Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome

Ivan P. Gorlov1,*, Marek Kimmel2,3 and Christopher I. Amos1

1Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA, 2Department of Statistics, Rice University, Houston, TX, USA and 3System Engineering Group, Silesian Institute of Technology, Akademicka 16, 44-100 Gliwice, Poland

Received November 10, 2005; Revised December 29, 2005; Accepted February 10, 2006

Using available information on the total absolute size of the coding region of the human genome, data on codon usage and pseudogene-derived mutation rates for different single nucleotide substitutions we have estimated, for the human genome, the potential numbers of mutation events capable to produce: (1) nonsense; (2) missense (radical and conservative); (3) silent; (4) splice; and (5) protein-elongating (those changing wild-type stop codon into an amino acid encoding codon) mutations. We used the NCBI dbSNP database to retrieve data on the observed number of polymorphisms of each category. The fraction of polymorphisms in each category among all potential events in the genome depends on the strength of selection: the higher the rate of polymorphism, the weaker the selection. We used nonsense mutations as a referent group. Compared with nonsense mutations, we found that the relative selection coefficient against protein-elongating mutations was 21%, and the relative selection was 12% against missense mutations. Radical missense mutations were found to be four times more deleterious compared to conservative ones. Surprisingly, we found that silent mutations on average are not neutral; with the average harmfulness of 3% of nonsense mutations. Silent mutations may be deleterious when they affect splicing by creating cryptic donor-acceptor sites or by disturbing exonic splicing enhancers (ESESs). The average selection coefficient against splice mutations was 48% of that against nonsense mutations. Converting the relative selection coefficients into absolute ones using data on loss-of-function mutations in Saccharomyces cerevisiae and Caenorhabditis elegans, or by analysis of the expected frequency of mutations in the human genome, suggested that genetic drift could play a role in population dynamics of conservative missense and silent mutations.

INTRODUCTION

Single-nucleotide polymorphisms (SNPs) are common sequence variants in the human genome (1). SNPs occur, on average, every 500–1000 bases (2–6). SNP density is lower in exonic than in intronic sequences, which suggests that purifying selection plays a role in the distribution of SNPs (7). Sunyaev et al. (8) compared synonymous and non-synonymous SNPs in human genes with the corresponding regions in murine homologues and concluded that most functional SNPs in the human genome are deleterious.

It long has been known that the substitution of an amino acid by amino acids with similar physicochemical properties is maintained during evolution more frequently than substitution of chemically different amino acids (9–11). This suggests that, on average, the substitution of an amino acid by a chemically similar one can be less deleterious compared with the substitution by a chemically different amino acid. Several classification schemes of amino acids were more recently developed (9,12,13). One of the first classifications of amino acids is Grantham’s chemical difference matrix (11), a multivariate combination of residue side chain composition,
polarity and volume differences, to quantify the severity of amino acid changes. A classification approach used Dagan et al. (10) a combination of Grantham's distance with polarity, volume and change and has been shown to correlate with intensity of selection.

Different categories of SNPs are expected to vary in deleteriousness; for example, missense mutations are more deleterious than silent mutations (14). It is of theoretical and practical interest to estimate the average deleteriousness of different functional categories of point mutations in coding regions of the human genome. In this study, we used the National Center for Biotechnology Information dbSNP database to estimate the relative selection coefficients against different functional categories of point mutations in coding regions of the human genome.

RESULTS

SNP retrieval

A total of 29 169 true (i.e. originated by point mutations) and frequency-validated SNPs in the coding regions of the human genome were retrieved from the dbSNP database. The numbers of the retrieved SNPs according to the functional categories of point mutations are shown in Table 1.

Average mutation rates

Relative mutation rates for the 12 variants of the point mutations derived from the analysis of processed pseudogenes are available (Table 2). We ascribed these mutation rates to each of the 576 possible wild-type (WT)-mutant codon pairs (Supplementary Material, Table S1) and computed the average relative mutation rates for the six functional categories used in our study (Table 3).

