Human SNPs Reveal No Evidence of Frequent Positive Selection

Liqing Zhang* and Wen-Hsiung Li†
*Department of Computer Science, Virginia Tech; and †Departments of Ecology and Evolution, University of Chicago

We compared the single-nucleotide polymorphisms (SNPs) in humans in 182 housekeeping and 148 tissue-specific genes. SNPs were divided into rare and common polymorphisms based on their frequencies. We found that housekeeping genes tend to be less polymorphic than tissue-specific genes for both rare and common SNPs. Using mouse as a second species for computing sequence divergences, we found no evidence of positive selection: for both housekeeping and tissue-specific genes, the ratio of nonsynonymous to synonymous common SNPs per site showed no significant difference from that of divergence. Similarly, we observed no evidence of positive selection for the 289 and 149 genes that have orthologs available for divergence calculation between humans and chimpanzees and between humans and Old World monkeys, respectively.

A comparison with previous SNP studies suggests that ~20% of the nonsynonymous SNPs in the human population are nearly neutral and that positive selection in the human genome might not be as frequent as previously thought.

Introduction

The abundance of single-nucleotide polymorphism (SNP) data in humans facilitates the inference of human population structure and detection of positive selection in the evolution of human genes (Cargill et al. 1999; Halushka et al. 1999; Fay, Wyckoff, and Wu 2001; Marth et al. 2003). Currently, the dbSNP database at the National Center for Biotechnology Information (NCBI) holds ~3.7 million nonredundant human SNPs. Studies have shown that the dbSNP database has a good coverage: ~50% to 95% of the SNPs discovered in individual studies are found in the dbSNP database (Altshuler et al. 2000; Marth et al. 2001; Reich, Gabriel, and Altshuler 2003). More recently, Jiang et al. (2003) did an extensive sampling of human SNPs in ~7,000 genes and found that the dbSNP database captures almost 60% of the SNPs with a frequency greater than 10% in their samples.

The neutral theory of molecular evolution predicts that the ratio of nonsynonymous to synonymous changes per site (A/S) should be the same for both polymorphisms and divergence (Kimura 1983). Therefore, the nature of SNPs can be inferred through the comparison of the A/S ratio of SNPs with that of species divergence data (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Templeton 1996). Because deleterious mutations can rarely reach a high frequency and because under the neutral theory advantageous mutations contribute mainly to sequence divergence but much less to polymorphisms owing to quick fixation, one can assume that most common SNPs are neutral and test for the presence of positive selection in sequence divergence between species by comparing the A/S ratio of sequence divergence with that of common SNPs. A significantly higher A/S ratio of divergence than that of common SNPs is evidence of positive selection. Based on this idea, Fay, Wyckoff, and Wu (2001) examined two SNP data sets containing 106 and 75 genes and estimated that a large fraction, 35%, of the nonsynonymous substitutions between human and the Old World monkeys have been fixed by positive selection.

In this study, we examined the SNP distribution in housekeeping and tissue-specific genes to compare the population dynamics of the two types of genes. Housekeeping genes are genes that are expressed in almost all tissues and are therefore the most basic set of genes for the physiology of an organism. Recent studies on housekeeping genes revealed that in comparison with genes that are expressed in a limited number of tissues, housekeeping genes evolve more slowly and are under stronger selective constraint (Hastings 1996; Duret and Mouchiroud 2000; Zhang and Li 2004). In this study, we tested the presence of positive selection in both types of genes. In addition, we estimated the proportions of neutral, deleterious, and beneficial mutations in the two types of genes and compared the results with previous studies.

Data and Methods

We downloaded the list of housekeeping and tissue-specific genes (a total of 905 genes) compiled by Eisenberg and Levanon (2003) based on the gene expression study of Su et al. (2002). Housekeeping genes are genes that are expressed in all the tested 47 tissues, and tissue-specific genes are genes expressed in only one or two tissues.

