the fastest evolving intronic (FEI) sites of adjacent genes. We determined empirically that the intronic sites showing the fastest rates of evolution are those nucleotides not close to exon boundaries (i.e., not close to intron splice control regions) and outside first introns. We confine our analysis of constraints to those sites that are unlikely to have been ancestrally part of a CG dinucleotide, because such sites are close to saturation between mouse and rat. The whole genome sequence of the mouse has recently been published (11), and the whole genome sequence of the rat is publicly available on GenBank at seven to eight times coverage. Mouse and rat are sufficiently closely related that it is possible to be confident in the orthology of noncoding DNA sequences. We use a probabilistic method for sequence alignment (P.D.K. and T. Johnson, unpublished work), based on an evolutionary model of indel evolution. We focus our analysis on noncoding DNA sequences associated with well-annotated loci and use estimates of levels of constraint to infer the fraction of deleterious mutations in noncoding DNA.

Methods

Compilation of DNA Sequence Data. We compiled coding and adjacent noncoding DNA sequences from orthologous mouse

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: FEI, fastest evolving intronic.

*To whom correspondence should be addressed. E-mail: p.keightley@ed.ac.uk.

and rat loci by random sampling from their respective whole genome sequence assemblies in GenBank. Using the mouse genome as the reference, we randomly selected chromosomes in proportion to their lengths in Mb, then randomly selected positions within chromosomes from a uniform distribution. The nearest locus to this position for which annotation evidence included at least one complete mRNA sequence in both species was chosen. The first 200 loci were sampled at random irrespective of their distances to the next coding sequences. To increase the sample size of loci with long intergenic regions, we sampled an additional 100 loci for which the annotation in both mouse and rat indicated that the nearest coding sequence was >6 kb away. The coding sequences were aligned by using CLUSTAL (20), and positions of gaps were examined and adjusted if necessary. Sequences of lengths of up to 6,000 bp from 5′ upstream and 3′ downstream regions of coding sequences were extracted, along with the first and last introns and one other randomly selected intron. Some genes contain extremely long introns, particularly those that are lowly expressed (21), so we initially focused our analysis of constraint on the first and last 1,000 bp, if intron length exceeded 2,000 bp, otherwise we analyzed complete intron sequences. A further data set of up to 12,000 bp from first introns was subsequently extracted. Only clearly orthologous intergenic and intronic sequences were analyzed; we used a moving window of 40 bp to check the degree of sequence divergence in putative alignments. In some cases we observed sharp jumps in the divergence to ≈60% (the divergence expected for alignment of nonorthologous mouse-rat sequences), whereas the typical mouse-rat divergence is ≈15%. We interpreted these as being caused by a long insertions or deletions or sequence assembly errors. Such obviously nonhomologous regions were excluded from the analysis.

Sequence Alignment. Noncoding DNA sequences were aligned by using a Monte Carlo alignment procedure, MCALIGN, which searches for the alignment of highest probability based on a specific evolutionary model of noncoding DNA sequence evolution (P.D.K. and T. Johnson, unpublished work). Briefly, the parameters of the model are θ, the rate of indels relative to the rate of nucleotide substitutions, and a vector w, the frequency distribution of indel lengths. These parameters are estimated empirically from other data (see below). The Monte Carlo procedure carries out an uphill search of the parameter space of plausible alignments by accepting or rejecting proposal alignments depending on their relative probabilities. New proposal alignments are generated by a set of indel shuffling routines.

The parameters for the alignment model (θ and w) were estimated from 27 orthologous intron sequences of the closely related mouse species *Mus domesticus*, *Mus spretus*, and *Mus caroli*, for which nucleotide and indel divergences are sufficiently low as to make alignments by heuristic methods practically unambiguous. In comparisons between *M. domesticus* and *M. spretus* (10 loci), θ was 0.188, and between *M. domesticus* and *M. caroli* (8 loci), θ was 0.125. A weighted average estimate of θ = 0.146 was used to parameterize the alignment model. The empirical distribution of indel lengths is shown in Fig. 1. To parameterize the alignment model a value for $w_1 = 0.565$ (the empirical value) was assumed; for $i > 1$, values of w_i were estimated by minimizing the sum of squares about a smoothing function, $w_i = \beta/\alpha^i$, where β is a normalizing constant and α is the smoothing parameter. The estimated value of α was 1.45. We used intronic data to parameterize the alignment model for intergenic DNA, which is an approximation. However, indels occur at a lower frequency in intergenic than intronic DNA, on average, and using this approximation will give alignments very close, on average, to the true alignments for the degree of sequence divergence between mouse and rat (unpublished data).