Table 1. Observed (Obs) and expected (Exp) fractions of the six functional categories of SNPs in the coding regions of the human genome

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of WT-mutant pairs</th>
<th>Exp fraction of SNPs (%) (adjustments)</th>
<th>Obs SNPs</th>
<th>Obs fraction/Exp fraction</th>
<th>Log of Obs fraction/Exp fraction*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>CU</td>
<td>CU and $\mu$</td>
<td>No.</td>
</tr>
<tr>
<td>Nonsense</td>
<td>23</td>
<td>4.0</td>
<td>3.9</td>
<td>3.6</td>
<td>88</td>
</tr>
<tr>
<td>Protein-elongating</td>
<td>23</td>
<td>4.0</td>
<td>0.2</td>
<td>0.2</td>
<td>21</td>
</tr>
<tr>
<td>Missense</td>
<td>392</td>
<td>68.1</td>
<td>73</td>
<td>66.9</td>
<td>13838</td>
</tr>
<tr>
<td>Radical missense</td>
<td>286</td>
<td>49.7</td>
<td>62.9</td>
<td>48.2</td>
<td>8378</td>
</tr>
<tr>
<td>Conservative missense</td>
<td>106</td>
<td>18.4</td>
<td>10.1</td>
<td>18.7</td>
<td>5460</td>
</tr>
<tr>
<td>Silent</td>
<td>138</td>
<td>23.9</td>
<td>22.9</td>
<td>29.4</td>
<td>15222</td>
</tr>
</tbody>
</table>

CU, codon usage; $\mu$, mutation rate.

Table 2. Relative frequencies of single-nucleotide substitutions estimated by the analysis of processed pseudogenes [after (47)]

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>—</td>
<td>3.4 ± 0.7</td>
<td>4.5 ± 0.8</td>
<td>12.5 ± 1.1</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>3.3 ± 0.6</td>
<td>—</td>
<td>13.8 ± 1.9</td>
<td>4.2 ± 0.5</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>4.2 ± 0.5</td>
<td>20.7 ± 1.3</td>
<td>—</td>
<td>4.6 ± 0.6</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>20.4 ± 1.4</td>
<td>4.4 ± 0.6</td>
<td>4.9 ± 0.7</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

The overall mutation rate for the 12 variants is 100%.

Table 3. Average relative mutation rates ($\mu$) for the six functional categories of SNPs in the coding regions of the human genome

<table>
<thead>
<tr>
<th>Category</th>
<th>Average relative $\mu$</th>
<th>Average relative $\mu/\mu_{\text{non}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsense</td>
<td>7.48 ± 1.48</td>
<td>1.00</td>
</tr>
<tr>
<td>Protein-elongating</td>
<td>5.93 ± 0.84</td>
<td>0.79</td>
</tr>
<tr>
<td>Missense</td>
<td>7.87 ± 0.25</td>
<td>1.05</td>
</tr>
<tr>
<td>Radical missense</td>
<td>7.59 ± 0.36</td>
<td>1.01</td>
</tr>
<tr>
<td>Conservative missense</td>
<td>8.61 ± 0.63</td>
<td>1.15</td>
</tr>
<tr>
<td>Silent</td>
<td>10.20 ± 0.59</td>
<td>1.36</td>
</tr>
</tbody>
</table>

$\mu_{\text{non}}$, $\mu$ for nonsense mutations.

Observed and expected proportions of the functional categories

The proportions of the potential sites for the functional categories of point mutations adjusted to codon usage and mutation rates are shown in Table 1. The proportions given in the columns 3–5 are expected when there is no variation in the strength of purifying selection against different categories of the mutations before and after adjustments for codon usage, mutation rates. The observed proportions were estimated on the basis of the observed numbers of SNPs corresponding to each functional category reported in the dbSNP database.

We found that the observed proportions of the nonsense, protein-elongating and radical missense mutations were lower than expected and that the observed relative proportions of the conservative missense and silent mutations were higher than expected (Table 1 and Fig. 1A). The variations in proportions were most easily interpreted when the log of the ratio of the observed to expected proportions was used (Fig. 1B). This pattern can be explained by inter-categorical differences in the strength of purifying selection: stronger purifying selection against nonsense, protein-elongating and radical missense mutations produces a deficit of these mutations and therefore tips their proportions downwards,
whereas the proportions of conservative missense and silent mutations are correspondingly inflated. Therefore, our results suggest inter-categorical variation in the strength of purifying selection.

