NEWS & VIEWS

EVOLUTIONARY GENOMICS

Come fly with us

Ewan Birney

The genomes of 12 fly species have been analysed comparatively. Why should we care? Because sequences that have resisted the selective forces of evolution from fly to human must have functional significance.

Geneticists and molecular biologists have always had a soft spot for the fruitfly *Drosophila melanogaster*, for this innocuous organism is continually providing insights into the biology of multicellular organisms. But in the 1990s, when the nematode worm *Caenorhabditis elegans* became the life and soul of the genomics party, and we humans were always the guests of honour, flies were in danger of being left off the guest list.

This started to change when Celera Genomics[1] sequenced the genome of *D. melanogaster* as a trial, before tackling the genomes of larger species. Now, in the era of evolutionary genomics, the sequencing of 10, and comparative analysis of 12, fly species — reported in this issue[2,3] and elsewhere in more than 40 companion papers — means that flies have overtaken other species to become a favourite organism of genomicists too.

Every aspect of an organism emerges and persists through evolution. Consequently, researchers have always used evolutionary analysis to understand genomes, in particular to identify protein-coding genes that are conserved between organisms. But evolutionary processes can be studied far more effectively than by merely cataloguing the gene content of a genome. Specifically, researchers can investigate two complementary evolutionary aspects: negative selection and positive selection. Stark *et al.*[2] (page 219) study negative selection, or the presence of functional genomic elements that, despite having undergone many random mutational events, have not changed in function (Fig. 1a). By contrast, the *Drosophila* 12 Genomes Consortium (Clark and colleagues, page 203)[3] investigate positive selection, or the acquisition of new functions in different species (Fig. 1b).

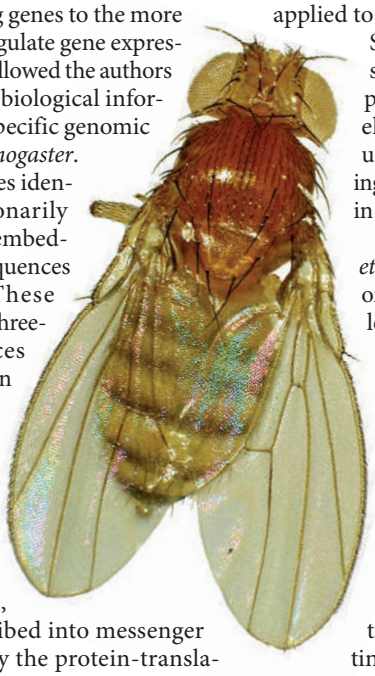
The remarkable diversity of fruitfly species makes them ideal organisms for such comparative analysis. Consequently, the authors studied closely related species such as *D. simulans* (pictured) and *D. sechellia* (which have a genetic distance equivalent to that between humans and closely related primates), as well as more distant drosophilids such as *D. grimshawi*. This is one of the many exotic Hawaiian species, and is physically 100 times bigger than its normal laboratory cousins, with a genetic distance between them equivalent to that between humans and lizards.

To discover functional elements and to refine our understanding of elements already known, Stark *et al.*[2] draw on most of our current knowledge of these elements, and use nature's own repertoire of mutations and selection. They consider all known classes of functional element — from the well-understood protein-coding genes to the more elusive motifs that regulate gene expression. These analyses allowed the authors to identify incorrect biological information ascribed to specific genomic sequences of *D. melanogaster*.

Stark and colleagues identify several evolutionarily conserved elements embedded in continuous sequences of coding DNA. These include stop codons (three-nucleotide sequences that signal termination of a protein sequence) and frameshift mutations, which throw the coding sequence out of step. It is hard to imagine that such gene structures — in which, for example, stop codons transcribed into messenger RNAs are ignored by the protein-translation machinery — are compatible with the normal rules of translation. So these findings strongly indicate the existence of additional, as yet unknown, mechanisms for the pre-translational processing of mRNAs, or alternative modes of translation.

MicroRNAs are short sequences of naturally occurring, single-stranded RNA that regulate gene expression. The authors next investigated genes for non-coding RNA sequences, such as microRNAs, and identify new microRNA sequences, thereby expanding the list of these regulatory sequences in *D. melanogaster* from 74 to 101.

