Influence of Effective Population Size on Genes under Varying Levels of Selection Pressure

Sankar Subramanian*

GeneCology Research Centre, The University of the Sunshine Coast, Sippy Downs, Queensland, Australia

*Corresponding author: E-mail: ssankara@usc.edu.au.

Accepted: February 21, 2018

Abstract

The ratio of diversities at amino acid changing (nonsynonymous) and neutral (synonymous) sites ($\omega = \pi_N/\pi_S$) is routinely used to measure the intensity of selection pressure. It is well known that this ratio is influenced by the effective population size (N_e) and selection coefficient (s). Here, we examined the effects of effective population size on ω by comparing protein-coding genes from *Mus musculus castaneus* and *Mus musculus musculus*—two mouse subspecies with different N_e. Our results revealed a positive relationship between the magnitude of selection intensity and the ω estimated for genes. For genes under high selective constraints, the ω estimated for the subspecies with small N_e (*M. m. musculus*) was three times higher than that observed for that with large N_e (*M. m. castaneus*). However, this difference was only 18% for genes under relaxed selective constraints. We showed that the observed relationship is qualitatively similar to the theoretical predictions. We also showed that, for highly expressed genes, the ω of *M. m. musculus* was 2.1 times higher than that of *M.m. castaneus* and this difference was only 27% for genes with low expression levels. These results suggest that the effect of effective population size is more pronounced in genes under high purifying selection. Hence the choice of genes is important when ω is used to infer the effective size of a population.

Key words: population size effect, deleterious mutations, amino acid diversity, gene expression and population genetics theory.

Introduction

The ratio of diversities ($\omega = \pi_N/\pi_S$) at nonsynonymous (π_N) and synonymous (π_S) sites of protein-coding genes reveal the intensity of natural selection on genes (Li 1997). Furthermore, ω suggests the fraction of nonsynonymous single nucleotide variations (SNVs) segregating in a population with respect to synonymous SNVs. This also suggests that a fraction of nonsynonymous SNVs has been eliminated from a population owing to their deleterious nature when $\omega < 1$. It is well known that ω is dictated by the product of effective population size (N_e) and selection coefficient (s) (Kimura 1983). Population genetic theories predict that ω estimated for large populations tend to be smaller than those obtained for small populations (Ohta 1993). This is because a much higher fraction of deleterious SNVs is removed in large populations compared with small populations owing to the difference in the efficacy of selection between them. Therefore, the overall variation in ω observed for different populations suggest the potential difference in their effective population sizes. Hence ω is routinely used to infer the difference in N_e between populations (Strasburg et al. 2011; Phifer-Rixey et al. 2012; Tsagkogeorga et al. 2012; Gayral et al. 2013; Harrang et al. 2013; Romiguier et al. 2014; James et al. 2017). In contrast, ω estimates also vary significantly between genes of a genome, which reflects the magnitude of selection on them (Bustamante et al. 2005). However, it is unclear how and to what extent the difference in effective population sizes influences the ω of various genes that are under different levels of selection constraints.

To examine this, we assembled the genome-wide SNV data from two subspecies of mouse: *Mus musculus castaneus* and *Mus musculus musculus*. These two species were estimated to be diverged ~500,000 years ago (Geraldes et al. 2008; Duvaux et al. 2011) and have lived in reproductive isolation (Good et al. 2008). The effective population sizes were estimated to be 580,000 (200,000–733,000) and 76,000 (25,000–120,000) (Salcedo et al. 2007; Geraldes et al. 2008; Halligan et al. 2010). Since these species diverged only recently and live in similar habitats (commensal with humans) but differ only in N_e our results are not confounded by the difference in other factors such as physiology, genetics, and ecology between the two groups compared. We first

examined the theoretical relationship between s and ω for different effective population sizes. Using genome-wide SNVs from the two mice we then investigated the empirical relationship by using the proportion of constrained sites and level of gene expression as proxies for the magnitude of selection (s).

Materials and Methods

Whole genome genotype data for 19 autosomes of *Mus musculus castaneus* and *Mus musculus musculus* were downloaded from the data repository (http://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/) of a previous study (Harr et al. 2016). The genome-wide SNVs data were available for 8 ($n = 16$) and 10 ($n = 20$) individuals of *M.m. musculus* and *M.m. castaneus*, respectively. Using the software program *SNPEff* (http://snpeff.sourceforge.net/) functional annotations were inferred (Cingolani et al. 2012). We then extracted only SNVs affecting coding regions and introns. We also downloaded the mouse reference genome (GRCm38/mm10) sequence and the annotation file (refGene.txt) from the UCSC genome browser data resource (https://genome.ucsc.edu/). Using the annotations, we extracted the protein-coding gene sequences and their chromosomal locations. The number of synonymous, nonsynonymous, and intron positions were also extracted from the annotation file. Using the above, we estimated the diversity at nonsynonymous, synonymous, and intron sites. For this purpose, we use the mean pairwise differences per site (π) (Tajima 1983). We also downloaded the mouse-rat genome alignment from the UCSC genome browser and extracted the alignments for each protein-coding gene. We estimated synonymous divergence for each gene using the likelihood based method implemented in the *codeml* program of the *PAML* package (Yang 2007).