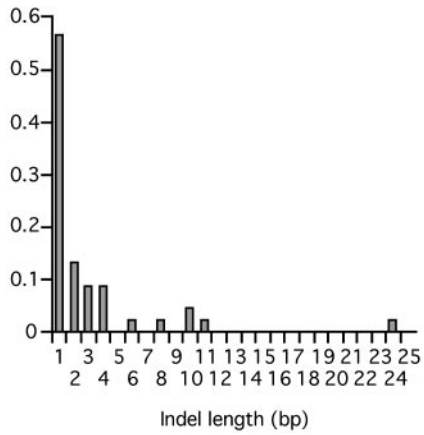

Fig. 1. Frequency distribution of indel length from between-species comparisons of closely related mouse species.

Masking of Microsatellite Repeats. Repeats of type $(XY)_n$, $(XYZ)_n$, $(XYY)_n$, $(XYYY)_n$, and $(XXYYY)_n$, where $n = 5$ in intronic or intergenic regions were excluded from the analysis, because their evolution is not driven by single nucleotide substitutions (22). Sequences adjacent to perfect microsatellite regions showing >80% homology to the specific repeat were also excluded. In addition, it was frequently observed that boundaries of microsatellites contained short stretches of obviously nonhomologous nucleotides, so 5l nucleotides adjacent to microsatellites, where l is the repeat length, were also excluded.

Calculation of Evolutionary Constraint. We calculated constraint in noncoding regions by extending a method previously developed for coding sequences (6). We used substitution rates at FEI sites to predict expected numbers (E) of substitutions in adjacent noncoding sequences (i.e., intergenic DNA, intronic splice sites, or first introns), under the assumption that point mutation rates of each possible kind are equal at FEI sites and the adjacent noncoding DNA sites. We calculated constraint by comparing E to numbers of observed substitutions (O):

$$C = 1 - O/E. \quad [1]$$

There are substantial differences in substitution rates between different kinds of nucleotide (23), so we needed to account for differences in substitution rates between FEI sites and adjacent noncoding regions. For each possible substitution type $i = 1.6$ (A↔T, A↔C, A↔G, T↔C, T↔G, C↔G), let k_i be the pairwise divergence in the FEI segment, i.e.,

$$k_i = \delta_i/N_i, \quad [2]$$

where d_i is the number of pairwise differences of type i, and N_i is the number of sites at which a change of type i could occur in one step (e.g., for A↔T changes, these sites are A/A, T/T, and T/A). The expected number of substitutions in an adjacent noncoding segment is,

$$E = \sum_{i=1}^{6} k_i M_i, \quad [3]$$

where M_i is the corresponding number of noncoding sites. This model assumes symmetric mutation rates and equivalent base composition in the FEI sites and the noncoding region of interest. However, analysis in which polarity of substitution was assigned via the relative probability of ancestry of each base gave very similar results (data not shown). The method to calculate

Table 1. Proportions of differences at nucleotides within and outside of CG dinucleotides at 4-fold and FEI sites

	Type of nucleotide change					
	A↔T	A↔C	A↔G	T↔C	T↔G	C↔G
Four-fold, within CG	—	0.182 (0.012)	0.468 (0.014)	0.468 (0.015)	0.199 (0.011)	0.149 (0.007)
FEI, within CG	—	0.160 (0.006)	0.385 (0.007)	0.382 (0.007)	0.163 (0.006)	0.105 (0.003)
Four-fold, outwith CG	0.0273 (0.0013)	0.0265 (0.0012)	0.0563 (0.0018)	0.0624 (0.0019)	0.0183 (0.0010)	0.0161 (0.0009)
FEI, outwith CG	0.0293 (0.0005)	0.0315 (0.0006)	0.0866 (0.0010)	0.0823 (0.0010)	0.0321 (0.0006)	0.0276 (0.0006)

