High proportions of deleterious polymorphisms in constrained human genes

Sankar Subramanian

Griffith School of Environment, Griffith University, Nathan Qld 4111, Australia

Keywords: human evolution, natural selection, deleterious polymorphisms and population genetics theory

Running head: Deleterious polymorphisms in human genomes

Abbreviation:

Title length: 72 characters (including spaces)
Abstract length: 145 words
Total length of text: 11,149 (including spaces)
Total page requirement: 3.5 pages
Number of figures: 1
Number of tables: 1
Number of references: 16

Address for correspondence:
Dr. Sankar Subramanian
Griffith School of Environment
Griffith University
170 Kessels Road
Nathan QLD 4111
Australia
Phone: +61-7-3735 7495
Fax: +61-7-3735 7459
E-mail: s.subramanian@griffith.edu.au
Abstract

Previous studies on human mitochondrial genomes showed that the ratio of intra-specific diversities at nonsynonymous-to-synonymous positions was 2-10 times higher than the ratio of inter-specific divergences at these positions, suggesting an excess of slightly deleterious nonsynonymous polymorphisms. However, such an overabundance of nonsynonymous SNPs was not found in human nuclear genomes. Here, genome-wide estimates using >14,000 human-chimp nuclear genes and one million SNPs from four human genomes showed a significant proportion of deleterious nonsynonymous SNPs (~15%). Importantly this study reveals a negative correlation between the magnitude of selection pressure and the proportion of deleterious SNPs on human genes. The proportion of deleterious amino acid replacement polymorphisms is 3.5 times higher in genes under high purifying selection compared to that in less constrained genes (28% Vs 8%). These results are explained by differences in the extent of contribution of mildly deleterious mutations to diversity and substitution.
A number of previous studies on animal mitochondrial genes reported an excess of intra-specific amino acid polymorphisms compared to fixed amino acid differences between species (Ballard and Kreitman 1994; Nachman et al. 1996; Rand and Kann 1996; Hasegawa, Cao, and Yang 1998; Wise, Srml, and Easteal 1998; Weinreich and Rand 2000; Gerber et al. 2001; Subramanian 2009). These studies compared the ratio of diversities ($\omega_r=p_N/p_S$) at nonsynonymous ($p_N$) and synonymous positions ($p_S$) with the ratio of divergences ($\omega_s=d_N/d_S$) at nonsynonymous ($d_N$) and synonymous positions ($d_S$). Under neutral evolution these two ratios are expected to be equal ($\omega_r=\omega_s$). While $\omega_r>\omega_s$ suggests an excess of amino acid polymorphisms relative to substitutions, $\omega_r<\omega_s$ indicates an excess of amino acid substitution relative to polymorphisms (Rand and Kann 1996). Studies on mitochondrial genes of human (Nachman et al. 1996; Hasegawa, Cao, and Yang 1998; Wise, Srml, and Easteal 1998; Subramanian 2009), mouse (Rand and Kann 1996) and fruit fly (Ballard and Kreitman 1994) found a much higher (2-10 times) $\omega_r$ computed within species compared to the inter-specific estimates of $\omega_s$. In contrast a previous study on nuclear genes observed a significantly higher $\omega_r$ than $\omega_s$ for only 17% of the genes while the same study found a higher $\omega_r$ for over 93% of the mitochondrial genes (Weinreich and Rand 2000). Notably large-scale comparative analysis of human-chimp nuclear genomes showed a similarity between $\omega_r$ (0.21-0.23) obtained from human population data and $\omega_s$ estimated using the inter-specific data (0.23) (Mikkelsen et al. 2005). The fundamental reason for this discrepancy between mitochondrial and nuclear genes is not known.

Population genetic theories predict that weakly deleterious mutations contribute more to diversity than to substitution (Kimura 1983). Since the majority of amino acid changing mutations
occurring in genes under high purifying selection are deleterious this leads to higher $\omega_R$ than $\omega_S$ as most of the deleterious mutations are eventually eliminated. It is well known that mitochondrial genes are highly constrained, which is evident from their very low $\omega_S$ [-0.04 for human-chimp comparison, (Subramanian 2009)]. Hence the observed higher $\omega_R$ in mitochondrial genes appears to be due to the presence of weakly deleterious mutations, which is expected under standard population genetics theories (Kimura 1983; Kryazhimskiy and Plotkin 2008). However nuclear genes are evolving under varying levels of selection constraint. Therefore to test the above prediction for nuclear genes it is important to split them into several groups based on the magnitude of selection constraint on them and examine the patterns of polymorphisms in each set of genes.

Firstly, to examine the genome-wide pattern I assembled a dataset of 14,939 human-chimpanzee orthologous genes and obtained over one million SNPs from four completely sequenced genomes belonging to two Europeans, a Chinese and an African. Evolutionary divergence between human and chimpanzee at nonsynonymous, synonymous, and intron positions were estimated. Nucleotide diversities at these positions were also estimated using the SNP data. To estimate $\omega_R$ and $\omega_S$, divergences at synonymous sites and introns were used as proxies for neutral evolution. The genomic average estimates revealed that $\omega_R$ is significantly higher than $\omega_S$ and this difference is 0.16% - 0.18% (Table 1).