Predicted average relative selection coefficients

Equation (7) was used to compute average relative selection coefficients. Table 4 shows these coefficients and their 99% confidence intervals for each functional category. We found that relative to the average selection coefficient against nonsense mutations, the coefficient was 21% against protein-elongating mutations, 3% against silent mutations and 12% against missense mutations. The average relative selection coefficient against radical missense mutations was four times higher than that against conservative missense mutations. The average relative selection coefficient against conservative missense mutations was slightly higher than that against silent mutations.

Deleteriousness of splice mutations

The dbSNP database contains reports on 95 frequency-validated splice mutations. Therefore, the proportion of polymorphisms for a splice site is \( \frac{95}{1,870,000} \approx 5.1 \times 10^{-5} \). Given that the proportion of polymorphic nonsense mutations among all potential nonsense mutations in the genome (\( P_{\text{pol}}^{\text{nonsense}} \)) is \( \sim 2.2 \times 10^{-5} \), the average relative selection coefficient against splice mutations is 48% (95% confidence interval 33–63%) that against nonsense mutations. Given the median length of an exon is \( \sim 120 \) nucleotides (15) and that there can be several sites in an exon that are important for splicing (16–19), disturbance of normal splicing by silent mutations can explain their relatively high deleteriousness (3% versus expected 0%).

DISCUSSION

Our approach gives an unbiased estimate of average equilibrium frequency of a mutant allele \( q_i \) (see Materials and Methods and Appendix). Confidence intervals for the estimated relative selection (\( s_{rel} \)) coefficients are very narrow for all functional categories analyzed except protein-elongating mutations (see Appendix and Table 4). We found substantial variation in the average relative selection coefficients against different functional categories of point mutations in coding regions of the human genome. After nonsense mutations, the strongest selection coefficient was for protein-elongating mutations (21% that for nonsense mutations). Protein-elongating mutations have an amino acid-encoding codon in place of a normal stop codon and are therefore likely to disturb protein structure and function, though not as much as nonsense mutations.

The relative strength of purifying selection against radical missense mutations was 17% that against nonsense mutations. Radical missense mutations change a WT amino acid into a

Table 4. Predicted relative selection coefficients against different functional categories of SNPs in the coding regions of the human genome

<table>
<thead>
<tr>
<th>Category</th>
<th>NPS</th>
<th>OBS</th>
<th>( P_{\text{pol}}^{\text{pol}} )</th>
<th>Average relative selection coefficient (%) and 99 CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsense</td>
<td>4,000,667</td>
<td>88</td>
<td>0.0000220</td>
<td>100 (referent)</td>
</tr>
<tr>
<td>Protein-elongating</td>
<td>198,333</td>
<td>21</td>
<td>0.0001059</td>
<td>20.8 (13.0–28.6)</td>
</tr>
<tr>
<td>Missense</td>
<td>74,465,667</td>
<td>13,838</td>
<td>0.0001858</td>
<td>11.8 (9.9–14.4)</td>
</tr>
<tr>
<td>Radical missense</td>
<td>64,194,267</td>
<td>8,378</td>
<td>0.0001305</td>
<td>16.9 (13.3–20.5)</td>
</tr>
<tr>
<td>Conservative</td>
<td>10,271,400</td>
<td>5,460</td>
<td>0.0005316</td>
<td>4.1 (3.1–5.1)</td>
</tr>
<tr>
<td>Silent</td>
<td>23,335,333</td>
<td>15,222</td>
<td>0.0006523</td>
<td>3.4 (2.5–4.3)</td>
</tr>
</tbody>
</table>

NPS, number of potential sites of the \( i \)th category; OBS, observed number of polymorphisms; \( P_{\text{pol}} \), fraction of polymorphisms estimated as a ratio of OBS to NPS.
mutant amino acid that is chemically different. These missense mutations are expected to disturb protein structure when they are located in sites important for protein folding or in functional sites. Conservative missense mutations were four times less deleterious than radical ones. The overall relative deleteriousness of missense mutations in the human genome was 12% that of nonsense mutations.