Entries are the proportions of nucleotide changes at corresponding categories of sites, where, for example, A↔C sites are A/C, C/A, A/A, or C/C in mouse/rat. Bootstrap SEMs are shown in parentheses.

constraint does not attempt to account for multiple hits. However, simulation results (data not shown) suggest that estimation bias is negligible for nucleotide divergence well in excess of that of mouse-rat (15%) and for substantial differences in GC content. Standard errors and confidence limits for C were calculated by bootstrapping the gene-specific values of O and E 10,000 times.

Calculation of Genomic Deleterious Mutation Rate. The contribution to the deleterious mutation rate from a DNA segment was calculated from the product of average deleterious mutation rate per site (u) and the number of nucleotide sites (s) in the segment. We subdivided sites into two classes: (i) sites preceded by C or followed by G, termed "CG-susceptible," or (ii) all other sites, termed "non-CG-susceptible." A weighted average of contributions from CG-susceptible sites (mutation rate = u_1; number of sites = s_1) and non-CG-susceptible sites (mutation rate = u_2; number of sites = s_2) was taken. The deleterious mutation rate per site (u) was calculated from the product of constraint in the corresponding segment (Eq. **1**) and the nucleotide divergences specific to non-CG-susceptible and CG-susceptible sites, calculated by using Kimura's two-parameter method (9). The contribution of intergenic or intronic DNA to the genomic deleterious mutation rate per diploid (U) was calculated by summing the average contributions of segments:

$$U = Z \sum_{i=1}^{segments} P_i l_i \frac{\sum_{j=1}^{loci} s_{1ij}u_{1ij} + s_{2ij}u_{2ij}}{\sum_{j=1}^{loci} s_{1ij} + S_{2ij}}, \qquad [4]$$

where l_i is the length of a segment (200 bp in the analyses carried out here), P_i is the fraction of loci that actually contain intergenic or intronic DNA in segment i, and Z is a constant to convert between the scale of nucleotide mutation rate and genomic mutation rate per generation: $Z = 35{,}000$ loci $\times$ 0.5 (generations per year)/13 $\times$ 10^6 (approximate age in years of *mus-rattus* divergence; ref. 24). The inclusion of the term P_i was necessary because intergene regions and introns vary in length, and the fraction of loci containing a specific DNA segment declines as the distance in bp from the coding sequence to the start of the segment increases. For intergenic regions, values of P_i were calculated from the first 200 loci sampled (which were assumed to be a random sample with respect to intergene length) and for introns from all loci sampled. Lengths of intergene regions used to calculate P_i were taken as one-half of actual intergene lengths. Mutation rates in CpG islands (regions of the genome, often close to the 5′ end of genes, that are unusually rich in CG dinucleotides) are an order of magnitude lower than most CG sites (25). In the analysis, we assumed that mutation rates at nucleotides within CG-susceptible sites in CpG islands were the same as rates in non-CG-susceptible nucleotides.

Delimiting of CpG Islands. The locations of CpG islands were estimated by using the CpG Plot/CpG Report utilities available from the European Bioinformatics Institute (www.ebi.ac.uk/index.html; ref 26). A CpG island was reported if the observed to expected ratio of C plus G to CpG exceeded 0.6, in regions of GC content >50% (27), in a succession of 10 × 100-bp windows. Islands of >200 bp only were reported.

Results and Discussion

In a preliminary analysis, we compared nucleotide substitution rates at synonymous sites with rates for various categories of noncoding DNA sites. The analysis showed that introns evolve substantially faster, on average, than intergenic DNA, and that the rate of nucleotide substitution in intron sequences close to exon boundaries is slower than intron sequences in general. The first 1,000 bp of the 5′ region of first introns evolve noticeably slower than introns in general. These preliminary results are not shown, but they are implicit in the results on selective constraints that follow. Based on the preliminary analysis, we used sequences in introns, excluding intron 1, and 6 bp at the 5′ end and 30 bp at the 3′ end of each intron, as the FEI sequences. These sequences are used as standards to infer mutation rates in the subsequent analyses.

