

Synchrotron Radiation Laboratories (beamline 9.2, M. Soltis), and European Synchrotron Radiation Facility (beamline ID29, W. Shepard) for beam time and help with data collection. This work is supported by the Pew Charitable Trusts through a Pew Scholar Award (C.S.R.), The Robert A. Welch Foundation grant AU-1524 (C.S.R.), NIH grant R01 AI054444 (C.S.R.), and Association pour la Recherche sur le Cancer (P.N.).

The GenBank accession number for *C. reinhardtii* sGC- β cloned in this work is AY343540. Coordinates and structure factors are available from the RCSB Protein Data Bank under accession code 1XBN.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1103596/DC1
Materials and Methods

Figs S1 to S10
Table S1
References and Notes

3 August 2004; accepted 27 September 2004
Published online 7 October 2004;
10.1126/science.1103596
Include this information when citing this paper.

Compensated Deleterious Mutations in Insect Genomes

Rob J. Kulathinal,¹ Brian R. Bettencourt,² Daniel L. Hartl^{1*}

Relatively little is known about the importance of amino acid interactions in protein and phenotypic evolution. Here we examine whether mutations that are pathogenic in *Drosophila melanogaster* become fixed via epistasis in other Dipteran genomes. Overall divergence at pathogenic amino acid sites is reduced. However, ~10% of the substitutions at these sites carry the exact same pathogenic amino acid found in *D. melanogaster* mutants. Hence compensatory mutation(s) must have evolved. Surprisingly, the fraction 10% is not affected by phylogenetic distance. These results support a selection-driven process that allows compensated amino acid substitutions to become rapidly fixed in taxa with large populations.

By mapping sequence space onto a fitness surface, the “fitness landscape” provides a powerful metaphor to understand how proteins evolve in populations. Sequence evolution may be visualized to traverse fitness peaks and valleys as certain sequence combinations are deleterious, advantageous, or neutral to an organism’s overall reproductive success. Using this landscape, Sewall Wright (1) championed the view that evolution occurs via epistatic or “coadapted” genetic interactions (2). In contrast, R.A. Fisher envisioned selection primarily acting on additive effects of individual loci (3). Whether epistatic interactions play an important role in the evolutionary trajectory of proteins—particularly in species with large effective population sizes—remains an open question.

The recently completed genomic sequence of *D. pseudoobscura* (4) offers a unique opportunity to approach this question by integrating the extensive mutant phenotypic information from *D. melanogaster* with a full set of orthologous gene sequences. These two fly species diverged between 40 and 50 million years ago (mya) and contain ample divergence information for comparative analyses. For nearly a century, *D. melanogaster* has been the target of extensive mutagenic screens, and a curated database of characterized mutant phenotypes and their corresponding molecular etiologies is freely accessible in the public domain (5).

We analyzed single amino acid residues that, when mutated in *D. melanogaster*, cause a

drastic fitness loss yet appear as the wild-type amino acid at its homologous site (termed “index site”) in the *D. pseudoobscura* protein. For such pathogenic substitutions to become fixed in another species, second-site or compensatory mutations must have coevolved (6). The number of compensated pathogenic deviations—or CPDs, using the parlance of Kondrashov *et al.* (6)—was compared to the number of index-site substitutions in *D. pseudoobscura* that contain an amino acid other than the one known to be pathogenic in *D. melanogaster* (Fig. 1). “%CPD” refers to the fraction of CPDs among substituted index sites. We sorted through all phenotypic mutants available from FlyBase (version 3.1) with the assumption that none of these mutants could persist in a natural population. From 35,311 “Gene” entries in FlyBase, we found 2245 single-site amino acid mutations that lead to a defined mutant phenotype, representing 525 unique genes. Most genes have between one and three different mutations; thus, our sample is not biased toward particular loci.

Of these genes, 475 had unambiguously aligned orthologs in *D. pseudoobscura* (4) representing 328,060 aligned amino acid sites. Overall, 77.8% of all amino acid sites were conserved between the two species. All instances in which a pathogenic amino acid site in *D. melanogaster* was present in the wild-type *D. pseudoobscura* protein were tabulated. To ensure site-specific orthology, we counted the number of conserved sites among 10 flanking amino acids on each side of the index site. Single-site insertion/deletion mutations were counted as one amino acid change. When a 50% identity criterion among flanking sites

was used, we found that among 1527 amino acid sites causing a phenotypic mutation in *D. melanogaster*, only 64 had substitutions in the *D. pseudoobscura* ortholog. Thus, substitutions among index sites were significantly less frequent than were random substitutions (95.8% versus 77.8%, $P < 0.0001$), an observation also reported in humans and mice (7), which suggests that selective constraints are maintained over time. Surprisingly, six of these index sites, or 1 in 10 substitutions at pathogenic sites, contained exactly the same amino acid that causes a deleterious phenotypic change in *D. melanogaster* [%CPD ~ 10% (Table 1)]. Similarly, using a 75% identity criterion (close to the average protein divergence estimated between these species), we found four CPDs among 31 index sites.

