Note

February 3, 2011

The abundance of deleterious polymorphisms in humans

Sankar Subramanian

Griffith School of Environment, Griffith University, 170 Kessels Road, Nathan Qld 4111, Australia
Running head: Temporal pattern of deleterious polymorphisms

Keywords: human evolution, deleterious mutations, disease-associated mutations, genome-wide association and population genetics theory

Address for correspondence:
Dr. Sankar Subramanian
Environmental Futures Centre
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

Here I show a gradual decline in the proportion of deleterious nonsynonymous SNPs (nSNPs) from tip to root of the human population tree. This study reveals that up to 48% of nSNPs specific to a single genome are deleterious in nature, which underscores the abundance of deleterious polymorphisms in humans.
The neutral theory of molecular evolution predicts that amino acid replacement polymorphisms are neutral and largely governed by mutation and genetic drift (Kimura and Ohta 1971). However, Ohta (1973) suggested that some amino acid polymorphisms could be slightly deleterious and that they are modulated by purifying selection and drift. According to her prediction, rare amino acid polymorphisms are, on average slightly deleterious and the number of these variants in a population is generally more than that expected under the strict neutral hypothesis (Ohta 1973). Later, Kimura (1983) proposed that deleterious mutations can contribute significantly to diversity, but at high frequencies they are selected against and prevented from becoming fixed. These predictions were examined in a number of studies using RFLP (Excoffier 1990; Merriwether et al. 1991; Whittam et al. 1986) or DNA sequence data (Tajima 1989), which indeed showed an excess in the number of rare variants than that expected using Watterson’s test (Watterson 1978) and/or Tajima’s test (Tajima 1989).

Using a different approach, studies on mitochondrial genes compared the diversity and divergence at synonymous and nonsynonymous sites and showed that the ratio of intra-specific diversities at nonsynonymous- to synonymous positions ($\omega=dN/dS$) was significantly higher than the ratio of inter-specific divergence at these sites (Ballard and Kreitman 1994; Nachman et al. 1994; Rand et al. 1994). This suggests an excess of nonsynonymous polymorphisms (nSNPs) in populations compared to fixed replacement mutations. The excess fraction of nSNPs could be enriched with rare alleles, which are potentially deleterious and natural selection eliminates them from the population over time. This pattern appears to be universal as it was observed by a number of studies on mitochondrial genes of human (Hasegawa et al. 1998; Nachman et al. 1996; Rand and Kann 1996), mouse (Nachman et al. 1994; Rand and Kann 1996), fruit fly
(BALLARD and KREITMAN 1994; RAND et al. 1994) as well as on nuclear genes of bacteria (ROCHA et al. 2006) and viruses (HOLMES 2003). Using a number of nuclear loci, Hughes et al. (2003; 2005) showed a reduced diversity at nonsynonymous sites of human genes suggesting purifying selection, which in turn, implies that slightly deleterious nSNPs are prevalent in human populations. Large-scale genome-wide studies on mitochondrial (KIVISILD et al. 2006; RUIZ-PESINI et al. 2004; SUBRAMANIAN 2009) and nuclear genes (BUSTAMANTE et al. 2005; CARGILL et al. 1999; WONG et al. 2003) also suggested the presence of low-frequency deleterious amino acid SNPs in humans. Using allele frequency spectrum of synonymous and nonsynonymous sites, several studies estimated the strength of selection on amino acid mutations and thus determined the proportion of deleterious amino acid mutations in humans (BOYKO et al. 2008; EYRE-WALKER et al. 2006).

However, it is important to estimate the fraction of deleterious nSNPs at different depths of the human population tree. This is because a much higher fraction of deleterious nSNPs is expected if individuals from the same population (or between closely related populations) are compared, as opposed to comparisons involving genomes from distantly related populations. For instance, we anticipate a relatively high fraction of deleterious SNPs among the polymorphisms shared between two Europeans than those shared between a European and an Asian or a European and an African. This is evident from a previous study that showed a significantly less fraction of deleterious nSNPs that are shared between African-Americans and European-Americans than those present in only one of the population (LOHMUELLER et al. 2008). This is because the fraction of deleterious SNPs is expected to be inversely proportional to the time of separation between populations as they are removed by purifying selection over time. Hence I examined
the temporal pattern of deleterious nSNPs at all internal and terminal branches of the human population tree using complete genomes belonging to four distinct human populations. To quantify the proportion of deleterious polymorphisms, two independent methods were used.