We found a relatively high deleteriousness of silent mutations in the human genome (3% of the average deleteriousness of nonsense mutations). Silent mutations do not change amino acid sequence and are generally believed to be selectively neutral, although in some cases, they can be extremely deleterious. It has been demonstrated that when silent mutations disturb a canonical splice site located near an exon–intron boundary, they can cause abnormal splicing and lead to the production of non-functional protein (16,17,20–22). Silent mutations can also disturb splicing by activating cryptic splice sites (18,23,24) or exonic splicing enhancers (25). Aside from the effects on splicing, silent mutations can disturb mRNA processing and transport (26,27). The widespread biases in codon usage can also cause non-neutrality of the silent mutations: if a silent mutation switches encoding to a codon with a low pool of corresponding tRNA, it can severely decrease the efficiency of translation and protein production (28,29). We found that the average relative selection coefficient against splice mutations was half that against nonsense mutations. This result is in good agreement with the estimates of selection coefficients for splice mutations obtained by comparing human and chimpanzee sequences (Dr A.S. Kondrashov (NCBI), personal report). Because there can be several regions in an exon important for splicing and because the average deleteriousness of a splice mutation was relatively high (half that of nonsense mutations), it seems not surprising, after all, that silent mutations appear to confer substantial deleteriousness.

Our approach provides estimates of relative selection coefficients. To convert relative estimates into absolute ones, we need to know the absolute selection coefficients against at least one functional category. We believe it reasonable to suggest that most of the nonsense mutations are loss-of-function mutations. Estimates of the average drop in fitness in loss-of-function mutations are available for Saccharomyces cerevisiae and Caenorhabditis elegans. Knocking out the 890 metabolic genes in S. cerevisiae and testing the knockouts in eight different conditions demonstrated the average drop in fitness to be ~10–20% (30). Eliminating the function of the more than 16 000 C. elegans genes by RNA interference (RNAi) demonstrated that >90% of them have no detectable phenotypic effect (31). The proportion of the RNAi embryonic lethal phenotype was only 5.5% (32). The other two groups of phenotype (growth defects and post-embryonic phenotypes) make up another 5% of all phenotypic effects of RNAi loss of functions (32). If we suggest that the proportion of essential genes in the human genome is similar to that in yeast and C. elegans genomes, then the average selection coefficient against nonsense mutations in humans will be about 0.1.

Alternatively, we can directly obtain an estimate of the absolute selection coefficient against nonsense mutations by using the proportion of polymorphic sites among all nonsense sites in the genome. We have estimated from the dbSNP data that the average mutant allele frequency for polymorphic nonsense mutations is ~0.2. Then, the average mutant allele frequency of nonsense mutations q_n nonsense will be q_n nonsense = 0.2 \cdot P_{pol}^{nonsense} + q_{pol}^{mon} (1 − P_{pol}^{nonsense}) where P_{pol}^{nonsense} is the proportion of polymorphic sites and q_{pol}^{mon} is the true population frequency of the nonsense mutations that were never reported as polymorphisms. If we assume q_{pol}^{mon} equals zero, then q_n nonsense = 0.2 \cdot P_{pol}^{nonsense} = 0.2 \cdot 0.00002 = 4 \times 10^{-6}. On the basis of the formula (1), and assuming that coefficient of dominance equals 0.2 (33–35), an estimate of absolute selection coefficient s will be s = \mu / h q = 2 \cdot 10^{-8} / 0.2 \cdot (4 \times 10^{-6}) = 0.025. This estimate of s is about 1/4 that obtained using data on average fitness reduction of loss-of-function mutation in S. cerevisiae and C. elegans. The estimate could be biased because there is insufficient sequencing of the genome at present to identify rare variants, so that the q_{pol}^{mon} is underestimated and s could, therefore, be overestimated. Nevertheless, the estimate that we obtain from this method is not dramatically different from the estimate obtained by the evaluation of model organisms.