The same analysis was then applied to the more distantly related *Anopheles gambiae* genome (divergence ~250 mya). We aligned 210,778 orthologous amino acid sites (insertions and deletions ignored) from 317 proteins, with an overall 52.4% identity. When a 50% flanking region cutoff was used (corresponding approximately to the average amino acid divergence between these species), there were 77 substitutions among 784

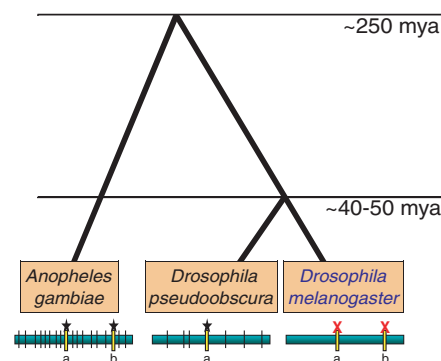


Fig. 1. Phylogenetic relationship of *D. melanogaster*, *D. pseudoobscura*, and *A. gambiae* and identification of compensated deleterious mutations. *D. melanogaster* is the reference species with site-specific mutational data on phenotypic mutants. Sites possessing amino acid residues that cause phenotypic deviations in *D. melanogaster* are called index sites (yellow bars). Other substituted sites are not informative (black vertical bars). We concentrate on index sites that contain substitutions (indicated as a and b) in at least one of the other species (denoted by a star). Index-site substitutions are of two types: the exact same pathogenic amino acid (CPDs) or other, non-exact amino acid substitutions.

¹Department of Organismic and Evolutionary Biology, ²Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed. E-mail: dhartl@oeb.harvard.edu

Table 1. Compensated amino acid substitutions from *D. melanogaster* versus *D. pseudoobscura* and *D. melanogaster* versus *A. gambiae*.

| <i>D. melanogaster</i> versus | Number of orthologs* | Protein identity | Index sites | Substituted amino acids | Exact amino acids substituted |
|-------------------------------|----------------------|------------------|-------------|-------------------------|-------------------------------|
| <i>D. pseudoobscura</i> | 475 | 77.8% | 1527 | 64 (4.2%) | 6 (9.4%) |
| <i>A. gambiae</i> | 317 | 52.4% | 784 | 77 (9.8%) | 7 (9.1%) |

*Phenotypic mutants with known single-amino acid mutations and identified orthologous sites. A 50% identity criterion among flanking sites was used in both comparisons.

index sites; 7 of these sites were CPDs. Although we again observed a significantly higher degree of conservation at index sites as compared to random sites in the *D. melanogaster*–*Anopheles gambiae* (*mel-Anoph*) comparison, index sites appear relatively less conserved over larger phylogenetic distance [*D. melanogaster*–*D. pseudoobscura* (*mel-pse*) 95.8% divergence versus *mel-Anoph* 90.2% divergence]. In contrast, among substituted index sites, the proportion with pathogenic amino acid changes remained remarkably constant over phylogenetic distance (%CPD ~ 10%; Table 1).

Using a somewhat different methodological approach on a set of 32 mammalian proteins, Kondrashov *et al.* (6) estimated that about 10% of all amino acid sites producing a pathogenic deviation in humans are present as the wild-type amino acid in at least one nonhuman mammal, independent of phylogenetic distance. Whereas we calculated %CPD from all possible and informative orthologs of three reference genomes, Kondrashov and colleagues used a large number of missense and nonsense mutations from a small but well-characterized subset of proteins to estimate %CPD among a variable set of organisms. We note that the quality of available mutational information is very different between the mammalian OMIM (Online Mendelian Inheritance in Man) and FlyBase databases, and the number of *Drosophila* genes with both missense and nonsense molecular information is insufficient to provide a comparable estimate of %CPDs using the same statistical methods as in Kondrashov *et al.* (6). From their results, the authors argue that, in mammals, nearby second-site compensatory mutation(s) must have rapidly coevolved with the index-site mutation to quickly traverse fitness valleys. Such valleys, as Wright (1) surmised, are important routes of evolutionary change in small isolated populations involving genotypic combinations that are deleterious in the presence of more fit neighboring genotypes. Their results effectively refute the hypothesis of the gradual and independent fixation of single compensatory mutations, because the accumulation of such modifiers would increase the fixation probability of the compensated deviations themselves, causing the fraction of CPDs among index sites to increase over phylogenetic distance.