Polymorphism data from ten complete human genomes belonging to Europeans, Asians, Africans and a Khoisan (an ancient African lineage) were obtained from different databanks (see, supporting information for materials and methods). Synonymous (sSNPs) and nonsynonymous SNPs (nSNPs) were grouped based on the pattern of sharing between the genomes following the known phylogenetic relationship between the human populations (Tishkoff et al. 2009). If a SNP is shared between a Khoisan and any other genome, it was considered to be the oldest (see branch A in figure 1). If a SNP is shared between a European and an African and not shared by Khoisan, it is considered to be common to Africans and non-Africans (branch B). Similarly, if a SNP is shared between a European and an Asian but not shared by an African or Khoisan, it is considered to be ancestral to Eurasians (branch C). Likewise, if a SNP is shared only between any two Europeans and not shared with any other genome, it was considered to be ancestral to Europeans (branch D). Finally, if a SNP is present in only one genome and not shared with any other genome, it is considered to be specific to that individual (all terminal branches such as E). These grouping assumed that convergent or parallel mutations are rare.

Typically, nSNPs are expected to be more abundant at terminal branches than internal branches of the tree as a fraction of them are removed over time. In contrast, sSNPs are expected to be present at roughly equal proportions at all nodes of the tree. Therefore, the ratio of nSNPs (A) to sSNPs (S) will normalize the time elapsed and therefore this ratio could be used to estimate the
excess fraction of nSNPs in each branch at various depths of the tree. Table 1 shows that A/S ratios are highest among unshared genome specific SNPs (branches E, G, I and J) and lowest among those SNPs that are shared with Khoisan genome (branch A). I then estimated the ratio of diversity at nonsynonymous to synonymous sites ($\omega$) for each branch. This measure is much clearer, as it is on a scale of 0-1. Figure 1 shows a steady decline in $\omega$ from the tip to the root of the tree with the $\omega$ estimate for the terminal branches (0.34-0.48) being up to twofold higher than that of the ancestral branch (0.26). The estimate of $\omega$ was smallest for the inter-species human-chimp comparison (0.25), which is significantly smaller than the $\omega$s estimated using intra-species comparisons ($P < 0.001$, using a Z test) (Table 1). These results are in conformity with the theoretical predictions suggesting that intra-species $\omega$ is expected to be higher than inter-species $\omega$ when $Nes < 0$ (Kimura 1983; Kryazhimskiy and Plotkin 2008; Peterson and Mase 2009).

McDonald and Kreitman (McDonald and Kreitman 1991) showed that under neutrality the $\omega$ estimated for human populations ($\omega_H$) is expected to be equal to the $\omega_{HC}$ estimated for the inter-specific (human-chimp) comparison i.e. $\omega_H = \omega_{HC}$. It is clear from table 1 that the $\omega$ estimated for the human population is always higher than that estimated for the human-chimp comparison i.e. $\omega_H > \omega_{HC}$. This is due to the presence of a fraction of deleterious nSNPs ($\delta$) in the populations. To subtract the fraction of deleterious SNPs the equation could be written as $\omega_H - (\omega_H \delta) = \omega_{HC}$. This equation could be simplified to estimate the fraction of deleterious nonsynonymous SNPs ($\delta$) as $\delta = 1 - (\omega_{HC}/\omega_H)$. The measure $\delta$ is the fraction of deleterious nSNPs that are segregating in the population. I estimated $\delta$ for each branch of the tree and as
expected, this measure systematically declined from the tip to the root of the tree (Figure 1). This is also clear from the negative relationship between $\delta$ and the relative age of the SNPs (Figure 2). Importantly, deleterious proportion is up to 7.9 times higher among the nSNPs specific to individual genomes (0.48 - branch $E$) than those shared with the ancient Khoisan genome (0.06 - branch $A$).