The ratio $\omega_R$ is the fraction of nonsynonymous mutations that is present in the population and $\omega_S$ is the proportion of replacement mutations that is fixed. If we assume that all neutral (and adaptive) nonsynonymous mutations are fixed then the excess nonsynonymous SNPs (compared
to replacement substitutions) are expected to be deleterious in nature. This proportion of deleterious replacement SNPs ($\delta$) can be estimated by $\delta = \frac{\omega_p - \omega_s}{\omega_p}$, which can be simplified as $\delta = 1 - \frac{d_{SP}}{d_{SN}}$. For the human nuclear genome $\delta$ was estimated to be 13%-15% (Table 1).

Next, in order to categorize human genes based on the magnitude of selection pressure on them, I used the $d_{SN}/d_S$ ratio estimated for the orthologous mouse-rat pair ($\omega_{S,Rodent}$). This is based on the assumption that on an average the magnitude of selective constraints on rodent genes reflects the extent of constraints on their orthologous hominid counterparts. By using rodent $\omega_S$ instead of the $\omega_S$ of human-chimp pair as a proxy for the magnitude of selection constraint on human genes, makes the variables $\delta$ and selection constraint ($\omega_{S,Rodent}$) statistically independent. For 11,020 human genes the orthologous mouse-rat sequences were obtained. Human genes were sorted based on the $\omega_{S,Rodent}$ estimated for the orthologous mouse-rat genes and were grouped into 20 groups with an equal number of genes in each group (551). For each group of genes, mean estimates of $\omega_F$ and $\omega_S$ were computed. Figure 1A shows that $\omega_F$ and $\omega_S$ of human genes positively correlate and the genes with a low $\omega_F$ also have a low $\omega_S$. However the relationship is not one-to-one. For highly constrained human genes ($\omega_{S,Rodent} = 0.002$) $\omega_F$ is $\sim 39\%$ higher than $\omega_S$. In contrast $\omega_F$ of less constrained human genes ($\omega_{S,Rodent} = 0.61$) is only 8% higher than the corresponding $\omega_S$. Then the fraction of deleterious replacement polymorphisms ($\delta$) was computed for each set of human genes. Figure 1B reveals a highly significant negative relationship (Spearman’s coefficient $\rho = -0.85$, $P = 0.0002$) between $\delta$ and selective constraint.
(\omega_S\text{-Rodent}). The proportion of deleterious replacement SNPs is 28.3% in highly constrained human genes whereas this fraction is only 7.7% in less constrained genes. Therefore \delta is 3.7 times higher in genes under high purifying selection compared to the genes under relaxed selective constraints. In the above analysis synonymous sites were used to estimate \omega_S and \omega_P, however similar highly significant relationship was obtained when intron sites were used (\rho = -0.69, P = 0.0028).

These results could be explained by the theoretical predictions of Kimura (1983), who demonstrated that when Nes (Ne and s are effective population size and selection coefficient respectively) increases, the contribution of weakly deleterious mutations to fixation declines much faster than their contribution to diversity (Kimura, 1983, figure 3.7). Kryazhimskiy and Plotkin (2008) illustrated that the difference between \omega_P and \omega_S increases when Nes deviates from 0 [Figure 1 of (Kryazhimskiy and Plotkin 2008)]. For instance, under high purifying selection (e.g. 2Nes =-2) \omega_S is very small (0.07), but \omega_P (0.47-0.52) is 6-7 times higher than \omega_S (calculated from the numerical data used to produce this figure). Conversely under relaxed selective constraint (e.g. 2Nes =-0.5) \omega_S is rather large (0.58), but the difference between \omega_P (0.84-0.87) and \omega_S (0.58) is small (only 40-50%).

Slightly deleterious SNPs at low frequencies behave like neutral SNPs and are strongly influenced by random sampling drift and thus they contribute significantly to diversity (Kimura 1983). However they are selected against at high frequencies and eventually prevented from being fixed and thus do not contribute to substitution. Since a high fraction of nonsynonymous
mutations in highly constrained genes are deleterious they are ephemerally maintained by drift and purged over time, which results in higher $\omega_P$ than $\omega_S$. In contrast most of the replacement SNPs in genes under relaxed constraints are largely neutral and become fixed (as evidenced from their high $\omega_S$), which leads to more similarity between $\omega_P$ and $\omega_S$. Hence the negative relationship between the fraction of deleterious mutations and selection intensity observed in this study is expected under these theoretical predictions.

These results and the theoretical predictions also explain the excess nonsynonymous SNPs reported in human mitochondrial genomes. Since mitochondrial genes are also under high selective constraints ($\omega_S = 0.04$), most of the replacement mutations in these genes are likely to be deleterious in nature. However the range of $\delta$ estimated for the mitochondrial genomes is 50%-90% (based on $\omega_P/\omega_S$ ratios of 2-10 respectively), which is much higher than that estimated for highly constrained nuclear genes (28%). Interestingly the effective population size of mitochondrial genome is known to be four fold smaller than that of nuclear genome and thus mtDNA mutations are much more influenced by genetic drift. Hence the low $\omega_S$ and high $\delta$ suggest a much higher selection coefficient ($s$) on mitogenes to compensate for the reduced $Ne$.