We obtained the basewise conservation scores (*PhyloP*) based on 59 vertebrate genomes with mouse (http://hgdownload.soe.ucsc.edu/goldenPath/mm10/ phyloP60way/) (Siepel et al. 2006). The score was available for each position of the mouse chromosomes. The *PhyloP* scores were then mapped to on to the protein-coding genes and we designated any position of a gene with a *PhyloP* score >2 as constrained. Based on this criterion, we estimated the proportion of constrained sites in each protein-coding gene. Our final data set included 16,285 reference protein-coding genes. These genes were grouped into 13 categories based on the fraction of constrained positions as (number of genes): <10 (1,219), 10–15 (1,295), 15–20 (1,727), 20–25 (1,952), 25–30 (2,097), 30–35 (1,998), 35–40 (1,799), 40–45 (1,425), 45–50 (1,094), 50–55 (755), 55–60 (493), 60–65 (277) and $>65\%$ (154).

We obtained the RNA-seq expression data (http://chromosome.sdsc.edu/mouse/download.html) from a previous large-scale study using 19 mouse tissues (Dunham et al. 2012). The values of Fragments Per Kilobase of transcript per million mapped reads (FPKM) for each gene from 19 tissues were averaged and a log-transformed mean was used for further analysis. We used only the genes that were expressed in all tissues. This data set included 6,733 mouse genes, which were sorted based on their FPKM values that represent the level of gene expression. These genes were then grouped by taking 500 genes in each of the 12 categories with FPKM values ranging: >25, 15–25, 11–15, 9–11, 7.4–9, 6.3–7.4, 5.4–6.3, 4.6–5.4, 3.9–4.6, 3.2–3.9, 2.6–3.2, 2.0–2.6, and the 13th category contains the remaining 733 genes with <2.0 FPKM.

To determine the significance of positive selection we examined whether the number of nonsynonymous SNVs per nonsynonymous site (*pN*) is significantly higher than the number of synonymous SNVs per synonymous site (*pS*) for a gene. For this purpose, we obtained the Jukes–Cantor variance for the two measures and used a *Z*-test to examine the significance. We used the nonparametric Spearman rank correlation to test the strength and significance of observed correlations.

For mutation rate analysis, we used the synonymous substitution rate as the proxy for mutation rate of each gene and sorted 16,285 genes based on this. We separated top and bottom 30% of the genes (4,886 in each category) and designated them as slow and fast evolving genes, respectively. Within each group, we further separated the genes with $>50\%$ and $<10\%$ constrained sites as constrained and relaxed genes. For recombination analysis, we obtained the fine-scale map of recombination rates created for the mouse genome by a previous study (Brunschwig et al. 2012), which could be mapped to 8,205 genes of our data set. We calculated the mean recombination rate for each protein-coding gene and sorted the genes based on this. We then separated the top and bottom 30% of the genes (2,461 in each category) and termed them as low (with a recombination rate <0.005 $N_e r$/kb) and high (>0.029 $N_e r$/kb) recombining genes, respectively. Within each group, we separated constrained and relaxed genes as mentioned earlier.

We examined the theoretical relationship between $N_e s$ and ω using Kimura's equation [3.18 in page 45 of (Kimura 1983)]:

$$\omega(H_T/H_{T,\,0}) = \frac{2(S - 1 + e^{-S})}{[S(1 - e^{-S})]}, \quad (1)$$

where $S = 4N_e s$. In the above equation, H_T is the sum of heterozygotes involving a mutant allele over all generations until either fixation or loss and $H_{T,0}$ denotes mutant with $s \rightarrow 0$. Hence H_T denotes neutral and nonneutral mutations and $H_{T,0}$ indicates only neutral mutations. For empirical data, nonsynonymous and synonymous mutations represent H_T and $H_{T,0}$ of equation 1. Because a nonsynonymous mutation could be deleterious or neutral but a synonymous or intron mutation is largely neutral in nature. Therefore, we used the ratio of

heterozygosity or diversity of nonsynonymous and synonymous mutations as the empirical equivalence of equation 1. Note that equation 1 assumes that the effective population is equal to the actual population ($N_e = N$).

