Selective Constraints Determine the Time Dependency of Molecular Rates for Human Nuclear Genomes

Sankar Subramanian* and David M. Lambert*

Environmental Futures Centre, Griffith University, Nathan, Queensland, Australia

*Corresponding authors: E-mail: s.subramanian@griffith.edu.au; d.lambert@griffith.edu.au.

Accepted: October 6, 2012

Abstract

In contrast to molecular rates for neutral mitochondrial sequences, rates for constrained sites (including nonsynonymous sites, D-loop, and RNA) in the mitochondrial genome are known to vary with the timeframe used for their estimation. Here, we examined this issue for the nuclear genomes using single-nucleotide polymorphisms (SNPs) from six complete human genomes of individuals belonging to different populations. We observed a strong time-dependent distribution of nonsynonymous SNPs (nSNPs) in highly constrained genes. Typically, the proportion of young nSNPs specific to a single population was found to be up to three times higher than that of the ancient nSNPs shared between diverse human populations. In contrast, this trend disappeared, and a uniform distribution of young and old nSNPs was observed in genes under relaxed selective constraints. This suggests that because mutations in constrained genes are highly deleterious, they are removed over time, resulting in a relative overabundance of young nSNPs. In contrast, mutations in genes under relaxed constraints are nearly neutral, which leads to similar proportions of young and old SNPs. These results could be useful to researchers aiming to select appropriate genes or genomic regions for estimating evolutionary rates and species or population divergence times.

Key words: rates of evolution, natural selection, time dependency, deleterious polymorphisms, population genetic theory.

Introduction

The rate of molecular evolution is a fundamental parameter in genetics and evolutionary biology. Although rates that are estimated using different timescales and methods are expected to be similar, recently studies suggested otherwise. For example, studies based on pedigree analyses have recorded higher rates of evolution compared with those estimated using phylogenetics methods (Parsons et al. 1997; Howell et al. 2003; Denver et al. 2004; Haag-Liautard et al. 2007; Millar et al. 2008). Comparative genomic studies have shown that molecular rates estimated using short timescales, for instance, those based on intraspecific data, are much higher than those estimated using interspecies data (Garcia-Moreno 2004; Ho et al. 2005, 2007; Burridge et al. 2008; Subramanian and Lambert 2011). A number of reasons have been suggested for this discrepancy. However, most of these involve biases or errors in estimation such as calibration errors, saturation effects on nucleotide positions, and phylogenetic and demographic model misspecification (Emerson 2007; Bandelt 2008; Debruyne and Poinar 2009; Henn et al. 2009; Ho et al. 2011). The higher intraspecific rates have also been attributed to artifacts such as sequencing errors, postmortem damage (in the case of ancient DNA), and ascertainment bias (reviewed by Ho et al. [2011]).

The major biological factor that produces a time-dependent pattern of molecular rates appears to be purifying selection (Endicott and Ho 2008; Subramanian et al. 2009; Ho et al. 2011; Subramanian and Lambert 2011). At short timescales, slightly deleterious polymorphisms will segregate in populations, and this standing variation results in a higher diversity. This is evident from a number of recent studies that have shown an overabundance of low-frequency nonsynonymous single-nucleotide polymorphism (nSNPs) in human protein-coding genes (Li et al. 2010; Nelson et al. 2012; Subramanian 2012; Tennessen et al. 2012). Thus, an elevated molecular rate is observed for the data from within a population or for closely related populations (Kimura 1983). In contrast, deleterious polymorphisms are removed over time, which results in reduced divergence between species or distantly related populations.