Estimates of the absolute selection coefficient for polymorphic nonsense mutations show that their deleteriousness is much lower than overall average for nonsense mutations. For the average frequency of the mutant allele of 0.2, the selection coefficient, based on formula (1), will be s = \mu / h q = 2 \cdot 10^{-8} / 0.2 \cdot 0.2 = 5 \times 10^{-7}. This means that polymorphic nonsense mutations are virtually neutral, whereas an average absolute selection coefficient against all nonsense mutations in the human genome is of the order of 10^{-1}–10^{-2}. Therefore polymorphic nonsense mutation is unlikely to explain mutational damage in man estimated based on the data on consanguineous marriages (36).

Our analysis is based on the assumption that level of polymorphism in the coding regions of the human genome is controlled by the purifying selection rather than drift. Genetic drift is expected to play a role when the selection coefficient is ~1/N_e. There is general agreement that the effective population size N_e for the human population is ~10^5 (37). Therefore, one can expect that random drift will be important when the selection coefficient is <10^{-4} (38,39). One can suggest that average effect of loss-of-function mutations is ~10^{-1} as indicated by observations in model organisms or from the direct computation we performed earlier. Considering nonsense mutations as a loss-of-function mutations, we can suggest that absolute selection coefficient against them to be ~10^{-1}. On the basis of our estimation of the relative selection coefficients, we can suggest that the absolute selection coefficient against all splice mutations is ~10^{-1}–10^{-2}, against all radical missense mutations is ~10^{-3} and an average selection coefficient against all conservative missense and silent mutations is ~10^{-3}–10^{-4}. Our estimates of selection coefficients do not exclude the possibility that genetic drift can play a role in the population dynamics of conservative missense and silent mutations, but would play a less important role for other mutations.

Fixation of the conservative missense and silent mutations due to genetic drift will result in decreasing the polymorphism that can be erroneously interpreted as an effect of purifying
selection. This means that our estimates of the relative selection coefficient for conservative missense and silent mutations can be upwardly biased. We believe, however, that this bias is unlikely to affect the conclusion that average deleteriousness of silent mutations is about several percent of that of nonsense mutations. Firstly, for genetic drift to play a major role in population dynamics of mutations, their deleteriousness should be much lower than critical $10^{-4}$. Our estimates, however, show that the deleteriousness of conservative missense and synonymous mutations is close to or slightly higher than $10^{-4}$. Secondly, average deleteriousness of silent mutations is comparable to that of conservative missense mutations, suggesting that silent mutations are on average slightly deleterious. Recently, Yampolsky et al. (40) have estimated a bird eye’s view distribution of missense mutations by their deleteriousness. Their analysis suggests that selection coefficient against majority of missense mutations should be $\sim 10^{-3} - 10^{-4}$ that is in good agreement with our estimates.

It is possible that nonsense mutations do not always lead to complete loss-of-function phenotype. Nonsense mutations located close to 3’ end of the coding region of the gene can retain some function. The mild effect of nonsense mutations can be relatively common in genes encoding large proteins with long non-functional or low functional N-terminus. An analysis of nonsense mutations in BRCA2 gene provides an example of a nonsense mutation with a mild effect on the protein function (41). Therefore, one can expect that average deleteriousness of nonsense mutations can be lower than deleteriousness of true loss-of-function mutations, suggesting that our estimates of absolute selection coefficients can be upwardly biased.

Our estimates can be used to predict the average fitness reduction (mutation load) attributable to new point mutations in coding regions of the human genome. There are $\sim 7.5 \times 10^7$ potential sites for missense mutations in the human genome. Given that the per-nucleotide, per-generation mutation rate is $\sim 2 \times 10^{-8}$ (42,43), the average number of new missense mutations per genome per generation will be $\sim 1.5$. According to our estimates, the absolute selection coefficient against missense mutations is $\sim 0.01$, the dominance coefficient is $\sim 0.2$ (33–35) and the average decrease in fitness attributed to the \textit{de novo} missense mutations is $\sim 0.3\%$ per generation. Similar calculations show that \textit{de novo} nonsense mutations can cause an $\sim 0.2\%$ reduction in fitness per generation, protein-elongating mutations $\sim 0.1\%$ and silent mutations $\sim 0.2\%$. Therefore, the overall total per-generation reduction in fitness attributable to point mutations in coding regions of the human genome is $\sim 1\%$.