For the 905 genes, we searched the SNP database at the NCBI: http://www.ncbi.nlm.nih.gov/SNP/, as of June 10, 2004. We were able to find 354 and 391 SNPs for the 182 housekeeping and 148 tissue-specific genes, respectively, that have SNP frequency information for synonymous and nonsynonymous polymorphisms. The average heterozygosity of an SNP, H, (http://www.ncbi.nlm.nih.gov/SNP/Hetfreq.html) can be used to obtain the allele frequencies. For each SNP, we calculated the minor allele frequency by the formula: \[ P_{\text{minor}} = \left(1 - \frac{1}{\sqrt{4 - 2H}}\right)/2 \]. We then classified each SNP into rare polymorphism if \( P_{\text{minor}} \) is less than 15% and common polymorphism otherwise (Fay, Wyckoff, and Wu 2001). Two other cutoff values (10% and 12%) were also used, but the results did not change qualitatively.

We obtained the mouse orthologs of the 330 human genes using the procedures in Zhang and Li (2004) and calculated the synonymous and nonsynonymous divergence using a maximum likelihood method as implemented in PAML (Yang 1997). We downloaded the entire database containing all the DNA sequences from the Old World monkeys in GenBank as of February 2005 and performed TFastX (Pearson and Lipman 1988), which translated the DNA database in six reading frames and compared the
resulting sequence database with the human query protein sequences. We kept the best hits when ≥50% of the DNA sequences from the Old World monkeys were aligned to the human genes. Chimpanzee coding sequences were downloaded from http://www.ensembl.org/ (Ensembl v30.2 as of March 20, 2005). We performed Fasta on the chimpanzee database using the human genes as query sequences and required the best hits having ≥50% alignment length of the complete gene sequences. All DNA sequences were then aligned based on the corresponding protein alignment.

### Results and Discussion

No Evidence of Frequent Positive Selection on Human Genes

Table 1 shows the frequency distribution of synonymous and nonsynonymous SNPs and the sequence divergence from the mouse orthologs for the 330 genes. There is no detectable difference in the synonymous SNP density for both rare and common SNPs between the housekeeping and tissue-specific genes studied. However, the nonsynonymous SNP density in tissue-specific genes is significantly (twice) higher than that in housekeeping genes for both rare and common SNPs (0.6 SNP/kb for tissue-specific genes vs. 0.3 SNP/kb for housekeeping genes: $\chi^2 = 24.5$, df = 1, $P < 0.001$) and common SNPs (0.4 vs. 0.2 SNP/kb: $\chi^2 = 13.6$, df = 1, $P = 0.0002$).

There is no evidence of positive selection (table 1) because the A/S ratio of common SNPs is slightly higher than that of divergence for housekeeping genes (0.12 vs. 0.10) and substantially higher than that of divergence for tissue-specific genes (0.30 vs. 0.21). In contrast, Fay, Wyckoff, and Wu (2001) found that the A/S ratio of common SNPs in 181 genes is significantly lower than that of divergence from the Old World monkeys.

To be directly comparable to the study of Fay, Wyckoff, and Wu, we searched for the orthologs of the 330 genes in the Old World monkeys and were able to obtain the orthologs for 104 housekeeping genes and 45 tissue-specific genes (table 2). The A/S ratio of common SNPs is only slightly higher than that of divergence for housekeeping genes (0.12 vs. 0.11) and only slightly lower than that of divergence for tissue-specific genes (0.24 vs. 0.25); in both cases the difference is not statistically significant (for housekeeping genes: $\chi^2 = 0.05$, df = 1, $P = 0.83$; for tissue-specific genes: $\chi^2 = 0.04$, df = 1, $P = 0.84$; and for the pooled genes: $\chi^2 = 0.17$, df = 1, $P = 0.68$). Therefore, the analysis of the 149 genes using the Old World monkeys as a second species also shows no evidence of positive selection.

In addition, we used the chimpanzee as the second species for divergence calculation; we were able to obtain 289 chimpanzee orthologs (table 3). The A/S ratio of common SNPs is lower than that of divergence (0.13 vs. 0.18 for housekeeping genes and 0.33 vs. 0.37 for tissue-specific genes); however, the difference is not statistically significant (for housekeeping genes: $\chi^2 = 2.49$, df = 1, $P = 0.11$; for tissue-specific genes: $\chi^2 = 0.45$, df = 1, $P = 0.50$; and for the pooled genes: $\chi^2 = 2.17$, df = 1, $P = 0.14$). Therefore, the analysis of the 289 genes using the chimpanzee as a second species also shows no evidence of positive selection.