Note that throughout this article, the rate of evolution refers to the observed or estimated divergence divided by calibration time. It does not denote a mutation rate, and hence, we do not intend to suggest that mutation rate varies with time.
Time-dependent rates have been typically reported for the D-loop region of the mitochondrial genome (Parsons et al. 1997; Lambert et al. 2002; Howell et al. 2003; Ho et al. 2005; Hay et al. 2008; Henn et al. 2009). Because the D-loop is hypervariable, the saturation effects on nucleotide sites confound the effects of purifying selection in this region. Hence, later studies (Endicott and Ho 2008; Subramanian et al. 2009; Subramanian and Lambert 2011) used coding genes of mitochondria to examine the time-dependency effect, because these are less prone to estimation errors. These studies showed that rates of molecular evolution at synonymous sites of protein-coding genes are similar across different timescales. In contrast, these studies found a strong time-dependent rate of evolution at nonsynonymous sites. Because synonymous sites are free from selection and nonsynonymous positions are under selective constraints, the observed pattern clearly points to the effects of purifying selection as predicted theoretically (Kimura 1983). This is also evident from the time-dependent decline in the ratio of divergences/diversities at nonsynonymous and synonymous sites (dN/dS or pN/pS) for protein-coding genes of human mitochondrial genes (Subramanian 2009) and nuclear genes of virus (Holmes 2003) and bacteria (Rocha et al. 2006).

Studies on the nuclear genomes of vertebrates did not reveal a clear-cut pattern of time dependency. Although earlier studies suggested a time-dependent pattern in microsatellites (Zhivotovsky et al. 2004, 2006) of the human nuclear genome, later studies based on next-generation sequencing did not identify such a pattern (Xue et al. 2009; Sun et al. 2012). This was further confirmed by a number of genome-wide studies, which showed that the evolutionary rates estimated using complete nuclear genomes of human pedigrees were similar to those estimated using human–chimpanzee species comparison (Conrad et al. 2010; Roach et al. 2010; Kong et al. 2012). However, as discussed earlier, time-dependent variation is expected only for constrained genomic regions. Therefore, in this study, we examined protein-coding genes in the human nuclear genome. It is well known that different nuclear genes are under different levels of selective constraint, depending on the relative importance of their functions. Hence, it would be interesting to examine the time-dependent pattern of evolutionary rates in genes, under various magnitudes of selection pressure. Therefore, we grouped nuclear genes of the human genome into four categories based on the intensities of selection pressure on them. We examined the pattern of time dependency using synonymous SNPs (sSNPs) and nSNPs from the complete genomes of six humans belonging to different populations.

Materials and Methods

DNA Sequence and Polymorphism Data

Protein-coding sequence alignments for humans (build 36) and chimpanzee (build 2) were obtained (for 16,750 known genes) from the University of California–Santa Cruz (UCSC) genome bioinformatics (http://genome.ucsc.edu/). Polymorphism data from six complete genomes belonging to a Khoisan (Schuster et al. 2010), two Yorubans (Bentley et al. 2008; Schuster et al. 2010), a European (Levy et al. 2007), a Chinese (Wang et al. 2008), and a Korean person (Kim et al. 2009) were obtained from UCSC genome bioinformatics and PSU Bioinformatics (http://main-genome-browser.bx.psu.edu/) data repositories. sSNPs and nSNPs of these genomes were determined using the chromosomal coordinates and gene boundary information. The number of sSNPs and nSNPs from each genome is given in table 1. The reference human genome was used to determine the ancestral state of each SNP. However, using the chimpanzee genome to orient the direction of SNP produced similar results (data not shown).

Results and Discussion

To examine nucleotide diversity at various temporal or evolutionary depths, we obtained SNP data from five complete human genomes belonging to a Chinese, a Korean, an European, a Yoruban (West African), and a Khoisan, an ancient African lineage. The phylogenetic relationship between the genomes of these humans (Tishkoff et al. 2009) is shown in figure 1. sSNPs and nSNPs were grouped based on the pattern of sharing, as described in the results (Subramanian 2012). To estimate the proportion of differences (pS or pA), the number of sSNPs or nSNPs belonging to each group was divided by the total number of synonymous or nonsynonymous positions in the genome. The binomial variance was used to compute the standard error.

