Supplementary file from

Sankar Subramanian

Effect of genetic drift on determinants of protein evolution

Biology Letters

Methods

Randomization test: We used a permutation or randomization test to examine whether $\delta_{\omega P}$ estimated using the mean estimates of genes from the two tails of the distributions was significantly higher than the δ_{LH} estimated from the means of two randomly drawn samples. For this purpose, we resampled two non-overlapping sets of genes (with sample sizes identical to those given in Table 1) and computed δ_{LH}. This was repeated for 10,000 times and calculated the proportion of resampled δ_{LH} estimates greater than or equal to the corresponding original δ_{LH}.

Tajima D estimation: We computed the *Tajima D* statistic for all synonymous SNVs from the whole genomes using equation 38 of Tajima (1989). For *M.m. castaneus, Tajima D* was –0.546 (n=20 and 250,200 SNVs) and for *M.m. musculus* it was 0.283 (n=16 and 60,921 SNVs). The 90% confidence intervals were obtained from Table 2 of Tajima (1989), which were computed using equation 47 of this paper under the assumption of a beta distribution with a mean and variance to be 0 and 1. The 90% confidence intervals for n=20 and n=16 were –1.584 to 1.710 and –1.583 to 1.709 respectively. Since the observed *Tajima D* values fall within these ranges they were not statistically significant ($P > 0.1$). We also estimated *Tajima D* and obtained confidence intervals based on a coalescence simulation using the software *DNASP* (Rozas et.al 2017). For this purpose, we generated the *hapmap* formatted files containing the chromosomal coordinates and genotypes of mouse SNVs and used them as the input file. Since this analysis could be performed only for a maximum of 2000 segregation sites we randomly selected 2000 biallelic positions. We first estimated *Tajima D*, which were -0.58 and 0.30 for *M.m. castaneus* and *M.m. musculus* respectively. We then conducted a coalescence simulation using 1000 replications for sample sizes of 20 and 16. We obtained a confidence interval of –1.66 to 1.69 for n=20. This program also calculated the probability for observing a *Tajima D* of –0.58 (for *M.m. castaneus*), which was $P = 0.31$. Similarly, the CI for n=16 was estimated to be –1.74 to 1.64 and the probability observing a *Tajima D* of 0.30 (for *M.m. musculus*) was $P = 0.66$.

Distribution of fitness effects: To estimate the distribution of fitness effects of nonsynonymous SNVs we used the program DoFE (Eyre-Walker et.al. 2006). We first ran the *lookupTableGenerator* using a sample size of 20 and used 10 and 1,000,000 for lower and upper limits of meanS with 100 steps. Similarly, 0 and 0.5 were used for lower and upper limits of Beta with 100 steps. This process was repeated using a sample size of 16 for *M.m. musculus*. We obtained the site frequency spectrum of synonymous and nonsynonymous SNVs of both mouse populations. We did not orient the SNVs as this will not influence the results (Eyre-Walker et.al. 2006). We then ran the DoFE program using the site frequency and the lookupTable data. This program grouped the nonsynonymous SNVs based on their fitness effects, which is expressed in terms of the product of effective population size (Ne) and selection coefficient (s) and estimated their relative proportions (Table S2 and Figure 2).

References

Eyre-Walker A et.al. (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173:891-900.

Rozas, J., et.al (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. Mol. Biol. Evol. 34: 3299-3302.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–95.

* - Since the distribution of the number of protein-protein interactions is discrete

the numbers of the genes in tails slightly differ