Fruit Fly Family Fun

Arend Sidow[1,*] and Philippe Lacroute[1]
[1]Department of Pathology and Department of Genetics, SUMC L235, Stanford CA 94305-5324, USA
*Correspondence: arend@stanford.edu
DOI 10.1016/j.cell.2007.12.003

A recent comparative analysis of the sequenced genomes of 12 *Drosophila* species (*Drosophila* 12 Genomes Consortium, 2007; Stark et al., 2007) reveals a comprehensive picture of the evolution of small animal genomes and greatly improves computational predictions of functional elements in the *D. melanogaster* reference sequence.

One hundred years ago, a fly at Columbia University sustained a molecular lesion in its germline DNA that would turn its progeny's eyes white and its home into the founding lab of modern genetics (Morgan, 1910). Since then, the relatives of this fly, and the infinite number of induced or spontaneous mutations that have afflicted its population's gene pool, have been used by Morgan's intellectual descendants to elucidate one genetic principle after another (Rubin and Lewis, 2000). Now the starting gun has been fired for the second century of fly genetics with the recent publication of the comparative analysis of several sequenced *Drosophila* species (*Drosophila* 12 Genomes Consortium, 2007; Stark et al., 2007). Do these papers herald another hundred years equally rich in the discovery of genetic principles, or do they foreshadow more "Kuhnian Normal Science" (Kuhn, 1962), which is based on prevailing theory and mostly confirms known principles while working out details?

Each paper serves a distinct purpose. One emphasizes insights into changes in the genomes and in their functional elements (*Drosophila* 12 Genomes Consortium, 2007), whereas the other utilizes signals of evolutionary constraint to improve annotation of the *D. melanogaster* genome (Stark et al., 2007). That the comparative analyses of these flies could address both sides of evolution—change and conservation—attests to the wisdom of the group that initially proposed these species for sequencing, and to the versatility of the genus *Drosophila*. The evolutionary divergence among the sequenced *Drosophila* genomes spans a wide range, from the equivalent of interprimate distances all the way to human-reptile distances, generating power for analysis of both constraint and change.

Over the entire phylogenetic tree relating these species (Figure 1A), one expects some fairly dramatic differences in genome organization. Indeed, genome size spans a two-fold range and genome rearrangements, though mostly confined to inversions within chromosome arms, have steadily accumulated since the last common ancestor of the sequenced genomes. The result is that gene order is scrambled between now distantly related flies. Mobile elements tell their usual story of boom and bust, with some transposons having been active in all lineages, and others being lineage specific.

Functional parts of the fly genome also exhibit the expected mix of evolutionary change and stasis. Maintenance of telomeres appears to be achieved in all lineages by transposition of mobile elements rather than by telomerase, though the specific elements can differ depending on evolutionary divergence. Most noncoding RNAs and proteins

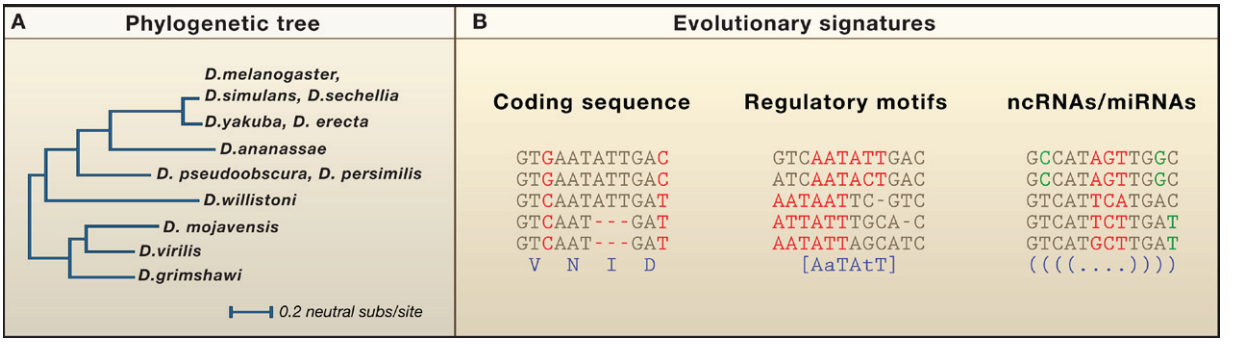

Figure 1. Phylogeny and Evolutionary Signatures in 12 *Drosophila* Genomes
(A) A compressed phylogenetic tree, drawn to scale with neutral divergence, to show the major lineages of the 12 sequenced *Drosophila* genomes.
(B) Key evolutionary signatures leveraged by Stark et al. (2007) to identify functional elements. For simplicity, only five hypothetical sequences are shown. Colored bases mark sites that have undergone diagnostic changes. Inferred structures of the functional elements (amino acid translation, regulatory motif consensus, and stem-loop structure of a noncoding RNA) are shown in blue. For noncoding RNAs, the tendency of microRNAs to allow changes in loops is indicated in red and the tendency of other noncoding RNAs to allow compensatory changes in stems is indicated in green.

have orthologs in all species, with only a relatively small fraction having undergone lineage-specific innovation or loss. Genes whose products interact more directly with the environment, such as those that function in detoxification, immunity, olfaction, and reproduction, appear to evolve at higher than average rates and some exhibit both increased gene turnover and evidence for positive selection. What emerges from this set of analyses is a coherent and comprehensive case study of how evolution shapes small metazoan genomes.