**MATERIALS AND METHODS**

The dataset

The dbSNP is the largest public database for genomic human SNP polymorphisms (44). The latest build of the database (build 125) includes 11 170 620 submissions. Different submitters target different regions of the human genome. The size of the targeted regions depends on the method used and can vary from 100 nucleotides to thousands of nucleotides (for expanding a long-template polymerase chain reaction system). The choice of the target sequence is not random, if we consider a specific submission, but it is approximately random when results of several millions of independent submissions are pooled together. Therefore, the dbSNP database can be described as a collection of polymorphisms detected in many randomly chosen and overlapping fragments of the human genome. Depending on the position of the point mutation in the codon, a SNP can result in a nonsense, missense or silent mutation. Probing all nucleotide positions in the coding regions of the human genome by all possible point mutations allows the estimation of the absolute number of potential sites for a specific functional category (e.g. nonsense mutations). In a large random sample of DNA fragments from the human genome, the proportions of potential sites for different functional categories are expected to follow the proportions of the potential sites in the whole genome.

The quality of a SNP characterization depends on the method used to detect it and varies from submittor to submitter (45). To reduce discovery errors, only frequency-validated SNPs (i.e. those reported by at least two independent submitters and at least one submission validated by a non-computational method) were used in our study. The retrieved SNPs were stratified into six functional categories: (i) nonsense mutations, which change an amino acid-encoding codon into a stop codon, (ii) protein-elongating mutations, which modify a wild-type (WT) stop codon into an amino acid-encoding one, (iii) missense mutations, which change a WT amino acid into a mutant one and (iv) silent mutations, which do not change the amino acid encoded. We stratified missense mutations as (v) radical or (vi) conservative by adopting the classification system used by Dagan et al. (10). Briefly, all amino acids were subdivided into three groups according to their charge: positive (R, H, K), negative (D, E) and uncharged (A, N, C, Q, G, I, L, M, F, P, S, T, W, Y, V). The amino acids were further subdivided by volume and polarity: special (C), neutral and small (A, G, P, S, T), polar and relatively small (N, D, Q, E), polar and relatively large (R, H, K), non-polar and relatively small (I, L, M, V), and non-polar and relatively large (F, W, Y). We considered radical missense mutations to be those that change amino acid categories (e.g. R → L) and conservative missense mutations to be those that do not change amino acid category (e.g. L → V). We also analyzed splice mutations (i.e. single-nucleotide substitutions in the two first and two last nucleotides of an intron).

For each SNP, we extracted nucleotide and amino acid variants from the original dbSNP ASN.1 file. SNPs with no data on the type of amino acid substitutions were excluded from our analysis. In some cases, more than one mutant variant was reported for a specific rsID. For example, two nucleotide variants of rs2280279, ACG(Thr) and TCG(Ser), were reported. In these cases, each variant was treated as an independent polymorphism.

**Estimation of the number of potential sites**

Point mutations produce three mutant variants per nucleotide, for example, ‘A’ can mutate into ‘C’, ‘G’ or ‘T’. Thus, nine
point mutations are possible per codon, and for 64 codons there will be \(64.9 = 576\) possible WT–mutant codon pairs. These 576 pairs are all possible variants of the point mutations in a codon (see Supplementary Material, Table S1). Depending on the WT and mutant codons, each pair can be assigned to one of the nonsense, protein-elongating, radical missense, conservative missense or silent mutation category. The proportions of the different categories of the WT–mutant codon pairs are equal to the proportions of the potential sites for the corresponding functional categories in the human genome conditional on no differences in codon usage or mutation rate. Different codons are used, however, in the human genome with different frequencies, for example, the frequency of CTG(Leu) is 40 times higher than the frequency of TAA(Ter). To deal with the differences in codon usage, we weighted the numbers of WT–mutant codon pairs by codon frequencies provided by Nakamura et al. (46). To adjust for the variation in mutation rates, we used Graur and Li’s (47) pseudogene-derived matrix of nucleotide substitutions (Table 2). The data in Table 2 reflect mutation rates unaffected by selection because there are no selection constraints on nucleotide replacements in pseudogenes.