The difference between the ωs from two populations was quantified as:

$$\delta_\omega = \frac{\omega_A - \omega_B}{\omega_A} \text{ or } \delta_\omega = \frac{\omega_{mus} - \omega_{cas}}{\omega_{mus}} \quad (2)$$

where ω_A and ω_B is estimated for the population with small and large effective population sizes (eg. *M.m. musculus* and *M.m. castaneus*), respectively.

Results

To understand the theoretical relationship between *s* and ω for different N_e, let us assume that we compare two populations A and B with 76,000 and 580,000 as their N_e, respectively (to represent the N_e of *M. m. musculus and M. m. castaneus*, respectively) and *s* is the mean selection coefficient on the nonsynonymous sites of a gene or a collection of genes. The theoretical relationship has been derived by Kimura (1983), which is shown equation 1. Using this formula, we assigned values for *s* ranging from 0 to -1.7×10^{-5} with an increment of -1.3×10^{-6} and computed ω_A and ω_B for the hypothetical populations with two different N_e mentioned earlier. Figure 1*A* shows that the difference between ω_A and ω_B is large when negative selection is high and this difference disappears when *s* approaches zero. When $s = -1.7 \times 10^{-5}$, ω estimated for the population A with small effective population size ($N_e s = 1.3$) was 3.9 times higher than that estimated for the large population B ($N_e s = 10$). However, this fraction becomes equal between the populations when *s* is very close to zero ($s \rightarrow 0$). This is further clear from the positive relationship between selection coefficient and the magnitude of difference between the ωs (δ_ω) (fig. 1*B*).

To examine the influence of population size on a genome scale, we estimated the nucleotide diversities using the whole genome data of *M.m. castaneus* and *M.m. musculus* populations. The genome-wide estimates are given in table 1. The effective population size difference between the two subspecies is evident from the difference in the number of SNVs. Although nonsynonymous diversity of *M.m. castaneus* was 2.3 times higher than that of *M.m. musculus*, synonymous and intron diversities estimated were 3.3 and 3.6 times higher for the former than the latter. The ω estimated for the whole genomes of *M.m. musculus* population was 31% and 37% (based on synonymous sites and introns, respectively) higher than those obtained for *M.m. castaneus* population. To examine the empirical relationship between selection intensity and ω, we first computed the proportion of constrained sites for each protein-coding gene as described in methods and this was used as a proxy for selection intensity (*s*). The diversities at nonsynonymous (π_N) and synonymous sites (π_S) were