The second paper of the pair focuses less on lineage-specific differences than on conservation (Stark et al., 2007), leveraging the comparative power of the 12 *Drosophila* genomes to improve the annotation of the *D. melanogaster* genome (Adams et al., 2000). The fundamental premise in such studies is that conservation implies function: given a collection of genomes with sufficient neutral divergence to have nonfunctional DNA that is significantly changed, any conserved genomic feature is likely to be a functional element. One might start with a simple metric of conservation such as sequence similarity, but that is a crude measure. Protein-coding genes, RNA genes, transcription factor-binding sites, and other types of functional elements each have characteristic patterns of evolutionary change and conservation (Figure 1B). The combination of computational screens that target the "evolutionary signatures" of specific kinds of functional elements along with complete genome sequences spanning a wide range of evolutionary distances provides substantial power to detect and characterize new functional elements.

Searching for constrained elements requires a multiple-sequence alignment. An accurate alignment is essential but challenging to construct, making the new alignment of the 12 sequenced *Drosophila* genomes a valuable resource for this and future studies. The next step is to develop computational screens for particular types of functional elements. This study focuses on protein-coding regions, noncoding RNA genes, and regulatory sequences.

The screen for protein-coding regions relies on two independent signatures: (1) frequencies of exon-specific codon substitution, such as higher proportions of synonymous substitutions and conservative amino acid changes (Grantham, 1974; Kimura, 1977) and (2) reading-frame conservation due to insertion and deletion lengths that are multiples of three (Kellis et al., 2003). Compared to these two signatures, DNA sequence similarity is a less sensitive signal of protein-coding sequences, particularly because the third position of each codon is relatively less conserved. The gene prediction model built from these metrics predicts hundreds of modifications to existing annotations for *D. melanogaster* genes, many of them subsequently verified experimentally. Conservation provides additional power to detect short exons, a challenging task when only a single genome sequence is available. The method also predicts many unexpected gene structures, including conserved stop codons and frame shifts within an exon.

Noncoding RNAs have a very different signature. In functional RNAs the paired regions of the secondary structure are highly conserved and have a higher-than-average proportion of compensatory substitutions (Rivas and Eddy, 2001). The screen for noncoding RNAs includes these metrics as well as the predicted score for RNA folding. MicroRNA elements have a more specific signature of conservation that makes detection feasible despite their small size: the sequence in the stem of the precursor structure is highly conserved, one arm even more so than the other, whereas the loop tolerates more substitutions. These distinctive features lead to the first genome-wide predictions of noncoding RNAs in *Drosophila.*

Regulatory sites are among the most challenging functional elements to predict because of their short length and relatively high degeneracy. Stark et al. (2007) propose a conservation score for putative regulatory motifs based on the total branch length of the species tree containing aligned or nearly aligned instances of the motif. The motifs need not be precisely aligned so as to allow for movement of a motif relative to its target. A second signature distinguishes binding sites for transcription factors from posttranscription regulatory sites such as microRNA targets in 3′ untranslated regions: the latter exhibit higher conservation on the transcribed strand because only the transcribed RNA product is functional, whereas transcription factor-binding sites show little asymmetry between strands. These metrics predict not just new motifs but also specific binding sites, some subsequently verified by a chromatin immunoprecipitation assay. In total, the comparative analyses result in thousands of revised and new *D. melanogaster* annotations.

The resources provided by the comparative genome sequencing of these species, a summary of which is provided in these two papers and in several companion papers, will surely facilitate fly research for decades to come. In that sense, the studies themselves as well as the sequencing are very much an example of Normal (genome) Science, though a satisfyingly comprehensive one at that.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). Science *287*, 2185–2195.

Drosophila 12 Genomes Consortium (2007). Nature *450*, 203–218.

Grantham, R. (1974). Science *185*, 862–864.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Nature *423*, 241–254.

Kimura, M. (1977). Nature *267*, 275–276.

Kuhn, T.S. (1962). The structure of scientific revolutions (Chicago, IL: University of Chicago Press).

Morgan, T.H. (1910). Science *32*, 120–122.

Rivas, E., and Eddy, S.R. (2001). BMC Bioinformatics *2*, 8.

Rubin, G.M., and Lewis, E.B. (2000). Science *287*, 2216–2218.

Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., et al. (2007). Nature *450*, 219–232.