Our results from three complete Dipteran genomes also reveal a high degree of compensatory mutational events. Epistatic interactions are evolving in insects, as the deleterious consequences of less-fit index site substitutions have become masked (or compensated) by at least one other mutation. By itself, the compensatory mutation(s) may have been neutral, beneficial, or even deleterious, and the series of mutational events could have occurred in either the *D. melanogaster* lineage or its sister lineages. Although the overall proportion of substituted index sites increases over phylogenetic distance, the %CPD remains constant and higher than expected. Therefore, compensated mutations must evolve frequently and regularly. We also found that most CPDs detected in this study were biochemically conservative amino acid changes appearing at a higher than expected frequency than other substitutions. Such conservative changes were observed at a significantly higher frequency among CPDs in the *mel-Anoph* lineage [$P = 0.029$ for *mel-Anoph*, $P = 0.071$ for *mel-pse* (Table 2)]. The conservative nature of substitutional change also suggests that compensatory mechanisms need not involve dramatic changes in fitness.

Using *A. gambiae* as an outgroup, we can also infer where in each gene tree specific mutational and compensatory events have arisen. Two protein-encoding genes, *carnation* and *Amylase proximal*, contain identified CPDs that are common to both the *D. pseudoobscura* and *Anopheles* data sets, indicating that molecular compensation must have evolved in the *melanogaster* branch (Fig. 1 and table S1). In these two cases, the ancestral amino acid at the index site causes a mutant phenotype in *D. melanogaster*, providing evidence for a deleterious mutational step in the compensatory process. One of these proteins, *Amylase proximal*, contains two CPD sites (T398A and D278N), each evolving in separate lineages. Four and five identified compensated sites appear to have evolved in the *D. pseudoobscura* and *Anopheles* lineages, respectively (table S1).

Even though Dipterans possess much larger effective population sizes than mammals (8), resulting in deleterious alleles persisting for shorter times due to efficient selection against them, we obtained a similar fraction of

Table 2. Composition of total substitutions versus compensated amino acid substitutions.

| <i>D. melanogaster</i> versus | Number of substitutions | Conservative |
|-------------------------------|-------------------------|-----------------|
| <i>D. pseudoobscura</i> | Total* | 583,423 45.3% |
| | CPD | 6 66.7% |
| <i>A. gambiae</i> | Total | 1,069,768 43.2% |
| | CPD | 7 71.4% |

*All available reciprocally best-hit protein orthologs.

compensated mutations as were found in mammals (6). These observations strongly support the contention that compensatory mutational evolution is independent of population size. When conditioned on other genetic backgrounds, the pool of mutations presently lethal in a *D. melanogaster* genetic background may become neutral (9, 10)—or possibly even advantageous—and rapidly become fixed by selection for favorable epistatic interactions. Such rapid fixation of compensatory events has been demonstrated in RNA secondary structure (11), and these results suggest an analogous process acting on protein structure. Recent analyses of selection at the molecular level also support the presence of considerable positive selection in *Drosophila* (12–14). In the future, it should be possible to combine mutational screens and fitness assays so as to finely dissect elements of compensation at the molecular level.

References and Notes

1. S. Wright, *Proc. Int. Congr. Genet. 6th* (1932), vol. 1, pp. 356–366.
2. T. Dobzhansky, *Genetics of the Evolutionary Process* (Columbia Univ. Press, NY, 1970).
3. R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
4. S. Richards *et al.*, *Genome Res.*, in press.
5. FlyBase Consortium, *Nucleic Acids Res.* 31, 172 (2003).
6. A. S. Kondrashov, S. Sunyaev, F. A. Kondrashov, *Proc. Natl. Acad. Sci. U.S.A.* 99, 14878 (2002).
7. R. H. Waterston *et al.*, *Nature* 420, 520 (2002).
8. R. K. Kliman *et al.*, *Genetics* 156, 1913 (2000).
9. M. Kimura, *J. Genet.* 64, 7 (1985).
10. W. Stephan, *Genetics* 144, 419 (1996).
11. C. O. Wilke, R. E. Lenski, C. Adami, *BMC Evol. Biol.* 3, 3 (2003).
12. J. C. Fay, G. J. Wyckoff, C. I. Wu, *Nature* 415, 1024 (2002).
13. N. G. C. Smith, A. Eyre-Walker, *Nature* 415, 1022 (2002).
14. S. A. Sawyer, R. J. Kulathinal, C. D. Bustamante, D. L. Hartl, *J. Mol. Evol.* 57, S154 (2004).
15. We thank members of the Hartl lab, especially D. Weinreich, for helpful discussions and the Computational Biology Group at the Bauer Center for Genomics Research for the use of their cluster. This work was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship (R.J.K.), a National Human Genome Research Institute (NHGRI) grant (P41-HG00739) to FlyBase (B.R.B.), and an NIH grant (GM068465) (D.L.H.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/1100522/DC1 Table S1

20 May 2004; accepted 22 September 2004
 Published online 21 October 2004;
 10.1126/science.1100522
 Include this information when citing this paper