Comparison of Proportions of Substitutions in CG Dinucleotide and Non-CG Dinucleotide Sites. Proportions of nucleotide differences at 4-fold degenerate sites and FEI sites are shown in Table 1, split according to whether or not nucleotides are part of CG dinucleotides in either species. The high fraction of differences at CG dinucleotide sites in both 4-fold and noncoding DNA implies that CG dinucleotide sites are close to saturation. Proportions of nucleotide differences within CG dinucleotides are higher at 4-fold sites than FEI sites, whereas proportions outside of CG dinucleotides are higher at FEI sites than 4-folds. However, this

Table 2. Proportions of nucleotide differences at non-CG-susceptible 4-fold and FEI sites

	Type of nucleotide change					
	A↔T	A↔C	A↔G	T↔C	T↔G	C↔G
Four-fold sites	0.0175 (0.0021)	0.0272 (0.0017)	0.0827 (0.0030)	0.0753 (0.0029)	0.0270 (0.0019)	0.0298 (0.0017)
FEI sites	0.0259 (0.0007)	0.0327 (0.0007)	0.0924 (0.0012)	0.0880 (0.0011)	0.0332 (0.0007)	0.0316 (0.0007)

Entries are defined as for Table 1. Bootstrap SEMs are shown in parentheses.

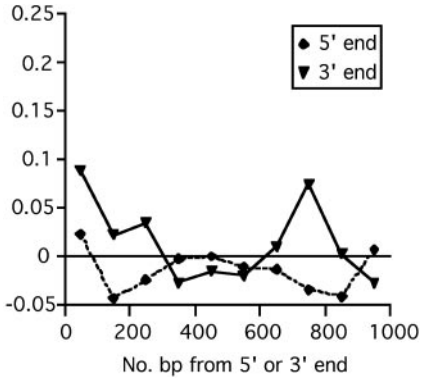

Fig. 2. Evolutionary constraint plotted against distance from the coding sequence (bp) in 100-bp blocks of the 5′ and 3′ ends of FEIs.

pattern is largely caused by ascertainment bias: the high selective constraints at amino acid sites in coding DNA cause ancestral CG sites to be more frequently correctly assigned to the CG category of sites than sites in relatively unconstrained intronic DNA. Conversely, ancestral CG sites in intronic DNA have a high probability of mutation away from CG in both species, so are more often incorrectly assigned to the category of non-CG sites than 4-fold sites. It is therefore inappropriate to simply exclude sites within CG dinucleotides. A less biased procedure was found to be to exclude CG-susceptible sites, those nucleotide sites that are preceded by C or followed by G in either species, and therefore have a high chance of being part of an ancestral CG dinucleotide. Simulations of noncoding DNA evolution including hypermutable CG dinucleotides suggested that such a procedure gives relatively unbiased estimates of constraint for cases of overall nucleotide divergence similar to mouse and rat (results not shown). In all subsequent analyses, this procedure was followed for calculating constraint.

Comparison of Substitution Rates at Non-CG-Susceptible Sites Between 4-Fold Sites and FEI Sites. Outside of CG-susceptible sites, fractions of nucleotide differences at 4-fold sites are consistently lower than at FEI sites (Table 2). It is notable that the fraction of A↔T changes at A/T sites is ≈30% lower at 4-fold sites than at FEI sites. Because A/T sites are four mutational changes from a CG-susceptible site, this finding suggests that the slower rate of substitution at 4-fold sites is unlikely to be a consequence of incorrect assignment of CG dinucleotide status. It is possible that the effect is a consequence of selection, although a role for selection at synonymous sites has been discounted (28). Slower rates of nucleotide substitution at 4-fold sites than noncoding sites have been reported in primates (29, 30).