According to the International Human Genome Sequencing Consortium, the total length of the coding sequences in the human genome is \(\sim 34\) Mb (48). If we allow that each nucleotide can produce three point mutations, then the total potential number of point mutations in the human genome is \(\sim 1.02 \times 10^9\). On the basis of this estimated proportion of potential sites for each functional category, we predicted the absolute number of potential sites for each functional category.

### Computation of selection coefficients

There is general agreement that point mutations in coding regions of the human genome are under purifying selection. For deleterious mutations, there should be equilibrium between a decrease in the number of mutations due to purifying selection and an increase in the number of mutations due to new mutations. Let us consider a locus with two alleles ‘A and ‘a’ in a large, randomly mating, sexually reproducing population. Let ‘A’ be the normal allele and ‘a’ the mutant allele. Let the genotypic values for fitness be 1 for AA, \(1 - hs\) for Aa and \(1 - s\) for aa, where \(h\) is the dominance coefficient and \(s\) (\(0 < s < 1\)) is the selection coefficient. At equilibrium, the frequency of the mutant allele will be

\[
q = \frac{\mu}{hs},
\]

where \(q\) is the equilibrium frequency of mutant allele a and \(\mu\) is the mutation rate (i.e. the proportion of A alleles that mutate to a alleles each generation) (49). It is widely believed that the distributions of \(h\) on metric phenotypic traits are leptokurtic and highly deleterious mutations tend to be nearly recessive (33–35). However, the distribution of \(h\) on fitness is less leptokurtic with coefficient of dominance similar for different mutations (50). Therefore, it is reasonable to assume that fitness mutations have the same \(h\) (close to 0.1) (51). For simplicity, we suggest that \(h\) is the same for the different functional categories of point mutations.

The average equilibrium frequency of the mutant alleles for the \(i\)th category \(\tilde{q}_i\) is a parameter of interest. We cannot use only SNPs with reported allele frequencies because the estimate will be upwardly biased. To have an unbiased estimate of \(\tilde{q}_i\), one should consider all potential sites of the \(i\)th functional category in the genome regardless of whether they are polymorphic (i.e. segregating). Let \(\bar{P}_i\) be the average frequency of the WT allele A and \(\tilde{q}_i\) the estimated average frequency of the mutant allele a.

\[
\bar{P}_i = 1 - \tilde{q}_i.
\]

The probability that a site will be monomorphic (i.e. non-segregating) after screening \(n\) randomly chosen chromosomes is

\[
P_{i}^{\text{mon}} = (1 - \tilde{q}_i)^n.
\]

Because screening equally targets all functional categories in the targeted DNA fragment, \(n\) should be the same for different functional categories of the point mutations.

We can rewrite Eq. (3) as

\[
\tilde{q}_i = 1 - \left(1 - P_{i}^{\text{mon}}\right)^{1/n}.
\]

\(P_{i}^{\text{mon}}\) can be estimated as a fraction of monomorphic sites among all potential sites of the \(i\)th category. From Eq. (4),

\[
\tilde{q}_i = 1 - \left(1 - P_{i}^{\text{pol}}\right)^{1/n},
\]

where \(P_{i}^{\text{pol}}\) is the proportion of polymorphic sites of the \(i\)th category. Because \(P_{i}^{\text{pol}}\) was very small \(\sim 10^{-5}\) for all functional categories analyzed (Table 4), we could simplify Eq. (5) as

\[
\tilde{q}_i \approx 1 - \left(1 - \frac{P_{i}^{\text{pol}}}{n}\right) \approx \frac{P_{i}^{\text{pol}}}{n}
\]

and rewrite Eq. (1) as

\[
\frac{P_{i}^{\text{pol}}}{n} = \frac{\mu_i}{h \cdot s_i} \quad \text{or} \quad s_i = \frac{\mu_i \cdot n}{P_{i}^{\text{pol}} \cdot h}
\]