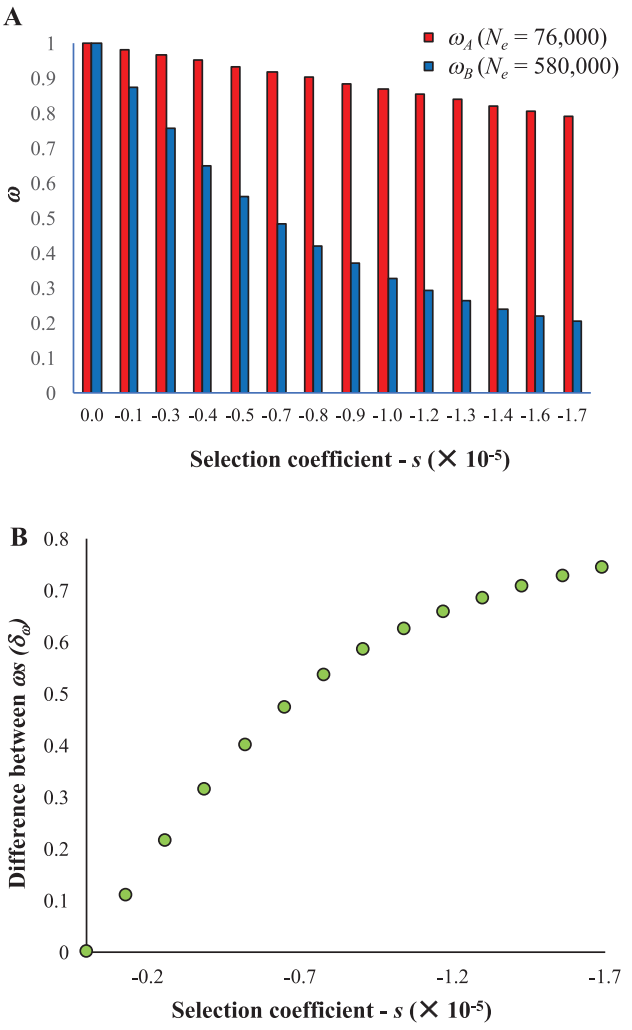

Fig. 1.—Theoretical relationship between the ratio of diversity at selected and neutral sites ($\omega = H_T/H_{T,0}$), effective population size (N_e) and selection coefficient (*s*). (*A*) Using equation 1 (see Materials and Methods), ω estimated for 13 different values of *s* ranging from 0 to -1.7×10^{-5} (increment of -1.3×10^{-6}) for two populations (*A* and *B*) with effective population sizes of 76,000 (ω_A) and 580,000 (ω_B) to represent those of *Mus musculus musculus* and *M.m. castaneus*, respectively. Note when $s = 0$ the equation becomes undefined (0/0) and therefore the first two columns were based on the assumption of *s* being infinitesimally small ($s \rightarrow 0$), which results in ω very close to 1 ($\omega \rightarrow 1$). (*B*) Relationship between selection coefficient (*s*) and the normalized difference between ω_A and ω_B (δ_ω).

used to represent evolution at constrained and neutral sites, which forms the empirical ratio ω (π_N/π_S). We grouped genes based on the proportion of constrained sties into 13 categories and obtained the mean ω for the genes belonging to each category. Cleary, the patterns of relationships in figure 2 are similar to those shown in figure 1, suggesting that the genome data provide the empirical proof for the theoretical relationship. For highly constrained genes (with >65% constrained sites) ω estimated for *M.m. musculus* (ω_{mus})

Table 1
Summary Statistics

	Mus musculus castaneus	*M.m. musculus*
SNVs		
Nonsynonymous	103763	33676
Synonymous	229655	55760
Intron	14831843	3266566
Diversity		
Nonsynonymous (π_N)	0.00076 ($\pm 6.1 \times 10^{-6}$)	0.00033 ($\pm 4.0 \times 10^{-6}$)
Synonymous (π_S)	0.0057 ($\pm 2.7 \times 10^{-5}$)	0.0017 ($\pm 1.5 \times 10^{-5}$)
Intron (π_I)	0.0037 ($\pm 2.1 \times 10^{-6}$)	0.0010 ($\pm 1.1 \times 10^{-6}$)
π_N/π_S	0.13 (0.0012)	0.19 (0.0028)
π_N/π_I	0.20 (0.0016)	0.32 (0.0039)
Difference between ω - δ_ω		
Using synonymous sites	0.31 (0.008) — $P<0.0001$	
Using intron	0.37 (0.011) — $P<0.0001$	

was 0.079 (± 0.017), which was three times higher than that estimated for *M.m. castaneus* [$\omega_{cas} = 0.026$ (± 0.004)] and the difference between the ωs was significantly >0 ($P = 0.0015$, one-tailed *Z*-test). Whereas, this difference between the ωs was only 18% [0.470 (± 0.027) vs. 0.388 (± 0.014)] for genes under relaxed selection pressure (<10% constrained sites) and it was statistically significant ($P = 0.0031$). This result is further supported by the positive relationship ($\rho = 0.97$, $P < 0.0001$) between the proportion of constrained sites and the normalized difference between ω_{mus} and ω_{cas} (δ_ω) estimated for *M.m. castaneus* and *M.m. musculus* (fig. 2*B*).

A number of earlier studies showed that highly expressed genes are under high selection pressure and expression level is a major determinant of protein evolution (Subramanian and Kumar 2004; Drummond et al. 2005; Yang et al. 2012). Based on this rationale, we used the level of gene expression as an independent proxy for selection intensity (*s*) and examined its relationship with ω. For this purpose, we obtained RNA-seq data from a previous study (Dunham et al. 2012). Based on the level of expression (in FPKM units) genes were grouped into 13 categories and the average ω was computed for genes belonging to each category. Our results based on expression levels were very similar to those obtained for the proportion of constrained sites (fig. 3). For highly expressed genes ω estimated for *M.m. musculus* [$\omega_{mus} = 0.130$ (± 0.014)] was 2.1 times higher than that estimated for *M.m. castaneus* [$\omega_{cas} = 0.062$ (± 0.006)] and the difference between the ωs was statistically significant ($P < 0.0001$). Whereas this difference between the ωs was only 27% for the genes with low expression levels [0.238 (± 0.015) vs. 0.174 (± 0.006)] and it was statistically significant ($P = 0.0073$). A highly significant positive correlation

