biology letters
Molecular evolution

Biol. Lett. (2009) **5**, 417–420
doi:10.1098/rsbl.2009.0155
Published online 8 April 2009

Opinion piece

Effective population size and the rate and pattern of nucleotide substitutions

Both the overall rate of nucleotide substitution and the relative proportions of synonymous and non-synonymous substitutions are predicted to vary between species that differ in effective population size (N_e). Our understanding of the genetic processes underlying these lineage-specific differences in molecular evolution is still developing. Empirical analyses indicate that variation in substitution rates and patterns caused by differences in N_e is often substantial, however, and must be accounted for in analyses of molecular evolution.

Keywords: effective population size; molecular evolution; substitution rate

1. INTRODUCTION

Nucleotide sequence data have been a great boon for the study of evolution. DNA sequences bring all organisms into the fold of comparative analyses, allowing us to jointly reconstruct the evolutionary histories of taxa that differ enormously in morphology and lifestyle. But while DNA is universal, its tempo and mode of evolution are not. It has become increasingly clear that the way in which a species' DNA evolves is affected by numerous aspects of its biology (e.g. Welch *et al.* 2008). One such aspect is effective population size (N_e), which is predicted to affect species' molecular evolution at many levels, from numbers of segregating nucleotide polymorphisms (Petit & Barbadilla 2008) to genome size and complexity (Lynch & Conery 2003; Hershberg *et al.* 2007). In this short review, however, I will focus on another level of molecular evolution affected by N_e: nucleotide substitutions. In particular, I will discuss how both the overall rate of nucleotide substitution and the ratio of non-synonymous to synonymous substitutions are likely to vary in lineages that differ in N_e.

2. WHAT IS EFFECTIVE POPULATION SIZE?

The simplest scenario under which change in allele frequencies can be studied is the Wright–Fisher model, which consists of a population of constant size N diploid individuals, with discrete generations, random mating and binomial distribution of offspring number per parent. In reality, all natural populations will deviate from the Wright–Fisher model in numerous ways. Wright therefore developed the concept of the effective population size, or N_e, which is the size of an idealized population that would experience the same effects of random sampling of alleles as the real population under consideration (Wright 1931; see also Charlesworth (2009) for a comprehensive review of subsequent theoretical developments).

The list of demographic or genetic factors expected to reduce N_e relative to N is long, and includes common phenomena such as skewed sex ratios, non-random mating, variance in reproductive success, fluctuations in census population size, some forms of population subdivision, and linkage between loci under selection (Charlesworth 2009). Even closely related species that vary in one or more of these traits may therefore have substantially different effective population sizes.

3. WHY SHOULD A SPECIES' N_e AFFECT ITS EVOLUTION?

N_e reflects the balance of power between selection and drift: in small populations, drift plays a greater role and selection (both positive and negative) is correspondingly less efficacious. A mutation is effectively neutral when the magnitude of its selective coefficient is less than or equal to the inverse of the effective population size (Kimura 1983), so as N_e decreases, mutations of larger and larger effects behave as neutral. In species with small N_e, therefore, increasing numbers of slightly deleterious mutations may drift to fixation rather than being removed by purifying selection, increasing the substitution rate for this class of mutations. By contrast, more slightly advantageous mutations are likely to be lost due to drift rather than being fixed by positive selection, decreasing the substitution rate for this second class of mutations in species with small N_e.

If advantageous mutations are rare, while a substantial proportion of mutations are slightly deleterious, then we should be able to detect an increase in overall substitution rate in lineages with small N_e compared with those with larger N_e (all else, including mutation rates, being equal). If we make the further assumption that non-synonymous mutations are more likely to be slightly deleterious than synonymous mutations, many of which are probably neutral (but see Chamary *et al.* 2006), the ratio of non-synonymous to synonymous substitution rates (ω) should also be greater in lineages with small N_e (Ohta 1992).

4. HOW GREAT SHOULD THE EFFECT BE?

The magnitude of the effect of a change in N_e on nucleotide substitutions is determined by the distribution of selective effects of mutations. To illustrate this, consider two lineages with different effective population sizes, the larger N_{eL} and the smaller N_{eS}. If we assume that advantageous mutations are rare and most of the mutations that go to fixation are slightly deleterious, then the difference in substitution rate between these lineages will be largely determined by the proportion of mutations that have selective coefficients between $1/N_{eL}$ and $1/N_{eS}$ (figure 1). This proportion, in turn, is determined by the distribution of selective effects.