Regulatory motifs are another type of functional element Stark *et al.* studied. The authors provide both an extensive 'dictionary' of such motifs and, for the first time in a genome-wide manner for an organism, a set of instances in which such motifs are putatively functional. Using genomics to identify cases of regulatory-motif activity is, indeed, an exciting new approach, and uses what the authors call 'branch length score'. This method takes into account the alignment and sequencing errors that are common in real data, and it can be applied to the whole of a phylogenetic tree.



S. J. CASTREZANA

Stark *et al.* carefully assess different statistical aspects of their method, providing a goldmine of functional elements that can be confidently used by molecular biologists studying flies and by laboratories interested in gene regulation.

Compared with the work of Stark *et al.*, Clark and colleagues' findings[3] on aspects of positive selection are of less direct use to molecular biologists working on *D. melanogaster*. Instead, their results provide for the first time a comprehensive set of genome-wide insights into how organisms arise during evolution. Statistically, the authors' analysis is not as powerful as that of Stark and colleagues. But this is not surprising, because their aim was to understand positive selection, which occurs in a non-continuous manner across the different *Drosophila* lineages, whereas the negative selection studied by Stark *et al.* is relatively constant and can easily be aggregated across the entire data set (Fig. 1).

Nevertheless, Clark and colleagues provide valuable insights into the evolution of *Drosophila* species. For example, by comparing the genomes of 12 drosophilid species, 10 of which they have sequenced and present in this issue, they show that, on both large and small scales, genomic rearrangements are extremely common in these genomes. They also find that about a third of the genes have undergone positive selection through mutations that affect the position of at least one amino acid. This suggests that positive selection

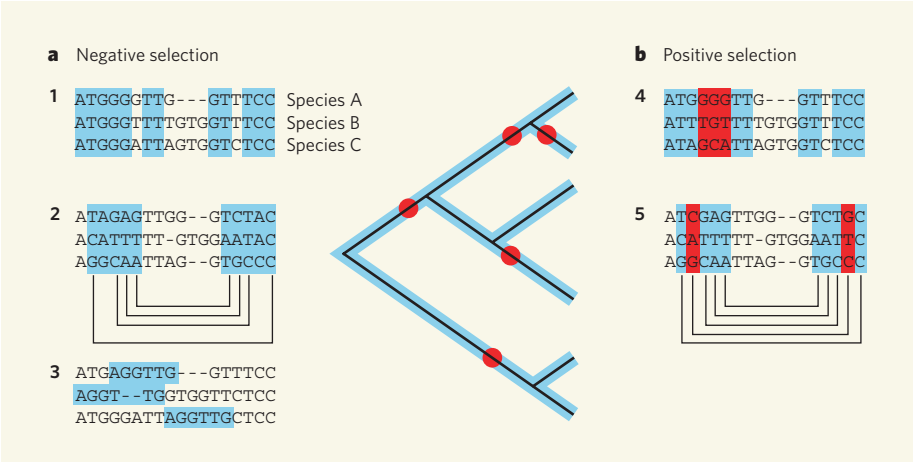


Figure 1 | Two types of evolutionary selection. **a**, For their analysis, Stark *et al.*[2] studied negative selection (blue), in which specific bases (from the four possible ones, A, T, C and G) remain roughly constant across the genomes of all lineages to ensure the conservation of functional genomic elements. Such an analysis uses three main methods: identifying conserved protein-coding sequences (**1**); identifying conserved paired bases in non-coding RNA genes (**2**); and identifying conserved specific motifs in the locale of the alignment (**3**). **b**, By contrast, Clark and colleagues[3] searched for cases of positive selection (red), which results in modification of specific bases in different species, leading to the acquisition of new functions. Two main methods are used to study positive selection: identifying fast-evolving codons embedded in a set of negatively selected codons (**4**), and searching for fast-evolving base pairs in the context of a non-coding RNA structure (**5**). The central tree, which indicates the phylogeny of the species (not all branches are shown), highlights the fact that, whereas negative selection is relatively continuous, positive selection is intermittent. Black lines in **2** and **5** show base pairing in the secondary structure. So although the positions paired may not show conservation on each blue column, the paired positions maintain a valid base pair.

occurs across many genes in a genome.

Codon-usage bias is the selective use by an organism of certain codons from a pool of codons that all specify a given amino acid, and it varies between different organisms. Clark and colleagues discover that, compared with other drosophilids, one species, *D. willistoni*, shows substantially reduced codon-usage bias.