Table 1

<table>
<thead>
<tr>
<th>Type of Genes</th>
<th>SNP Type</th>
<th>A (nonsynonymous SNPs per site)</th>
<th>S (synonymous SNPs per site)</th>
<th>A/S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housekeeping (182 genes)</td>
<td>Rare</td>
<td>0.0003 (70/226727)</td>
<td>0.0016 (119/76132)</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>0.0002 (43/226727)</td>
<td>0.0016 (122/76132)</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Divergence (human-mouse)</td>
<td>0.0669 (15165/226727)</td>
<td>0.6658 (50692/76132)</td>
<td>0.10</td>
</tr>
<tr>
<td>Tissue-specific (148 genes)</td>
<td>Rare</td>
<td>0.0006 (116/189271)</td>
<td>0.0017 (116/69707)</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>0.0004 (72/189271)</td>
<td>0.0012 (87/69707)</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Divergence (human-mouse)</td>
<td>0.1634 (30919/189271)</td>
<td>0.7858 (54774/69707)</td>
<td>0.21</td>
</tr>
<tr>
<td>All 330 genes</td>
<td>Rare</td>
<td>0.0004 (186/415998)</td>
<td>0.0016 (235/145839)</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>0.0003 (115/415998)</td>
<td>0.0014 (209/145839)</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Divergence (human-mouse)</td>
<td>0.1108 (46084/415998)</td>
<td>0.7232 (105466/145839)</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Type of Genes</th>
<th>SNP Type</th>
<th>A (nonsynonymous SNPs per site)</th>
<th>S (synonymous SNPs per site)</th>
<th>A/S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housekeeping (104 genes)</td>
<td>Rare</td>
<td>0.0005 (34/73660)</td>
<td>0.0024 (61/25283)</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>0.0003 (19/73660)</td>
<td>0.0022 (55/25283)</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Divergence (human-OWM)</td>
<td>0.0184 (1355/73660)</td>
<td>0.1643 (4153/25283)</td>
<td>0.11</td>
</tr>
<tr>
<td>Tissue-specific (45 genes)</td>
<td>Rare</td>
<td>0.0011 (42/38253)</td>
<td>0.0022 (31/3953)</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>0.0005 (20/38253)</td>
<td>0.0022 (31/3953)</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>Divergence (human-OWM)</td>
<td>0.0258 (987/38253)</td>
<td>0.1033 (1441/3953)</td>
<td>0.25</td>
</tr>
<tr>
<td>All 149 genes</td>
<td>Rare</td>
<td>0.0007 (76/111913)</td>
<td>0.0023 (92/39236)</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Common</td>
<td>0.0003 (39/111913)</td>
<td>0.0022 (86/39236)</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Divergence (human-OWM)</td>
<td>0.0209 (2342/111913)</td>
<td>0.1426 (5594/39236)</td>
<td>0.15</td>
</tr>
</tbody>
</table>
There are several explanations for finding no evidence of frequent positive selection. First, using mouse as a second species might mask the signal of adaptive evolution due to the long divergence time. However, using either the Old World monkeys or the chimpanzee as the second species, we also found no evidence of frequent positive selection.

Second, the SNP data set could be biased toward detection of more nonsynonymous SNPs than synonymous SNPs, leading to an elevated A/S ratio of common polymorphisms. However, this is very unlikely for two reasons: (1) the coverage of the dbSNP database for SNPs with frequencies more than 15% can be as high as 95%, and more importantly, there is no bias toward nonsynonymous or synonymous SNPs (Jiang et al. 2003), and (2) the A/S ratio of common SNPs in both types of genes does not appear abnormally high, and A/S = 0.19 in the 330 genes is very close to the value of 0.20 in Fay, Wyckoff, and Wu (2001).

Therefore, it seems that the study of Fay, Wyckoff, and Wu (2001) was biased toward genes that tend to be frequently positively selected. The genes in their study have been documented to cause or have strong evidence of association with certain types of genetic disease in humans (Cargill et al. 1999; Halushka et al. 1999). We do not know why these candidate genes for particular diseases should be biased toward the presence of positive selection. At any rate, more studies are needed to estimate the prevalence of positive selection in the human genome.