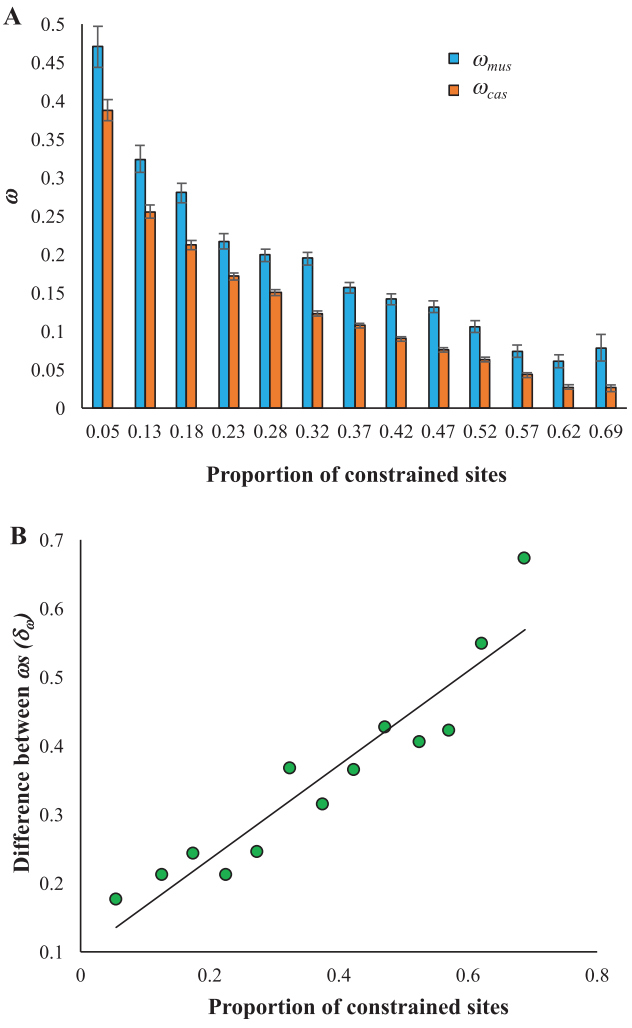

Fig. 2.—Empirical relationship between the proportion of constrained positions and ω estimated for *Mus musculus musculus* (ω_{mus}) and *M.m. castaneus* (ω_{cas}) using nonsynonymous and synonymous sites. (*A*) Protein-coding genes were grouped into 13 categories based on the proportion of constrained sites (see Materials and Methods) and average ω computed for genes belonging to each category are shown. We used the fraction of constrained sites as a proxy for selection intensity (*s*). Error bars shows the standard error of the mean. The difference between ω_{mus} and ω_{cas} was statistically significant for all categories (at least $P < 0.01$). (*B*) Relationship between the proportion of constrained sites in protein-coding genes and normalized difference between ω_{mus} and ω_{cas} (δ_ω) is shown. This relationship is highly significant ($\rho = 0.97$, $P < 0.0001$). Best fitting regression line is shown.

($\rho = 0.95$, $P < 0.0001$) between expression levels and the normalized difference between ω_{mus} and ω_{cas} (δ_ω) provides confirmatory support for our results (fig. 3*B*).

Discussion

Our results suggest that the influence of effective population size is more pronounced in genes under high selection intensity. In this study, we first used the proportion of constrained