Divergence and Diversity Estimation

The numbers of synonymous and nonsynonymous positions and substitutions in each gene were calculated using the codeml program of the software PAML (Yang 2007). Pairwise evolutionary distances at synonymous and nonsynonymous sites for the human–chimpanzee pair were estimated using the Jukes–Cantor method. sSNPs and nSNPs were grouped based on the pattern of sharing, as described in the results (Subramanian 2012). To estimate the proportion of differences (pS or pA), the number of sSNPs or nSNPs belonging to each group was divided by the total number of synonymous or nonsynonymous positions in the genome. The binomial variance was used to compute the standard error.
Table 1
Number of SNPs and Estimates of Ratios for each Branch of the Human Tree

<table>
<thead>
<tr>
<th>Branch on the Tree</th>
<th>Nonsynonymous SNPs (A)</th>
<th>Synonymous SNPs (S)</th>
<th>A/S</th>
<th>pN/pS (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dN/dS = 0–0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAECK</td>
<td>423</td>
<td>2,309</td>
<td>0.18</td>
<td>0.067 (0.0035)</td>
</tr>
<tr>
<td>AECX</td>
<td>137</td>
<td>534</td>
<td>0.26</td>
<td>0.093 (0.0089)</td>
</tr>
<tr>
<td>ECK</td>
<td>122</td>
<td>371</td>
<td>0.33</td>
<td>0.120 (0.0125)</td>
</tr>
<tr>
<td>CK</td>
<td>558</td>
<td>1,049</td>
<td>0.53</td>
<td>0.194 (0.0101)</td>
</tr>
<tr>
<td>dN/dS = 0.1–0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAECK</td>
<td>721</td>
<td>1,552</td>
<td>0.46</td>
<td>0.169 (0.0076)</td>
</tr>
<tr>
<td>AECX</td>
<td>168</td>
<td>327</td>
<td>0.51</td>
<td>0.187 (0.0177)</td>
</tr>
<tr>
<td>ECK</td>
<td>169</td>
<td>235</td>
<td>0.72</td>
<td>0.262 (0.0264)</td>
</tr>
<tr>
<td>CK</td>
<td>571</td>
<td>702</td>
<td>0.81</td>
<td>0.296 (0.0167)</td>
</tr>
<tr>
<td>dN/dS = 0.2–0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAECK</td>
<td>2,406</td>
<td>2,695</td>
<td>0.89</td>
<td>0.325 (0.0091)</td>
</tr>
<tr>
<td>AECX</td>
<td>626</td>
<td>636</td>
<td>0.98</td>
<td>0.358 (0.0202)</td>
</tr>
<tr>
<td>ECK</td>
<td>461</td>
<td>434</td>
<td>1.06</td>
<td>0.387 (0.0259)</td>
</tr>
<tr>
<td>CK</td>
<td>1,571</td>
<td>1,188</td>
<td>1.32</td>
<td>0.481 (0.0185)</td>
</tr>
<tr>
<td>dN/dS = 0.6–1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAECK</td>
<td>875</td>
<td>571</td>
<td>1.53</td>
<td>0.558 (0.0300)</td>
</tr>
<tr>
<td>AECX</td>
<td>244</td>
<td>165</td>
<td>1.48</td>
<td>0.538 (0.0542)</td>
</tr>
<tr>
<td>ECK</td>
<td>160</td>
<td>96</td>
<td>1.67</td>
<td>0.607 (0.0783)</td>
</tr>
<tr>
<td>CK</td>
<td>506</td>
<td>286</td>
<td>1.77</td>
<td>0.644 (0.0476)</td>
</tr>
</tbody>
</table>

African–Eurasian, and Khoisan–other humans, respectively (Gronau et al. 2011). Hence, these times provide relative ages for the SNPs as <36.5, 36.5–51.0, 51.1–132.5, and >132.5 for CK, ECK, AECX, and SAECK, respectively.