Ohta (1977) assumed that the distribution of selection coefficients for new mutations was exponential. Under this distribution, and given a realistic mean

One contribution of 11 to a Special Feature on 'Whole organism perspectives on understanding molecular evolution'.

Received 25 February 2009
Accepted 16 March 2009

This journal is © 2009 The Royal Society

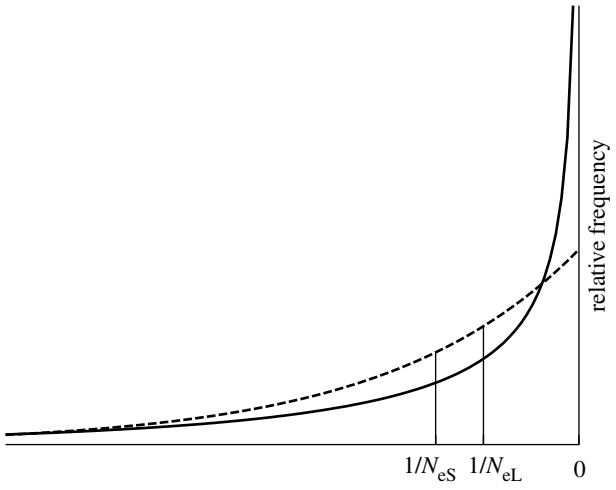

Figure 1. The distributions of fitness effects modelled by Ohta (1977) (exponential or gamma with $\beta=1$, dashed curve) and Kimura (1979) (gamma with $\beta=0.5$, solid curve). In a small population, with effective population size N_{eS}, mutations with selection coefficients between $1/N_{eS}$ and zero will be effectively neutral. Fewer mutations, those with selection coefficients between $1/N_{eL}$ and zero, will be effectively neutral in a larger population with N_{eL}. The proportion of mutations that have selective coefficients between $1/N_{eS}$ and $1/N_{eL}$ will be greater under a gamma distribution of fitness effects with $\beta=1$ than with $\beta=0.5$ for most regions of parameter space.

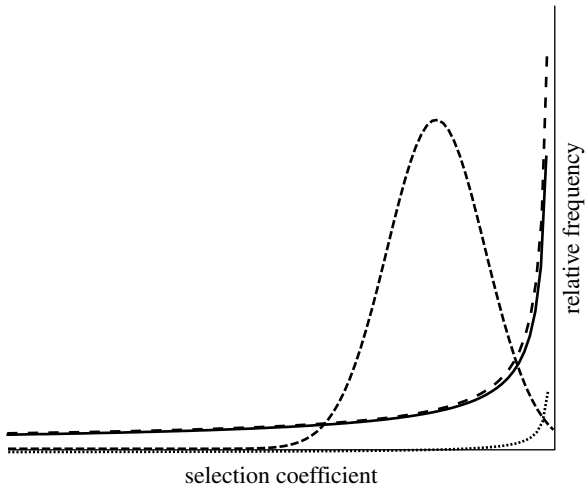

Figure 2. Example distributions of fitness effects estimated from different datasets, including lognormal for *Drosophila miranda* and *Drosophila pseudoobscura* (Loewe & Charlesworth 2006; dotted curve), strongly leptokurtic gamma for *Drosophila melanogaster* (solid curve) and human (spaced dashed curve) nuclear genes (Keightley & Eyre-Walker 2007), and normal for primate mitochondrial genes (Nielsen & Yang 2003; closed dashed curve).

strength of selection, a substantial proportion of mutations have fitness effects of the order of $1/N_e$ for many natural populations, and the effect of a change in population size on the rate of molecular evolution is expected to be quite large. This model was modified by Kimura (1979) who proposed that negative selection coefficients followed a more leptokurtic distribution. For a given strength of selection, fewer mutations will typically fall in the range from $1/N_{eL}$ to $1/N_{eS}$ under this distribution, and so the difference in substitution rate between lineages with different N_e will also be less, although a negative correlation between N_e and fixation rate is still predicted.