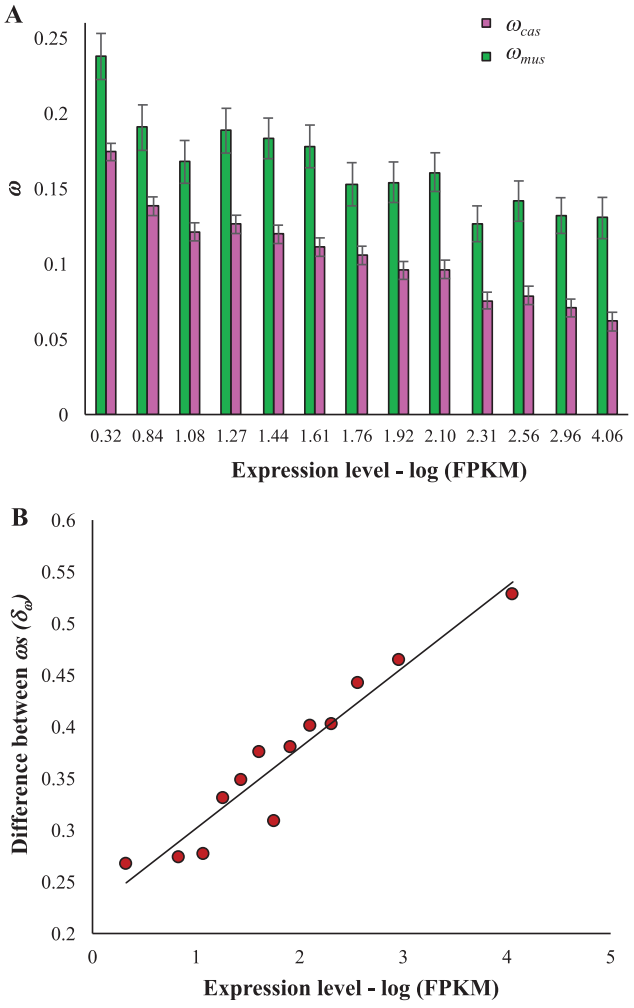

Fig. 3.—Relationship between gene expression levels and ω estimated for *Mus musculus musculus* (ω_{mus}) and *M.m. castaneus* (ω_{cas}) using nonsynonymous and synonymous sites. (*A*) Genes were grouped into 13 categories based on their expression levels represented by the Fragments Per Kilobase of transcript per million mapped reads (FPKM) units (see Materials and Methods) and the average ω estimated for genes belonging to each category are shown. Here, we used the level of gene expression as an independent proxy for selection intensity (*s*). Error bars shows the standard error of the mean. The difference between ω_{mus} and ω_{cas} was statistically significant for all categories (at least $P < 0.01$). (*B*) Relationship between the gene expression levels and normalized difference between ω_{mus} and ω_{cas} (δ_ω) is shown ($\rho = 0.95$, $P < 0.0001$). Best fitting regression line is shown.

sites as a proxy for selection intensity (*s*), which is straightforward. Since the level of gene expression is known to correlate with selection intensity (Subramanian and Kumar 2004; Drummond et al. 2005; Yang et al. 2012) we used this as an independent proxy for *s*. However, both measures produced almost identical patterns of relationships with ω. The pattern of our population diversity based results is similar to that reported based on divergence between species (Subramanian 2013). Overall, the higher ω observed for

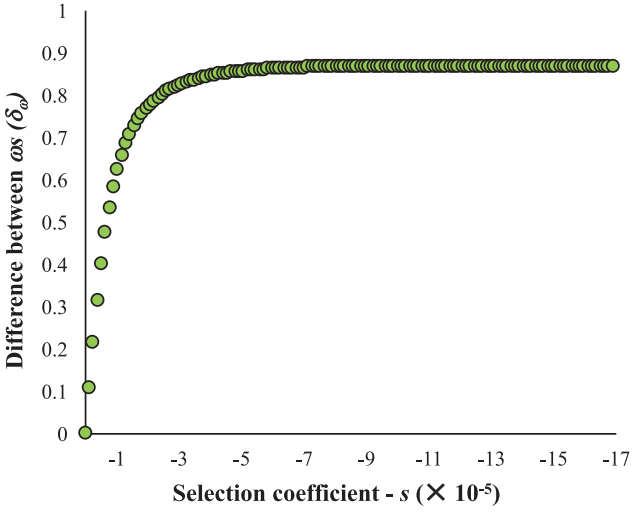

Fig. 4.—Theoretical relationship between selection coefficient (*s*) and normalized difference between ω_A and ω_B (δ_ω) is shown for higher values of *s*. Note that δ_ω changes only when the values of *s* fall between the nearly neutral range of 0 to −2.

M.m. musculus than *M.m. castaneus* suggests a greater fraction of deleterious mutations segregating in the former. This is due to the fact that selection is not efficient in purging deleterious mutations in small populations. We used synonymous sites as a proxy for neutral evolution. However, previous studies suggested that a fraction of synonymous sites could be under selective constraints (Chamary et al. 2006). If negative selection is assumed in synonymous sites the magnitude of this effect will be more pronounced in populations with large effective sizes and in this case in *M.m. castaneus*. This will in turn expected to further increase the difference between ωs (δ_ω) from the two mouse populations. Hence this assumption will make our results more conservative. However, we addressed this issue by replacing introns for synonymous sites to estimate neutral diversities and found similar results (supplementary figs. S1 and S2, Supplementary Material online).

In figure 1, the theoretical relationship was shown only for small values of *s* (0 to -1.7×10^{-5}). However, the corresponding $N_e s$ values are −10 and −1.3 for *M.m. castaneus* and *M.m. musculus*, respectively. Previous studies on the distribution of the fitness effects of mutations in *M.m. castaneus* populations suggested that almost 77–80% of the mutations with $N_e s > 10$ were lethal or highly deleterious and ~20% of them ($N_e s < 10$) were nearly neutral in nature (Halligan et al. 2010, 2013; Kousathanas and Keightley 2013). It is well known that only nearly neutral or slightly deleterious mutations are influenced by N_e and both neutral and highly deleterious mutations are independent and are not modulated by effective population sizes. This is very clear from figure 4, which shows a plateau or an asymptote for δ_ω when $s > -2.0 \times 10^{-5}$. Due to this reason, we chose to show only the nearly neutral range in figure 1. Because the mutations/