We then grouped human genes based on the level of selective constraint on them. To quantify this, we used dN/dS ratios estimated from the human–chimp divergence at nonsynonymous- (dN) and synonymous sites (ds). Because synonymous sites are free from selection, this ratio reveals the extent of selective constraint on amino acids. For each group of genes, we estimated the ratio (pN/pS) of nSNPs (pN) to sSNPs (pS) per site using the polymorphisms shared between different branches of the human population tree (table 1). First, we examined the distribution of SNPs in the highly constrained human genes. This revealed a clear negative relationship between the age of SNPs (based on the extent of sharing) and pN/pS ratios (fig. 2A). For example, the pN/pS ratio of the Asian-specific SNPs (CK) was 2.9 times higher than that estimated using the SNPs shared with the ancient Khoisan genome (SAECK). This is a perfect example for time dependency, as the latter (>132.5 Kyr) is more than three times older than the former (<36.5 Kyr).

Interestingly, the distribution of pN/pS of genes gradually changes with decreasing selection pressures (fig. 2A–D). The distribution observed for highly constrained genes was strongly right skewed (fig. 2A). In contrast, it is largely uniform for the genes under relaxed selective constraints (fig. 2D) where the difference between pN/pS ratios of young and old SNPs disappeared (Z test, P < 0.12).

In addition to using dN/dS ratios, we also grouped human genes based on a different measure namely the Genome Evolutionary Rate Profiling (GERP) score described previously (Cooper et al. 2005). A GERP score is determined using multiple sequence alignments of orthologous genes and the level of selective constraints on each position. The score is quantified using the deficit of substitutions compared with a neutral site. We estimated the average GERP score for each gene and grouped genes based on their mean GERP scores. The results based on average GERP score (supplementary fig. S1, Supplementary Material online) were very similar to those showed in figure 2. For instance, constrained genes with a mean GERP score > 3.5 showed a clear-cut pattern of time dependency, which was absent for genes under relaxed selective constraints (average GERP score < 0.5).

Theoretical studies suggested that deleterious mutations contribute to the diversity of a population, but they are prevented from reaching higher frequencies and are purged over time (Kimura 1983). The results of this study strongly support this prediction. Because mutations in constrained genes are highly deleterious in nature, they do not typically spread through populations. As they are removed over time, we observe a very strong time-dependent effect on these genes (fig. 2A). In contrast, mutations on genes under relaxed selective constraints are typically only mildly harmful to humans and, therefore, segregate in human population for a long time. Particularly mutations in the most rapidly evolving genes with a dN/dS > 0.6 (fig. 2D) are almost neutral with negligible effects on the fitness of humans. This results in the similarity of pN/pS ratios across various timescales.

Previous studies suggested a population bottleneck in European and Asian populations (Marth et al. 2004; Li and Durbin 2011), and thus, a much smaller effective population size is expected for non-African populations than Yoruban
and Khoisan populations. Hence, this bottleneck effect might have drifted some deleterious SNPs to higher frequencies, and this might have some influence on the pattern observed for constrained genes (fig. 2A). To address this issue, we examined the pattern of nSNPs in the African lineage, which is not bottlenecked in the past (Gronau et al. 2011; Li and Durbin 2011). For this purpose, we included another Yoruban genome and examined $p_N/p_S$ ratios for the branches of the human population tree constructed using the genomes from two Yorubans, a European, a Chinese, and a Khoisan (fig. 3). For constrained human genes ($dN/dS < 0.1$), we found a significantly higher $p_N/p_S$ ratio for Yoruban-specific SNPs (0.15) than that estimated using the SNPs shared between Yorubans and Eurasians (0.09), and this is higher than that computed for those shared between the ancient Khoisan and other genomes (0.06) (fig. 3B). In contrast, for genes under weak purifying selection ($dN/dS > 0.6$), $p_N/p_S$ ratios of old and young SNPs were similar ($P = 0.44$) (fig. 3E). Figure 3B clearly supports a time-dependent pattern of deleterious SNPs rather than any bottleneck effect in the African lineage. Furthermore, a previous study suggested that the common ancestral population of all humans (SECAF) was smaller than the population ancestral to Yorubans (or Africans) and Eurasians (ECAF), which was in turn smaller than the population of Yorubans (AF) (Gronau et al. 2011). Hence, it seems likely that population expansion occurred throughout the African lineage. If population size is assumed to modulate the $p_N/p_S$ ratios, then a pattern opposite to that shown in figure 3 is expected: high $p_N/p_S$ for SECAF and low for AF.