Neither of these distributions were chosen on the basis of biological data (Gillespie 1991), but a number of empirical estimates of the distribution of fitness effects of deleterious mutations have recently been made. Results vary between datasets and between taxa, with the estimated distributions including normal (Nielsen & Yang 2003), lognormal (Loewe & Charlesworth 2006) and strongly leptokurtic gamma distributions (Keightley & Eyre-Walker 2007) (figure 2). These estimates are based on the data from relatively few species, but indicate that the distribution of mutant effects is likely to vary between taxa. Adding further complexity, recent experimental work has suggested that a species' distribution of fitness effects is dynamic, and may change as organismal fitness and/or effective population size change (Silander *et al.* 2007).

The prediction of increased rate of evolution in species with small N_e relies on the assumption that advantageous mutations are rare: positive selection is less efficacious in small populations, so fixation of advantageous mutations will be reduced rather than increased in species with low N_e. Slightly advantageous mutations are in fact likely to be relatively common (Charlesworth & Eyre-Walker 2007), but theoretical work that incorporates positive selection on such mutations shows that a negative correlation between overall rate of substitution and effective population size is still predicted (Ohta 1992). More problematically, some studies have suggested that, far from being rare, strongly advantageous mutations may comprise a substantial proportion of those mutations that contribute to substitution in humans and *Drosophila* (Eyre-Walker 2006), and this may further weaken the inverse relationship between N_e and substitution rate.

5. WHAT DO THE DATA SAY?

An increase in either overall substitution rate or ω in taxa with long-term low N_e has been shown for a broad range of species. For example, island endemic animal species, which are likely to experience a reduction in N_e compared with their mainland relatives due to both the bottleneck during island colonization and long-term restriction in range size, show significantly increased ω values (Woolfit & Bromham 2005). Endosymbiotic bacteria and fungi, which live within invertebrate hosts and undergo severe bottlenecks with each transmission to the next host generation, have higher substitution rates and values of ω than their free-living relatives (Woolfit & Bromham 2003; Moran *et al.* 2008). Also, hominids have higher values of ω, genome-wide, than other mammalian lineages with larger N_e (Kosiol *et al.* 2008).

We see the same patterns repeated across genomic regions that differ in N_e. Genes in regions of low recombination have reduced N_e due to Hill–Robertson interference, in which linkage between weakly selected loci reduces the efficacy of selection at any one locus (Hill & Robertson 1966); such genes show increased values of ω (Haddrill *et al.* 2007) and reduced fixation of beneficial mutations (Presgraves 2005).

By contrast, Charlesworth & Eyre-Walker (2007) have shown that lineages which have undergone an expansion in N_e may experience a transient, though potentially substantial, increase in substitution rate before the rate of evolution decreases to below the level it was before the increase in N_e. This temporary increase in substitution rate is due to the fixation by positive selection of slightly advantageous mutations that had previously been effectively neutral. They tested for such an effect in sequences from taxa that had probably undergone population expansion after colonizing the mainland from an island and found a significant increase in ω, supporting their prediction. Furthermore, Bachtrog (2008) recently analysed divergence data from 91 genes in two species of *Drosophila* that differ substantially in N_e, and found no evidence that N_e is a major determinant of the rate of adaptive evolution for these data, possibly due to recent changes in N_e or differences in the distribution of fitness effects of mutations between taxa.

6. WHAT NEXT?

It is clear that N_e may have substantial effects on the rates and patterns of nucleotide substitution, but predicting the precise form of those effects is far from simple. Nonetheless, some obvious implications for evolutionary analyses can be extrapolated from these results. For example, as even closely related species may differ substantially in N_e (e.g. Ramos-Onsins *et al*. 2004), assuming that changes in evolutionary rate along lineages are rare, is unlikely to be an appropriate model for estimating divergence dates. Similarly, when performing comparative analyses of selection in different lineages or genes, the possibility that variation in ω is due to differences in N_e must be considered alongside selective explanations.

To move beyond these caveats and begin to incorporate N_e into analyses of molecular evolution more quantitatively, we must obtain better estimates of the effective population sizes and distributions of fitness effects of both deleterious and advantageous mutations for many more taxa. Such analyses require substantial amounts of sequence data. Next-generation sequencing technology is making this increasingly tractable, although the effort involved in both sample collection and computational analysis of the data is likely to remain substantial. The return on investment would be great, however, as estimates of these parameters are essential not only to fully understand this major driver of molecular rate variation, but to answer questions in a host of other evolutionary fields ranging from conservation biology to quantitative genetics (Keightley & Eyre-Walker 2007).