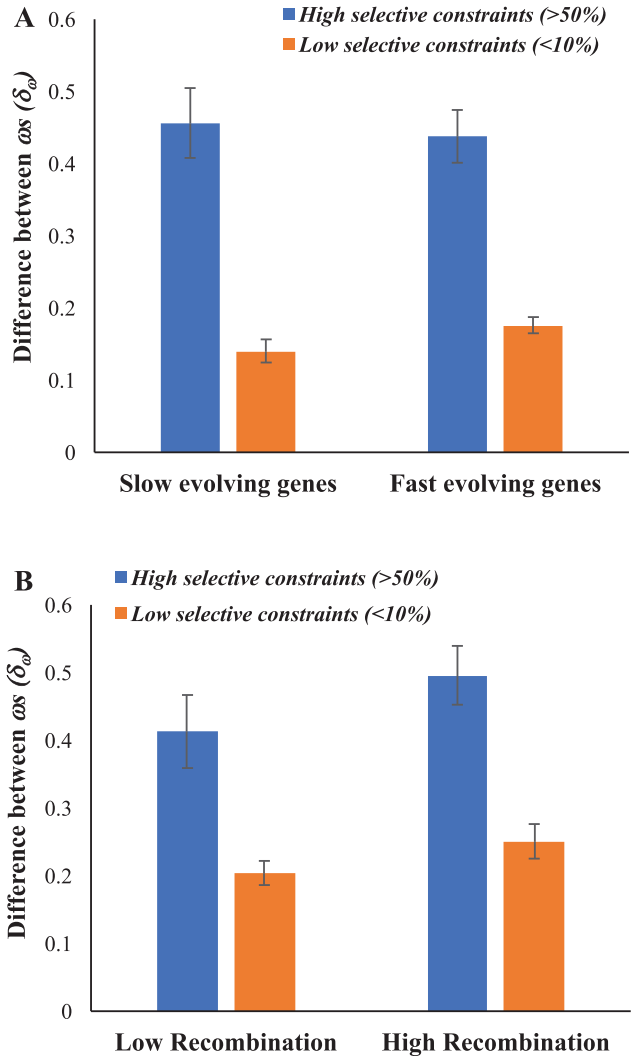

Fig. 5.—Normalized difference between ω_{mus} and ω_{cas} (δ_{ω}) estimated for genes under high and low selective constraints using nonsynonymous and synonymous sites. (*A*) Genes evolving under high and low substitution rates (see Materials and Methods). (*B*) Genes present in high and low recombining regions. All differences between δ_{ω} of genes under high and low selective constraints were statistically significant at least $P < 0.0001$. Error bars show the standard error of the mean.

variations associated with the observed difference in the ωs estimated for *M.m. castaneus* and *M.m. musculus* were predominantly nearly neutral or slightly deleterious in nature.

Theoretical relationship shown in figure 1 was also based on simple assumptions and did not consider the influence of other factors such as mutation and recombination rate difference between genes, which might result in different N_e for genes. To examine this effect, first we used the rate of substitution at synonymous sites as the proxy for mutation rate and sorted genes based on the synonymous divergence between mouse and rat. We then obtained the top 30% of the genes with slowest evolutionary rate and within this category we estimated δ_{ω} for the genes under high and low selective constraints (see Materials and Methods). Similar estimates were obtained for the bottom 30% of the genes with fastest evolutionary rate. The difference between ωs of *M.m. musculus* and *M.m. castaneus* (δ_{ω}) estimated for slow-evolving constrained genes was 3.2 times higher than that obtained for slow-evolving relaxed genes (0.46 vs. 0.14) (fig. 5*A*). Similarly, this difference for fast-evolving constrained genes was 2.5 times higher than that estimated for fast-evolving relaxed genes (0.438 vs. 0.176). Similar results were obtained when diversity at introns (instead of synonymous sites) were used to estimate ω (supplementary fig. S3*A*, Supplementary Material online). These results revealed that the effect of effective population size was more pronounced in constrained than in relaxed genes and the magnitude of this effect was largely similar in the fast and slow mutating genes. Therefore, difference in mutation rate between genes is unlikely to affect the main results of this study.

Next, to examine the effects of recombination we obtained the fine-scale map of recombination rates from a previous study (Brunschwig et al. 2012) and computed the mean recombination rate for each mouse protein-coding gene. Similar to the previous analysis we sorted genes based on their recombination rates and obtained the top and bottom 30% of the genes with low and high recombination rates, respectively. The difference between ωs (δ_{ω}) of *M.m. musculus* and *M.m. castaneus* estimated for low-recombining constrained genes was 2 times higher than that obtained for low-recombining relaxed genes (0.4 vs. 0.2) (fig. 5*B*). Similarly, this difference for high-recombining constrained genes was also two times higher than that estimated for high-recombining relaxed genes (0.5 vs. 0.25). Comparable results were also obtained when diversity at introns was used to estimate ω (supplementary fig. S3*B*, Supplementary Material online). The above results suggest that the magnitude of the effects of N_e on ω (or δ_{ω}) was similar for genes located in high and low recombining regions. Therefore, variation in the rate of recombination between genes do not affect the findings and conclusions of this study.