To obtain an independent estimate of the fraction of deleterious nSNPs, I used *PolyPhen-2*, which predicts the possible impact of an amino acid substitution on the structure and function of a human protein using physicochemical and comparative information (Adzhubei *et al.* 2010). Based on severity, this software predicts an amino acid change as either benign or damaging. The proportion of nSNPs damaging to the protein ($\rho$) was then estimated using nSNPs specific to each branch of the tree. It is interesting to note that the pattern of $\rho$ on the human population tree is almost identical to that of $\delta$ (Table 1 and Figure 1&2). Among the oldest nSNPs that are shared with the Khoisan, only 10% were predicted to be damaging to protein structure/function (branch $A$) while this fraction was 33% for the genome-specific nSNPs.

Although the declining patterns of $\delta$ from tip to root are very clear in African as well as non-African lineages, the $\delta$ estimates of non-African branches (0.29-0.48) (Figure 2A&B) are higher than those of African ones (0.21-0.38) (Figure 2C). This could be due to the effect of population bottlenecks that presumably had occurred in European and Asian populations (Li and Durbin 2011; Marth *et al.* 2004). During this period some of the deleterious SNPs might have drifted to higher frequencies, which are expected to decrease during population expansion [see, (Hughes
et al. 2005). Since the African population has had an uninterrupted (or mildly bottlenecked) population history (Li and Durbin 2011), such an effect is expected to be minimal.

The results show that the deleterious proportion of nSNPs is can be up to 48% of the amino acid variants specific to a single genome. However, this is an underestimate as this analysis was based only on a small dataset of 10 genomes. Adding more genomes to this analysis will redefine the pattern of sharing of SNPs as a number of nSNPs that are currently identified as lineage or genome specific will likely be shared among diverse populations. However a large proportion of these shared nSNPs are expected to be neutral as deleterious SNPs are unlikely to perpetuate in a population for a long time. Hence, by adding more genomes, a large proportion of neutral nSNPs are likely to be shifted to deeper internal branches of the human population tree. Although this will lead to a reduction in the number of nSNPs (observed in this study) in the terminal branches of the population tree, a much higher (than reported here) proportion of them are likely to be deleterious in nature. Furthermore, the fraction of deleterious SNPs estimated here is based on the assumption that adaptive evolution in hominids is rare. If this assumption is not valid then the fractions of deleterious SNPs reported here are underestimates. This is because adaptive evolution results in higher rate of nonsynonymous substitution compared to nonsynonymous diversity. Hence by excluding the adaptive fraction while estimating $\delta$ will lead to lower inter-species $\omega$ (than that reported here), which will result in higher $\delta$ estimates.

A previous study using two populations (African-Americans and European-Americans) showed that the deleterious fraction of population specific nSNPs was 52 - 77% higher compared to that
shared between the two populations (LOHMUELLER et al. 2008). However using a much deeper human population tree this study showed that the deleterious fraction of genome-specific nSNPs is 3.3 – 7.9 times higher compared to that shared between all humans. Although the temporal pattern reported in this study is based on nonsynonymous coding SNPs, an identical pattern is expected for all noncoding SNPs such as those present in promoters, silencers, enhancers, splice-junctions and untranslated regions (UTRs) that are under selective constraint. The findings of this study suggest that SNPs that are exclusively present in a single genome or closely related genomes are more likely to be associated with a disease as they are under strong purifying selection (HUGHES et al. 2003). Hence genome-wide association studies could focus on these SNPs, particularly when the disease causing variants have large phenotypic effects (CIRULLI and GOLDSTEIN 2010).

ACKNOWLEDGEMENTS

The author is grateful to David Lambert and thanks Leon Huynen, Robert Friedman, Michael Nachman and two anonymous reviewers for valuable comments.

FOOTNOTES

Supporting information is available online at: http://www.genetics.org/. Polyphen data is available at http://www.mediafire.com/?5jila01rh49dmgm.