Therefore, population size effect might not explain our results shown in figures 2 and 3.

This study demonstrated how purifying selection on genes influences the pattern of time dependency of molecular rates. Previous studies on human mitochondrial genes also suggested enrichment of nonsynonymous polymorphisms in the tips of the human population tree, compared with the internal nodes (Kivisild et al. 2006; Subramanian 2009; Pereira et al. 2011). Furthermore, Pereira et al. (2011) showed a predominance of pathogenic mutations in the younger branches of the human tree. Previous studies based on pedigree analysis of D-loop regions reported much higher rates compared with phylogenetic studies, which suggests the presence of deleterious mutations (Parsons et al. 1997; Howell et al. 2003; Millar et al. 2008). In contrast, pedigree analyses based on complete nuclear genomes (Conrad et al. 2010; Roach et al. 2010; Kong et al. 2012; Sun et al. 2012) of humans showed that the rates obtained from these analyses were not significantly different from the interspecies rate estimated using pseudogene data from human and chimp comparison (Nachman and Crowell 2000). Because the majority (>95%) of the sites of human genome are under neutral evolution, the similarity of the pedigree-based rate with the interspecific rate appears to be due to the predominance of neutral mutations.

In this study, we showed the variations in the rate of evolution observed between different timescales could largely be attributed to purifying selection. However, some of the variations could be due to other biological factors such as changes in population sizes. For instance, if the population size ($N_e$) of...
A species has been declining over time, this might mimic the pattern of time dependency as \( pN/pS \) ratios could be modulated by \( N_e \). However, time dependency has been observed in a large number of species from virus to vertebrates, which suggests that this could be a universal phenomenon. This seems a reasonable view because we do not expect a population decline in the large number of species examined to date. Although multiple substitutions, homoplasy, or back mutations are of biological origin, they might be missed particularly in mutational hotspots because of the lack of methods to identify them. Therefore, long-term divergences are subjected to a higher rate of underestimation than short-term distances, and this might also produce a time-dependent effect. However, we used nuclear genomes of human for which the evolutionary rate is very low (<1 per site per billion years [Conrad et al. 2010; Roach et al. 2010; Kong et al. 2012; Sun et al. 2012]), and therefore multiple, back, or parallel substitutions are unlikely to influence our results as our comparisons involve only closely related species and populations.

Our results suggest that although the observed rates of molecular evolution vary for different timescales, this variation is limited to genes or genomic regions under selection. Importantly, we showed that the extent of variation is determined by the magnitude of selection on these regions. Therefore, to estimate molecular evolutionary rates or divergence times between species/populations, it is advisable to use the genes and genomic regions that are under minimal selective constraints.

**Fig. 3.** (A) Phylogeny of human genomes belonging to different populations. Abbreviations are SNPs specific to Yorubans (AF), those shared between Yorubans and Eurasians (ECAF), and Khoisan and other genomes (SECAF). \( pN/pS \) ratios were estimated for three branches at different depths of the tree. Human genes were grouped into four categories mentioned in figure 2. Error bars denote the standard error of the mean. In (E), the differences in \( pN/pS \) ratios between any two categories were not statistically significant (Z test, \( P > 0.20 \)).
Subramanian and Lambert

Supplementary Material

Supplementary figure S1 is available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgment

This work was supported by the Australian Research Council and Griffith University.

Literature Cited


Associate editor: George Zhang