Evolutionary Constraint in Intronic DNA. In FEIs, the average level of constraint is zero, by definition, because FEIs are assumed in this analysis to be the neutrally evolving standard against which constraint is measured. We tested for variation about this average by calculating mean constraint in 100-bp segments of the FEIs (i.e., the complete FEI data set was used to calculate constraint specific to intronic segments; Fig. 2). Mean constraint is nonsignificantly different from zero along the whole 1,000-bp length at the 5′ end of FEIs and is also nonsignificantly different from zero at the 3′ end of FEIs, with the exception of the first 100 bp at the 3′ end ($P < 0.001$; presumably associated with intronic splice control), and a marginally significantly constrained region at 700–800 bp of the 3′ end ($P = 0.02$). We examined constraint in more detail near 5′ and 3′ splice control regions (Table 3). As expected, there is a strong signal of purifying selection at the dinucleotides adjacent to the 5′ and 3′

Table 3. Estimates of mean selective constraint in intron sequences

Intronic DNA data set	Constraint
5′ bases 1–2	1.0 (0.0)
5′ bases 3–6	0.57 (0.044)
5′ bases 7–10	0.025 (0.076)
3′ bases 1–2	1.0 (0.0)
3′ bases 3–16	0.31 (0.031)
3′ bases 17–30	0.15 (0.040)
Intron 1, 5′ end	0.10 (0.017)
Intron 1, 3′ end	0.0056 (0.016)

In the analysis of intron 1, up to 6,000 bp at the 3′ or 5′ ends were analyzed. If a first intron was <12,000 bp long, the intron sequence was divided equally at the central nucleotide between data sets of 5′ and 3′ sequences. SEMs are shown in parentheses.

splice sites (which are invariant), and in the sequences within 6 bp and ≈30 bp of the 5′ and 3′ ends, respectively, known from previous work to be intimately involved in intron splicing and to be conserved across taxa (31). It has recently been shown that there is higher frequency of transcriptional regulatory sequences in first introns than introns in general (32). Analysis of our data set also supports this observation by revealing constrained sequences in intron 1, located within ≈2 kb of the 5′ end (Table 3 and Fig. 3*A*). The 3′ ends of first introns evolve at a similar rate to FEIs (Table 3).

Evolutionary Constraint in Intergenic DNA. Evolutionary constraint in intergenic regions is moderately strong close to the 5′ and 3′ ends of coding sequences, then drops off surprisingly slowly as a function of distance from the gene (Fig. 3*B*). Some 5′ and 3′ intergenic regions are extremely strongly conserved: ≈5% of loci contain runs of 100 bp within 200 bp of the start or stop codon that are identical between mouse and rat (average sequence

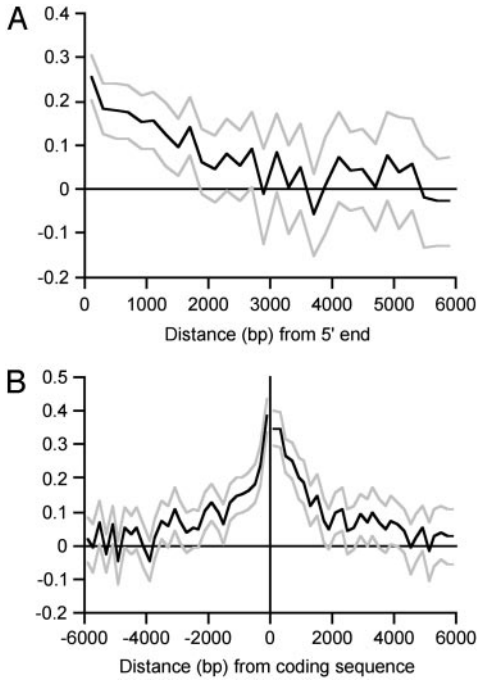

Fig. 3. Evolutionary constraint plotted against distance from the coding sequence (bp) calculated in 200-bp blocks of the 5′ end of first introns (*A*) and in intergenic regions (*B*). The upper and lower 95% confidence limits are shown in light gray.