Proportions of Neutral and Deleterious Mutations

Because both beneficial and deleterious mutations contribute little to common polymorphisms in the population, under the assumption that most synonymous mutations are neutral, we can use the A/S ratio of common polymorphisms to estimate the proportion of neutral nonsynonymous mutations. The proportion of neutral nonsynonymous mutations in the 330 genes is 0.19 in the current study (table 1), similar to the value of 0.2 in Fay, Wyckoff, and Wu (2001). However, our results (tables 1–3) reveal that the proportion of neutral nonsynonymous mutations is only ~0.12 in housekeeping genes but ~0.30 in tissue-specific genes. Thus, the two types of genes are, on average, subject to different intensities of selective constraint.

Taken together, the comparison between studies gives us several insights. First, the proportion of neutral nonsynonymous mutations varies greatly among genes, ranging from 12% to 30%, so that expression breadth among tissues is an important factor in determining the proportion of neutral nonsynonymous mutations. Second, the estimate of 20% as the proportion of neutral nonsynonymous polymorphisms in the human population seems reasonable because the 330 genes used in our study are from two extremes in terms of expression breadth; the estimate for 181 genes in Fay, Wyckoff, and Wu (2001) was also 20%. Third, the A/S ratio of common SNPs in tissue-specific genes is about two times that in housekeeping genes, and this ratio from polymorphism data agrees well with the divergence data (table 1; Zhang and Li 2004), suggesting that although selective constraint on these genes may fluctuate over time, the average effect remains quite constant. Fourth, the majority (~80%) of nonsynonymous mutations are removed from the population by purifying selection and contribute little to evolution and species divergence (Fay, Wyckoff, and Wu 2001; Ho et al. 2005).

Because rare polymorphisms contain mainly neutral and slightly deleterious mutations, the difference between the A/S ratios of rare and common polymorphisms approximates the proportion of deleterious nonsynonymous polymorphisms. Based on this idea, we estimated that the proportion of slightly deleterious nonsynonymous mutations (here we used the same frequency, i.e., less than 5%, to define the slightly deleterious mutations, as in Fay, Wyckoff, and Wu 2001) is 14% in housekeeping genes and 12% in tissue-specific genes, lower than the 23% estimated in Fay, Wyckoff, and Wu (2001). The difference could be partially due to the observation that deleterious nonsynonymous SNPs exist in the population in low frequencies and may thus be underrepresented in the current dbSNP database owing to the poor coverage of low-frequency SNPs (Jiang et al. 2003). However, the different estimates in the proportion of deleterious nonsynonymous polymorphisms could also reflect a real difference between the genes included in the two studies, similar to what we observe for the proportion of neutral nonsynonymous mutations.

Concluding Remarks

The distribution of A/S ratios for genes in each of the three groups (housekeeping genes, tissue-specific genes, and genes in Fay, Wyckoff, and Wu 2001) is shown in figure 1. Similar to the comparison of A/S ratios for polymorphism data, genes in Fay, Wyckoff, and Wu (2001) show an average A/S ratio for divergence from the
human–mouse comparison that lies in between the A/S ratios of housekeeping and tissue-specific genes (the Fay, Wyckoff, and Wu average A/S = 0.15 vs. housekeeping = 0.10 and tissue-specific = 0.21). Statistical tests show that, on average, housekeeping genes have significantly lower A/S ratios than tissue-specific genes (Mann-Whitney test: $P = 0$) and genes in the study of Fay, Wyckoff, and Wu (Mann-Whitney test: $P = 0$) and that genes in Fay, Wyckoff, and Wu (2001) have slightly significant lower A/S values than tissue-specific genes (Mann-Whitney test: $P = 0.048$). The observation leads us to postulate that adaptive evolution tends to occur more frequently in genes with intermediate selective constraints. Our rationale is that housekeeping genes are essential and subject to strong purifying selection, so that the proportion of advantageous mutations is low. On the other hand, although tissue-specific genes are, on average, subject to weaker selective constraint, due to their narrow expression breadth, the beneficial effect of an advantageous mutation may not confer a large selective advantage to its carriers. In comparison, genes with intermediate selective constraints have a higher chance of advantageous mutation than do housekeeping genes and a larger fitness impact than do tissue-specific genes. One can test this hypothesis by examining the population genetics of different groups of genes classified by the expression breadth.

Acknowledgments

We thank Justin Fay and Chung-I Wu for comments and Hung-Yi Wang for sharing the sequence data. The work was supported by a start-up fund to L.Z. at Virginia Tech and National Institutes of Health grants to W.-H.L.

Literature Cited


Naruya Saitou, Associate Editor

Accepted August 15, 2005