LITERATURE CITED


Table 1. Genome-wide human SNPs and parameter estimates

<table>
<thead>
<tr>
<th>Branch on the tree</th>
<th>Nonsynonymous SNPs (A)</th>
<th>Synonymous SNPs (S)</th>
<th>A/S</th>
<th>$\omega = \frac{dN}{dS}$</th>
<th>Fraction of deleterious nSNPs ($\delta$)</th>
<th>Damaging nSNPs</th>
<th>Benign nSNPs</th>
<th>Fraction of damaging nSNPs ($\rho$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC*</td>
<td>62373</td>
<td>92246</td>
<td>0.68</td>
<td>0.25 (0.001)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>6494</td>
<td>9019</td>
<td>0.72</td>
<td>0.26 (0.004)</td>
<td>0.06 (0.017)</td>
<td>503</td>
<td>4598</td>
<td>0.10 (0.004)</td>
</tr>
<tr>
<td>B</td>
<td>3028</td>
<td>3723</td>
<td>0.81</td>
<td>0.30 (0.007)</td>
<td>0.17 (0.025)</td>
<td>382</td>
<td>1982</td>
<td>0.16 (0.008)</td>
</tr>
<tr>
<td>C</td>
<td>2003</td>
<td>2181</td>
<td>0.92</td>
<td>0.33 (0.010)</td>
<td>0.26 (0.032)</td>
<td>329</td>
<td>1221</td>
<td>0.21 (0.010)</td>
</tr>
<tr>
<td>D</td>
<td>1210</td>
<td>1232</td>
<td>0.98</td>
<td>0.36 (0.014)</td>
<td>0.31 (0.043)</td>
<td>217</td>
<td>758</td>
<td>0.22 (0.013)</td>
</tr>
<tr>
<td>E</td>
<td>5172</td>
<td>3956</td>
<td>1.31</td>
<td>0.48 (0.010)</td>
<td>0.48 (0.024)</td>
<td>1364</td>
<td>2782</td>
<td>0.33 (0.007)</td>
</tr>
<tr>
<td>F</td>
<td>293</td>
<td>309</td>
<td>0.95</td>
<td>0.35 (0.028)</td>
<td>0.29 (0.085)</td>
<td>55</td>
<td>186</td>
<td>0.23 (0.027)</td>
</tr>
<tr>
<td>G</td>
<td>2333</td>
<td>1843</td>
<td>1.27</td>
<td>0.46 (0.014)</td>
<td>0.47 (0.034)</td>
<td>576</td>
<td>1327</td>
<td>0.30 (0.011)</td>
</tr>
<tr>
<td>H</td>
<td>411</td>
<td>478</td>
<td>0.86</td>
<td>0.31 (0.021)</td>
<td>0.21 (0.069)</td>
<td>77</td>
<td>261</td>
<td>0.23 (0.023)</td>
</tr>
<tr>
<td>I</td>
<td>4013</td>
<td>4284</td>
<td>0.94</td>
<td>0.34 (0.007)</td>
<td>0.28 (0.023)</td>
<td>878</td>
<td>2382</td>
<td>0.27 (0.008)</td>
</tr>
<tr>
<td>J</td>
<td>5565</td>
<td>5140</td>
<td>1.08</td>
<td>0.39 (0.008)</td>
<td>0.38 (0.021)</td>
<td>1535</td>
<td>3118</td>
<td>0.33 (0.007)</td>
</tr>
</tbody>
</table>

* - Human-Chimpanzee comparison
FIGURE LEGENDS

**Figure 1.** Population phylogeny of human genomes used in this study. At each branch, estimates of the following are given in parenthesis: i) Ratio of nonsynonymous to synonymous SNPs per site (\( \omega \)) ii) Fraction of deleterious nonsynonymous SNPs (\( \delta \)) iii) Proportion of damaging nonsynonymous SNPs (\( \rho \)). *Estimates for terminal branches are the average of five, two and two corresponding genomes of Europeans, Asians and Africans respectively. Materials and methods are provided in the supporting information.

**Figure 2.** The proportion of deleterious nonsynonymous polymorphisms (\( \delta \)) and the fraction of damaging amino acid polymorphisms (\( \rho \)) estimated for each branch of the human population tree shown in Figure 1. The negative relationship between deleterious nSNPs and relative time is shown for the A) European lineage B) Asian lineage C) African lineage. Error bars indicate the standard error of the mean.
Figure 1
Subramanian
Figure 2
Subramanian