Table 4. Estimates of selective constraint in coding, intronic, and intergenic DNA of rodents, and contributions to the genomic deleterious mutation rate (*U*) per diploid genome per generation

DNA category	Nucleotide sites per locus	Mean constraint, SEM	Contribution to *U*
Coding	1,125*	0.87 (0.009)	0.22¶
5′ intronic splice regions	44.4†	0.73 (0.027)	0.0071
3′ intronic splice regions	222†	0.29 (0.024)	0.012
Intron 1, 5′ end	3,307‡	0.10 (0.017)	0.049
5′ intergenic	5,596§	0.093 (0.013)	0.074
3′ intergenic	5,271§	0.12 (0.015)	0.079

Estimates were made under the assumption that there are 35,000 protein coding loci in the mouse genome (10, 11).

*The average length of rodent coding sequences is ≈1,500 nt, and about three-quarters of sites in coding sequences generate an amino acid change if mutated.

†Blocks totaling 6 and 30 nt near 5′ and 3′ splice junction sites, respectively, show significant evidence of selective constraint (Table 1), and there are an average of 7.4 introns per locus (11).

‡Blocks of up to 6,000 bp (excludes splice control regions).

§Blocks of up to 6,000 bp upstream or downstream from the coding sequence were analyzed.

¶Estimate based on ref. 8, but assuming 35,000 rather than 80,000 loci, calculated under the assumption that mice and rats diverged 13 million years ago (24) and have two generations per year (8).

divergence ≈15%). Mean constraint has dropped to levels close to zero by ≈4,000 bp from the coding sequence (Fig. 3*B*).

Contribution of Noncoding DNA to Overall Deleterious Mutation Rate. Under the assumption that there are 35,000 rodent genes (10, 11), we calculated the contributions of coding, intronic, and intergenic DNA to *U* (Table 4). In the set of loci analyzed, evolutionary constraint at amino acid sites calculated by a method as described (6) is 0.87 (SEM = 0.009), which is a typical value for rodent loci (33), and the contribution to *U* is 0.22. The overall estimate of *U* in noncoding DNA, summing over contributions from intronic and intergenic DNA, is also 0.22. This estimate for noncoding DNA is conservative for several reasons. (*i*) It does not include the contribution from constrained nucleotides outside the 6-kb 5′ and 3′ intergenic segments analyzed. This contribution is likely to be small, however, because ≈95% of well-characterized gene regulatory regions in murine intergenic regions are within 2 kb of promoters (11). (*ii*) The estimate will be too low if there are substantial selective constraints in FEIs. (*iii*) It does not include a contribution from indels. (*iv*) Our estimates of numbers of constrained nucleotides do not include sites under weak selection (with selection coefficients close to $1/N_e$). Such weakly selected mutations contribute to the mutation load (34, 35) and can have an appreciable probability of fixation, but the fraction of mutations with effects close to $1/N_e$ is relatively small for many reasonable distributions of selection coefficients.

In rodents, an overall estimate for *U* is ≈0.44 (Table 4). However, *U* is positively correlated with generation time (8), and *U* could be considerably higher for longer-lived taxa such as hominids. For example, an estimate for the mean level of constraint at amino acid sites in a sample of human and chimpanzee genes is 0.69 (33), and the generation time for hominids is ≈20 years (6). These estimates suggest that *U* for amino acid sites of protein-coding genes in hominids is ≈1.5 (8). If the proportion of deleterious mutations in noncoding DNA is similar among mammalian taxa, a genomic estimate for *U* (including point mutations in both coding and noncoding DNA) in hominids is therefore 3.0. Under a multiplicative model, the resulting mutation load (95%) is so high as to imply that nonmultiplicative effects of mutations are important in reducing the load in hominids.

The high frequency of deleterious mutations in intergenic DNA contrasts sharply with the low frequency of regulatory mutations associated with human Mendelian genetic diseases (≈1% of point mutations; ref. 36). This finding suggests that deleterious mutations in noncoding DNA are predominantly quantitative in nature and could be an important source of quantitative trait variation and of the burden of complex genetic disease in human populations. Human complex trait association mapping programs may therefore gain enhanced efficiency by concentrating markers in the regions of high constraint indicated by our study.