I thank the editors and three anonymous reviewers for their perceptive and extremely helpful comments on the manuscript.

Megan Woolfit*
School of Biological Sciences, University of Queensland, Brisbane 4072, Australia
*m.woolfit@uq.edu.au

Bachtrog, D. 2008 Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster*, two species with different current effective population sizes. *BMC Evol. Biol.* **8**, 334. (doi:10.1186/1471-2148-8-334)

Chamary, J. V., Parmley, J. L. & Hurst, L. D. 2006 Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108. (doi:10.1038/nrg1770)

Charlesworth, B. 2009 Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205. (doi:10.1038/nrg2526)

Charlesworth, J. & Eyre-Walker, A. 2007 The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proc. Natl Acad. Sci. USA* **104**, 16 992–16 997. (doi:10.1073/pnas.0705456104)

Eyre-Walker, A. 2006 The genomic rate of adaptive evolution. *Trends Ecol. Evol.* **21**, 569–575. (doi:10.1016/j.tree.2006.06.015)

Gillespie, J. H. 1991 *The causes of molecular evolution*. Cambridge, UK: Cambridge University Press.

Haddrill, P. R., Halligan, D. L., Tomaras, D. & Charlesworth, B. 2007 Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* **8**, R18. (doi:10.1186/gb-2007-8-2-r18)

Hershberg, R., Tang, H. & Petrov, D. A. 2007 Reduced selection leads to accelerated gene loss in *Shigella*. *Genome Biol.* **8**, R164. (doi:10.1186/gb-2007-8-8-r164)

Hill, W. G. & Robertson, A. 1966 Effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294.

Keightley, P. D. & Eyre-Walker, A. 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251–2261. (doi:10.1534/genetics.107.080663)

Kimura, M. 1979 Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl Acad. Sci. USA* **76**, 3440–3444. (doi:10.1073/pnas.76.7.3440)

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

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R. & Siepel, A. 2008 Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**, e1000144. (doi:10.1371/journal.pgen.1000144)

Loewe, L. & Charlesworth, B. 2006 Inferring the distribution of mutational effects in *Drosophila*. *Biol. Lett.* **2**, 426–430. (doi:10.1098/rsbl.2006.0481)

Lynch, M. & Conery, J. S. 2003 The origins of genome complexity. *Science* **302**, 1401–1404. (doi:10.1126/science.1089370)

Moran, N. A., McCutcheon, J. P. & Nakabachi, A. 2008 Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* **42**, 165–190. (doi:10.1146/annurev.genet.41.110306.130119)

Nielsen, R. & Yang, Z. H. 2003 Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol.* **20**, 1231–1239. (doi:10.1093/molbev/msg147)

Ohta, T. 1977 Extension to the neutral mutation random drift hypothesis. In *Molecular evolution and polymorphism* (ed. M. Kimura), pp. 148–167. Mishima, Japan: National Institute of Genetics.

Ohta, T. 1992 The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286. (doi:10.1146/annurev.es.23.110192.001403)

Petit, N. & Barbadilla, A. 2008 Selection efficiency and effective population size in *Drosophila* species. *J. Evol. Biol.* **22**, 515–526. (doi:10.1111/j.1420-9101.2008.01672.x)

Presgraves, D. C. 2005 Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr. Biol.* **15**, 1651–1656. (doi:10.1016/j.cub.2005.07.065)

Ramos-Onsins, S. E., Stranger, B. E., Mitchell-Olds, T. & Aguade, M. 2004 Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. *Genetics* **166**, 373–388. (doi:10.1534/genetics.166.1.373)

Silander, O. K., Tenaillon, O. & Chao, L. 2007 Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. *PLoS Biol.* **5**, 922–931. (doi:10.1371/journal.pbio.0050094)

Welch, J. J., Bininda-Emonds, O. R. P. & Bromham, L. 2008 Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol. Biol.* **8**, 53.

Woolfit, M. & Bromham, L. 2003 Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol. Biol. Evol.* **20**, 1545–1555. (doi:10.1093/molbev/msg167)

Woolfit, M. & Bromham, L. 2005 Population size and molecular evolution on islands. *Proc. R. Soc. B* **272**, 2277–2282. (doi:10.1098/rspb.2005.3217)

Wright, S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 97–159.