**Methods**

Alignments of human-chimpanzee genomes were obtained from the UCSC genome browser ([http://genome.ucsc.edu/](http://genome.ucsc.edu/)). Using gene location and intron-exon boundary information, alignments for individual genes were extracted. Only the genes that have intron(s) were included. Single nucleotide polymorphism (SNP) data for the complete genomes of two Europeans (Levy et al. 2007; Wheeler et al. 2008), a Chinese (Wang et al. 2008) and an African
(Bentley et al. 2008) were obtained from the corresponding repositories mentioned in the publications. Since all these SNP data were mapped on to the reference human genome build 36 it was straightforward to find the variants from all four genomes. For each gene, divergence between human-chimp at nonsynonymous, synonymous, and intron sites were estimated using the Jukes-Cantor method and the number of synonymous and nonsynonymous positions were obtained using PAML (Yang 2007). Similar procedure was used to compute the nucleotide diversity (π) at these positions between the four human genomes. To reduce the variance all or a set of genes were concatenated to compute average estimates. Protein-coding gene sequences of mouse and rat were downloaded from the GenBank. Reciprocal BLAST search was used to identify orthologs of human, mouse and rat and genic dN/dS estimates for the mouse and rat pair were estimated using PAML.

**Acknowledgment**

The author is grateful to David Lambert and acknowledges the support from AISRF grant and Griffith University. I thank John McDonald for pointing out a confounding methodological artifact. Also I thank Joshua Plotkin for providing the numerical data and Leon Huynen, Konstantin Popadin and an anonymous reviewer for critical comments.
**Figure Legends**

**Figure 1.** (A). Mean estimates of $\omega_s$ (grey) and $\omega_p$ (black). Human genes were sorted based on the $d_{SNP}/d_S$ ($\omega_{S-Rodent}$) of the orthologous genes of mouse-rat pair and then they were split into 20 groups, each group with an equal number of genes (551). The X-axis shows the average $\omega_s$ estimated for the mouse-rat pair ($\omega_{S-Rodent}$). Average estimates of $\omega_p$ (humans) and $\omega_s$ (human-chimp) were obtained for each set of human genes are shown in the Y-axis.

(B) Relationship between the magnitude of selective constraint on human genes and the proportion of deleterious nonsynonymous SNPs ($\delta$). The mean $\omega_{S-Rodent}$ of rodents was used as a proxy to quantify the relative selection intensity on human genes. The correlation is highly significant (Spearman’s coefficient $\rho = -0.85$, $P = 0.0002$). The best-fitting linear regression line is shown.
Table 1. Genomic estimates of divergence, diversity and their ratios

<table>
<thead>
<tr>
<th>Types of sites/ratio</th>
<th>Number of Polymorphisms/Substitutions</th>
<th>Divergence*/Diversity*/Ratio (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsynonymous divergence ($d_N$)</td>
<td>66037</td>
<td>0.00344 (1.3 × 10^{-5})</td>
</tr>
<tr>
<td>Synonymous divergence ($d_S$)</td>
<td>92453</td>
<td>0.01322 (4.3 × 10^{-5})</td>
</tr>
<tr>
<td>Inronic divergence ($d_I$)</td>
<td>8391989</td>
<td>0.01217 (4.1 × 10^{-6})</td>
</tr>
<tr>
<td>Nonsynonymous diversity ($p_N$)</td>
<td>6665</td>
<td>0.00035 (4.2 × 10^{-6})</td>
</tr>
<tr>
<td>Synonymous diversity ($p_S$)</td>
<td>7984</td>
<td>0.00113 (1.3 × 10^{-5})</td>
</tr>
<tr>
<td>Inronic diversity ($p_I$)</td>
<td>988726</td>
<td>0.00106 (1.1 × 10^{-6})</td>
</tr>
</tbody>
</table>

$\omega_S = d_S/d_S$ 0.260 (0.0013)

$\omega_S = d_S/d_I$ 0.283 (0.0011)

$\omega_F = p_N/p_S$ 0.306 (0.0051)

$\omega_F = p_N/p_I$ 0.327 (0.0040)

$\omega_F/\omega_S$ 1.18, $P < 0.0001$

$\omega_F/\omega_I$ 1.16, $P < 0.0001$

Proportion of deleterious replacement SNPs ($\delta$ based on $d_S$) 15%

Proportion of deleterious replacement SNPs ($\delta$ based on $d_I$) 13%

* - substitutions per site
References


Figure 1

A

Mean $\omega_{S-Rodent}$ (Mouse-Rat)

Mean $\omega_{P or S}$ (Human-Chimp)

B

Proportion of deleterious SNPs ($\delta$)

Mean $\omega_{S-Rodent}$ (Mouse-Rat)