We thank Adam Eyre-Walker and Alex Kondrashov for helpful discussions and two reviewers for useful comments on an earlier version.

1. Crow, J. F. & Simmons, M. J. (1983) in *The Genetics and Biology of Drosophila*, eds. Ashburner, M., Carson, H. L. & Thompson, J. N. (Academic, London), Vol. 3C, pp. 1–35.
2. Crow, J. F. (2000) *Nat. Rev. Genet.* **1,** 40–47.
3. Muller, H. J. (1950) *Am. J. Hum. Genet.* **2,** 111–176.
4. Kondrashov, A. S. (1988) *Nature* **336,** 435–440.
5. Whitlock, M. C. (2002) *Genetics* **160,** 1191–1202.
6. Eyre-Walker, A. & Keightley, P. D. (1999) *Nature* **397,** 344–347.
7. Nachman, M. W. & Crowell, S. L. (2000) *Genetics* **156,** 297–304.
8. Keightley, P. D. & Eyre-Walker, A. (2000) *Science* **290,** 331–333.
9. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
10. International Human Genome Sequencing Consortium (2001) *Nature* **409,** 860–921.
11. Mouse Genome Sequencing Consortium (2002) *Nature* **420,** 520–562.
12. Clark, A. G. (2001) *Genome Res.* **11,** 1319–1320.
13. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).
14. Kondrashov, A. S. & Crow, J. F. (1993) *Hum. Mutat.* **2,** 229–234.
15. Bergman, C. M. & Kreitman, M. (2001) *Genome Res.* **11,** 1335–1345.
16. Jareborg, N., Birney, E. & Durbin, R. (1999) *Genome Res.* **9,** 815–824.
17. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, F. A. & Kondrashov, A. S. (2001) *Trends Genet.* **17,** 373–376.
18. Hare, M. P. & Palumbi, S. R. (2003) *Mol. Biol. Evol.* **20,** 969–978.
19. Thorne, J. L., Kishino, H. & Felsenstein, J. (1991) *J. Mol. Evol.* **33,** 114–124.
20. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
21. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. (2002) *Nat. Genet.* **31,** 415–418.
22. Calabrese, P. & Durrett, R. (2003) *Mol. Biol. Evol.* **20,** 715–725.
23. Li, W.-H., Wu, C.-I. & Luo, C.-C. (1984) *J. Mol. Evol.* **21,** 58–71.
24. Jaeger, J. J., Tong, H. & Denys, C. (1986) *C. R. Acad. Sci. Ser. II* **302,** 917–922.
25. Bird, A. (1986) *Nature* **321,** 209–213.
26. Rice, P., Longden, I. & Bleasby, A. (2000) *Trends Genet.* **16,** 276–277.
27. Antequera, F. & Bird, A. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11995–11999.
28. Smith, N. G. C. & Hurst, L. D. (1999) *Genetics* **152,** 661–673.
29. Chen, F.-C. & Li, W.-H. (2001) *Am. J. Hum. Genet.* **68,** 444–456.
30. Hellmann, I., Zollner, S., Enard W., Ebersberger, I., Nickel, B. & Paabo, S. (2003) *Genome Res.* **13,** 831–837.
31. Sharp, P. A. (1994) *Cell* **77,** 805–815.
32. Majewski, J. & Ott, J. (2002) *Genome Res.* **12,** 1827–1836.
33. Eyre-Walker, A., Keightley, P. D., Smith, N. G. C. & Gaffney, D. (2002) *Mol. Biol. Evol.* **19,** 2142–2149.
34. Ohta, T. (1973) *Nature* **246,** 96–98.
35. Kondrashov, A. S. (1995) *J. Theor. Biol.* **175,** 583–594.
36. McKusick, V. A. (1998) *Mendelian Inheritance in Man: Catalogs of Human Genes and Genetic Disorders* (Johns Hopkins Univ. Press, Baltimore), 12th Ed.