If we assume that the selection coefficient is highest against nonsense mutations among all point mutations and arbitrarily set it to a value of 1, then the relative selection coefficient \(s_{i,\text{rel}}\) for the other functional categories can be computed as

\[
s_{i,\text{rel}} = \frac{s_i}{s_{\text{non}}} = \frac{P_{i}^{\text{pol}} \cdot h \cdot n \cdot \mu_{\text{non}}}{P_{i}^{\text{pol}} \cdot n \cdot \mu_{\text{non}}} = \frac{\mu_i \cdot P_{i}^{\text{pol}}}{P_{i}^{\text{pol}} \cdot \mu_{\text{non}}}
\]

Therefore, the relative selection coefficient for point mutations of the \(i\)th category equals the ratio of the proportion of polymorphic nonsense sites to the proportion of polymorphic sites of the \(i\)th category adjusted for mutation rates. The ratio of the absolute mutation rates is the same as the ratio of the relative mutations rates; therefore, we can use the ratio \(\mu_i/\mu_{\text{non}}\) from Table 3 for Eq. (7).

### Analysis of splice mutations

To validate our hypothesis of a relatively high deleteriousness of silent mutations resulting from their effect on splicing, we estimated the average selection coefficient against splice mutations in the human genome. It is well known that the
first two nucleotides (donor site) and the last two nucleotides (acceptor site) in an intron are almost invariantly GT and AG, respectively. There are about 20,000–25,000 genes in the human genome (48) and an average of 7.8 exons per gene (52). Therefore, there are \( \approx 23,000 \cdot 6.8 \cdot 4 \) 625,600 nucleotides in the splice sites of the human genome, which suggests that there are 625,600 \( \cdot 3 \approx 1,877,000 \) potential splice mutations in the genome. We used dbSNP data to identify SNPs located in splice sites.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG Online.

**ACKNOWLEDGEMENTS**

We thank two anonymous reviewers and also A.S. Kondrashov and S. Sunyaev for discussion and critical comments. M.K. was supported in part by the NCI grant CA 75432 and C.A. was supported by ES09912 and P01 CA 34936.

**Conflict of Interest statement.** The authors declare no conflicts of interest.

**REFERENCES**

expression for the confidence interval of the estimated relative selection coefficient

\[
S_{i,\text{rel}} = \frac{\mu_{i} \cdot P_{i}^{\text{pol}}}{P_{i}^{\text{pol}} \cdot \mu_{\text{non}}}
\]

where

\[
P_{i}^{\text{pol}} = \frac{N_{i}^{\text{pol}}}{N_{\text{non}}}
\]

and \(N_{i}^{\text{pol}}\) is the observed number of polymorphic (segregating) sites for nonsense mutation, \(N_{\text{non}}\) is the total potential number of nonsense mutations in the human genome, \(N_{i}^{\text{pol}}\) is the observed number of polymorphic (segregating) sites for mutation of the \(i\)th category and \(N_{\text{non}}\) is the total potential number of mutations of the \(i\)th category in the human genome. Frequencies \(P_{i}^{\text{non}}\) and \(P_{i}^{\text{pol}}\) have distributions, which can be derived in the usual way based on the binomial distributions of the numerators (53). Therefore, the standard deviation of \(S_{i,\text{rel}}\) can be estimated on the basis of the delta method (53).

\[
\sigma(S_{i,\text{rel}}) = \frac{\mu_{\text{non}} N_{i}}{N_{\text{non}}^{\text{pol}} N_{\text{non}}^{\text{pol}} (1 - N_{\text{non}}^{\text{pol}} / N_{\text{non}})} + \frac{(N_{i}^{\text{pol}})^{2} N_{\text{pol}}^{\text{pol}} (1 - N_{i}^{\text{pol}} / N_{i})}{(N_{i}^{\text{pol}})^{4}}
\]

The 95% confidence intervals for \(S_{i,\text{rel}}\) will then have the form

\[
(S_{i,\text{rel}}) \pm 1.96 \sigma(S_{i,\text{rel}}).
\]

The approximation holds whenever the standard deviation is relatively small, which seems to be an acceptable assumption in our case.