EVOLUTIONARY BIOLOGY

Fruitfly genome is not junk

Alexey S. Kondrashov

A comparison of two fruitfly genomes shows that much of their non-coding DNA is controlled by either negative or positive selection, dealing a double blow to the neutral theory of molecular evolution.

Once upon a time, the world seemed simple when viewed through the eyes of evolutionary biologists. All genomes were tightly controlled by various forms of natural selection. DNA encoded functional genes, and most mutations that occurred were rejected through negative selection. Those exceptional mutations that were beneficial substituted for the original gene variant (allele) and spread through the evolving populations by positive selection. And polymorphisms — where several alleles coexist within a population — were maintained by yet another, balancing, form of selection.

This idyllic world begun to crumble in 1968, when Kimura[1] made his modest proposal that most allele substitutions and polymorphisms do not substantially affect an organism's fitness and are governed, not by positive or balancing selection, but by random drift. Kimura still allowed for negative selection to eliminate most new mutations, so this proposal can be regarded as 'weak neutralism'. However, a decade later the onset of large-scale genome sequencing led to the discovery of genes that were so degraded as to be no longer functional (pseudogenes) and of other junk DNA. This led to 'strong neutralism', which claims that regions of genomes that do not encode proteins consist mostly of functionless DNA, ignored by all forms of selection[2]. Indeed, current estimates of the fraction of functionally important segments in mammalian non-coding sequences range from 10–15% (ref. 3) to just 3% (ref. 4).

On page 1149 of this issue, however, Andolfatto[5] reports a strikingly pre-neutralist pattern for a sample of fruitfly genomes, those of *Drosophila melanogaster* and *Drosophila simulans*. First, he compared non-coding nucleotide sites with synonymous sites — protein-coding sites where, because of the redundancies in the nucleotide triplet code, a substitution would not alter the encoded amino acid. The synonymous sites were used as a paradigm of neutrality, for want of a better one. Andolfatto found reduced levels of interspecies divergence and of intraspecies polymorphism within *D. melanogaster*, suggesting that around 50% of non-coding sites in *Drosophila* are affected by negative selection more strongly than synonymous sites. So because synonymous sites are also subject to some negative selection in *Drosophila*[6], most of the fly's non-coding sequences must be under some functional constraint.

Second, a substantial fraction of those nucleotide substitutions that do occur in non-coding *Drosophila* sequences are driven by positive selection. This conclusion follows from the results of the McDonald–Kreitman test, a statistical analysis that detects positive selection acting on certain kinds of nucleotide sites from the excess of substitutions, relative to polymorphisms, at these sites[7]. So, if Andolfatto's results are confirmed by further genome-scale analyses, neither strong nor even weak neutralism describes the evolution of the *D. melanogaster* and *D. simulans* genomes.

Can the neutral theory survive this double blow? Easily, because, unlike the fruitfly genomes, mammalian genomes are certainly full of junk. Although some originally junk sequences can be recruited to perform a function, and so become subject to selection[8,9], there is little doubt that substitutions and polymorphisms are, indeed, effectively neutral in the bulk of mammalian non-coding DNA.

At least two classes should therefore be recognized among the genomes of multicellular eukaryotes, which have long non-coding regions. Negative selection in mammals, and more generally in vertebrates, is so weak or inefficient (presumably because of their low effective population sizes) that even long segments of junk DNA often spread through the population. As a result, bloated and mostly neutral mammal-like genomes evolve.

By contrast, genomes of *D. melanogaster* and its close relatives (although not all of *Drosophila*), and probably those of many other species, are protected from the rampant accumulation of junk DNA by efficient selection. In *D. melanogaster*, individual transposable elements (jumping DNA segments) are mostly kept at low frequencies[10], introns (segments of non-coding DNA within genes) are comparatively short, and pseudogenes are few[11]. In such species, selection maintains lean and mostly functional melanogaster-like genomes.

Although the neutral theory will survive Andolfatto's demonstration that there is pervasive selection in *Drosophila*, its original justification may not. Kimura[1] claimed that most substitutions must be neutral because positive selection driving all of them would incur too high a fitness cost. However, it is now known that the 'lag load'[12] associated with even rapid adaptive evolution is not necessarily very high[13]. So Andolfatto's conclusion that one selection-driven substitution has occurred about every ten generations since *D. melanogaster* and *D. simulans* diverged seems reasonable theoretically. If flies can do it, so could others, invalidating Kimura's original argument.

In *Drosophila*, the relatively junk-free regions between genes, which probably regulate gene expression, seem to be a major target of positive selection. It is therefore fair to assume that the accumulation of beneficial mutations also had a major role in the evolution of functional segments in the intergenic regions in mammal-like genomes. However, nature is queerer than we may think. Since the human–chimpanzee divergence, functionally important intergenic segments have incorporated many deleterious mutations, rather than beneficial ones, perhaps because of the relatively recent decline of the effective population size in hominids[14]. We do not yet know how quickly beneficial mutations accumulate in functional segments of mammal-like genomes in the lineages where efficiency of selection did not decline.

The total number of functionally important nucleotides in the genome, T, is crucially important for estimating the genomic deleterious mutation rate U, a key parameter in evolutionary genetics[15]. Indeed, in a diploid organism, $U = 2T\mu$, and μ, the per nucleotide mutation rate, can be measured directly. Andolfatto's data and analysis suggest that, when U is estimated for *D. melanogaster*, μ must be multiplied by about 2×10^8. Recently, $\mu = 2 \times 10^{-8}$ was reported in the nematode worm *Caenorhabditis elegans*[16], and analogous data for *Drosophila* are expected soon. These data will have profound implications for a variety of outstanding problems in evolutionary biology, in particular with regard to the evolution of sex. It is truly amazing how little we know quantitatively about mutation and selection in the genomes of even the most well-studied organisms. ■

Alexey S. Kondrashov is at the National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA.
e-mail: kondrashov@ncbi.nlm.nih.gov

1. Kimura, M. *Nature* **217,** 624–626 (1968).
2. Li, W.-H., Gojobori, T. & Nei, M. *Nature* **292,** 237–239 (1981).
3. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. *Trends Genet.* **17,** 373–376 (2001).
4. Mouse Genome Sequencing Consortium. *Nature* **420,** 520–559 (2002).
5. Andolfatto, P. *Nature* **437,** 1149–1152 (2005).
6. Akashi, H. *Genetics* **144,** 1297–1307 (1996).
7. Smith, N. G. & Eyre-Walker, A. *Nature* **415,** 1022–1024 (2002).
8. Jordan, I. K. *et al. Trends Genet.* **19,** 68–72 (2003).
9. Silva, J. C. *et al. Genet. Res.* **82,** 1–18 (2003).
10. Bartolomé, C., Maside, X. & Charlesworth, B. *Mol. Biol. Evol.* **19,** 926–937 (2002).
11. Harrison, P. M. *et al. Nucl. Acids Res.* **31,** 1033–1037 (2003).
12. Maynard Smith, J. *The Evolution of Sex* (Cambridge Univ. Press, 1978).
13. Kondrashov, A. S. *J. Theor. Biol.* **107,** 249–260 (1984).
14. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. *PLoS Biol.* **3,** e42 (2005).
15. Charlesworth, B. & Charlesworth, D. *Genetica* **102/103,** 3–19 (1998).
16. Denver, D. R. *et al. Nature* **430,** 679–682 (2004).