The authors also show that genes encoding proteins involved in olfaction and immunity — the usual suspects for positive selection among protein-coding genes — have evolved faster than the rest of the genome. Rapid evolution was also seen in genes that regulate specific aspects of *Drosophila* physiology, such as insecticide resistance. During its long association with humans, *Drosophila* has endured radical changes to its environment, ranging from the introduction of insecticides to the transfer of species through human migration. We can therefore probably expect many interesting studies attempting to correlate genomic changes with such events in the fly's evolutionary history.

What are the broader implications of the findings of these two studies[2,3], particularly for further study of the human genome? In the case of negative selection, the evolutionary-genomics approach taken by Stark *et al.* clearly provides impressive insights into functional elements that are conserved across a clade (a group of related organisms). The proposed Mammalian Genome Project[4], which is well under way, is likely to have roughly the same statistical power as Stark and colleagues' data set. This would mean that we have a collection of powerful exploratory methods that can be applied to large-scale genomic analysis in mammals, and that are complementary to experimental techniques. In theory, there should be no qualitative difference in generating results for the *Drosophila* and mammalian clades using these methods. Nonetheless, the larger size of mammalian genomes, and the fact that there are potentially more fluctuations in the rate of neutral evolution, both across the genome of one species and between genomes of different species, may pose some interesting problems to be overcome.

Researchers are concerned that data obtained using methods based on evolutionary-genomic analysis do not entirely overlap with those obtained through experimental discovery methods, such as ChIP-chip and ChIP-seq, which generate comprehensive *in vivo* maps of transcription-factor binding sites and other functional DNA elements. In particular, these experimental techniques often define a set of elements that are not identified as conserved by the sensitive criteria of evolutionary-genomic analysis. As discussed previously[5] and by Stark *et al.*, this mismatch seems to be consistent across species and analyses performed by different laboratories. So it probably reflects our lack of understanding of how seemingly neutral evolutionary processes give rise to new, biochemically active elements before selection kicks in, rather than the existence of a large portion of lineage-specific elements, or defects in the methods used.

The analysis of positive selection by Clark and colleagues[3] is undoubtedly the broadest and most detailed investigation performed in any clade of multicellular organisms. Their study emphasizes the fact that, to understand differences between species, and thus how evolution leads to adaptive changes, we must improve the methods we use, and look at larger data sets and a broader range of species. This argument favours both sequencing the genomes of more species — now a realistic prospect given the advent of radically cheaper sequencing technologies — and determined efforts to carry out experimental studies on other members of each clade. Such studies are essential to any attempts to correlate sequence changes with changes in functional elements, and so test any new methods developed.

For the drosophilids, the next phase should entail sequencing the genomes of yet more fruitflies and other members of the order Diptera, thereby adding to the sequenced genomes of the drosophilids discussed here and their distant cousins, mosquitoes[6,7]. Moreover, more sequences should be generated at the population level — that is, we should sequence several individuals of the same species to gather the raw material for classical population-genetic analysis, which can be used for comparison with evolutionary data. Attempts to generate such resources are well under way for some drosophilid species. Finally, concerted efforts to obtain new experimental results in other species, beyond the experimental workhorse *D. melanogaster*, are needed for comparison with data obtained through evolutionary analysis.

Clark and colleagues' findings suggest that, to understand the fascinating adaptive changes among primates, including those unique to humans, we probably need to sequence the genome of every extant primate (and, where possible, any extinct primates with recoverable DNA), using optimal sequencing strategies to obtain both population-level data and accurate genome sequences. Basic molecular-biological studies on cell lines from selected primate species will also be needed to correlate sequence changes with changes in functional elements.

Returning to the present, the data presented and analysed by Stark *et al.*[2] and Clark and colleagues[3] provide the first significant example of the power of evolutionary genomics, which will be a central research theme for the next decade. It also means that genomicists can finally join their geneticist and molecular-biologist colleagues in the fruitfly fan club. ■

Ewan Birney is at the European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
e-mail: birney@ebi.ac.uk

1. Adam, M. D. *et al. Science* **287,** 2185–2195 (2000).
2. Stark, A. *et al. Nature* **450,** 219–232 (2007).
3. *Drosophila* 12 Genomes Consortium *Nature* **450,** 203–218 (2007).
4. www.broad.mit.edu/mammals
5. The ENCODE Project Consortium *Nature* **447,** 799–816 (2007).
6. Holt, R. A. *et al. Science* **298,** 129–149 (2002).
7. Nene, V. *et al. Science* **316,** 1718–1723 (2007).