The results of this study are under the assumption that the fraction of adaptive nonsynonymous segregating variations is negligible. This is because when adaptive sweep occurs nonsynonymous SNVs will be quickly fixed and do not contribute to the segregating variation. We examined this using our data and found that 147 and 362 genes (*M.m. castaneus* and *M.m. musculus*, respectively) had more number of nonsynonymous SNVs per nonsynonymous site than synonymous SNVs per synonymous site (or $\omega > 1$). . However, the difference was statistically significant only for 4 and 5 genes, respectively (using a *Z* test—see Materials and Methods). Finally, the theoretical relationship shown in this study (fig. 1) is based on the assumption of no dominance and independence between mutations. However, empirical data include the effects of dominant mutations and interactions/epistasis between mutations. Hence these caveats should be noted while inferring the empirical results of this study.

Based on previous theoretical and empirical predictions, comparative population genetic studies generally assume the difference in ω observed between two populations to reflect the variations in their effective population sizes (Strasburg et al. 2011; Phifer-Rixey et al. 2012; Gayral et al. 2013; Romiguier et al. 2014; James et al. 2017). However, our results suggest that the magnitude of this difference is dependent upon the genes being compared. As we have shown that comparing genes under relaxed selective constraints will underestimate the actual difference in the effective population sizes. Therefore, it is important to consider the selection intensity on the genes when comparing ω between populations to infer effective population size. Although our results are based on protein-coding genes the findings will hold true for other constrained noncoding regions such as 3′ and 5′ untranslated regions, splice sites, up-, and downstream regulatory elements.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The author acknowledges the support from the Australian Research Council (LP160100594) and the University of the Sunshine Coast.

Literature Cited

Brunschwig H, et al. 2012. Fine-scale maps of recombination rates and hotspots in the mouse genome. Genetics 191(3):757–764.

Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437(7062):1153–1157.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 7(2):98–108.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: sNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80–92.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102(40):14338–14343.

Dunham I, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74.

Duvaux L, Belkhir K, Boulesteix M, Boursot P. 2011. Isolation and gene flow: inferring the speciation history of European house mice. Mol Ecol. 20(24):5248–5264.

Gayral P, et al. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet. 9(4):e1003457.

Geraldes A, et al. 2008. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol Ecol. 17(24):5349–5363.

Good JM, Handel MA, Nachman MW. 2008. Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. Evolution 62(1):50–65.

Halligan DL, et al. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 9(12):e1003995.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 6(1):e1000825.

Harr B, et al. 2016. Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. Sci Data 3:160075.

Harrang E, Lapegue S, Morga B, Bierne N. 2013. A high load of non-neutral amino-acid polymorphisms explains high protein diversity despite moderate effective population size in a marine bivalve with sweepstakes reproduction. G3 (Bethesda) 3:333–341.

James J, Castellano D, Eyre-Walker A. 2017. DNA sequence diversity and the efficiency of natural selection in animal mitochondrial DNA. Heredity (Edinb) 118(1):88–95.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kousathanas A, Keightley PD. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. Genetics 193(4):1197–1208.

Li W-H. 1997. Molecular evolution. Chicago: Sinauer Associates Inc.

Ohta T. 1993. An examination of the generation-time effect on molecular evolution. Proc Natl Acad Sci U S A. 90(22):10676–10680.

Phifer-Rixey M, et al. 2012. Adaptive evolution and effective population size in wild house mice. Mol Biol Evol. 29(10):2949–2955.

Romiguier J, et al. 2014. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. J Evol Biol. 27(3):593–603.

Salcedo T, Geraldes A, Nachman MW. 2007. Nucleotide variation in wild and inbred mice. Genetics 177(4):2277–2291.

Siepel A, Pollard KS, Haussler D. 2006. New methods for detecting lineage-specific selection. In: Apostolico A, Guerra C, Istrail S, Pevzner PA, Waterman M, editors. Research in computational molecular biology. RECOMB 2006. Lecture Notes in Computer Science. Berlin: Springer. p. 190–205.

Strasburg JL, et al. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol Biol Evol. 28(5):1569–1580.

Subramanian S. 2013. Significance of population size on the fixation of nonsynonymous mutations in genes under varying levels of selection pressure. Genetics 193(3):995–1002.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168(1):373–381.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105(2):437–460.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biol Evol. 4(8):740–749.

Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A. 109(14):E831–E840.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Associate editor: George Zhang