

Adaptive genic evolution in the *Drosophila* genomes

Joshua A. Shapiro*, Wei Huang[†], Chenhui Zhang[†], Melissa J. Hubisz[‡], Jian Lu*, David A. Turissini*, Shu Fang*, Hurng-Yi Wang*, Richard R. Hudson*, Rasmus Nielsen[§], Zhu Chen^{†||}, and Chung-I Wu^{*||**}

Departments of *Ecology and Evolution and [†]Human Genetics, University of Chicago, Chicago, IL 60637; [‡]Department of Genetics, Chinese National Human Genome Center at Shanghai, Shanghai 201203, China; [§]Institute of Biology, University of Copenhagen, DK-1100 Copenhagen, Denmark; ^{||}State Key Laboratory of Medical Genomics and Shanghai Institute of Hematology, Rui-Jin Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, 201203, China; and ^{**}International Center for Evolutionary and Genomic Studies, Sun Yat-Sen University, Guangzhou 510275, China

Contributed by Zhu Chen, December 18, 2006 (sent for review August 28, 2006)

Determining the extent of adaptive evolution at the genomic level is central to our understanding of molecular evolution. A suitable observation for this purpose would consist of polymorphic data on a large and unbiased collection of genes from two closely related species, each having a large and stable population. In this study, we sequenced 419 genes from 24 lines of *Drosophila melanogaster* and its close relatives. Together with data from *Drosophila simulans*, these data reveal the following. (i) Approximately 10% of the loci in regions of normal recombination are much less polymorphic at silent sites than expected, hinting at the action of selective sweeps. (ii) The level of polymorphism is negatively correlated with the rate of nonsynonymous divergence across loci. Thus, even under strict neutrality, the ratio of amino acid to silent nucleotide changes (A:S) between *Drosophila* species is expected to be 25–40% higher than the A:S ratio for polymorphism when data are pooled across the genome. (iii) The observed A/S ratio between species among the 419 loci is 28.9% higher than the (adjusted) neutral expectation. We estimate that nearly 30% of the amino acid substitutions between *D. melanogaster* and its close relatives were adaptive. (iv) This signature of adaptive evolution is observable only in regions of normal recombination. Hence, the low level of polymorphism observed in regions of reduced recombination may not be driven primarily by positive selection. Finally, we discuss the theories and data pertaining to the interpretation of adaptive evolution in genomic studies.

McDonald–Kreitman test | selection | polymorphism

Recent studies based on DNA sequence data from large numbers of genes have increasingly suggested the prevalence of adaptive evolution in coding (1–5) as well as noncoding (6, 7) regions. The extent to which positive selection influences DNA polymorphism and divergence appears to be incompatible with the Neutral Theory of Molecular Evolution (8). This theory posits that the overall pattern of DNA evolution can be accounted for by mutation, genetic drift, and negative selection. It does not deny the operation of positive selection on some loci but only asserts that the overall pattern of genomic evolution can be explained without invoking adaptive evolution. Presumably, adaptive changes at any given time involve too small a fraction of the genome to be a statistically significant factor, despite their overwhelming biological significance.

The evidence used to test the Neutral Theory can be classified as divergence among species (9–11), polymorphism within species (12–14) or a combination of these (15, 16). The combined approach, as exemplified by the McDonald–Kreitman (MK) test and its derivatives, can separate the effects of negative and positive selection and is especially informative about adaptive evolution. Many such studies have concluded that positive selection may play a significant role in driving amino acid substitutions in the human and *Drosophila melanogaster* lineages (1–5).

However, as pointed out (4), the MK test requires all sites to be from the same genealogy (i.e., no recombination among sites). When there has been recombination between sites, the genealogies of each segment will differ, and the possibility arises that there will be an association between the rate of evolution in a segment and the

time to its most recent common ancestor. Such an association, even if weak, may substantially affect the results of the MK test. To the best of our knowledge, no report has attempted to estimate such a correlation, although variants of the MK test insensitive to such a relationship have been developed (3, 4). Hence, many estimates of adaptation by the MK test may require recalibration. For noncoding regions, recalibration may be an even more challenging task because of recombination between different types of sites.

Previous applications of the MK test were based on a relatively small sample of genes, which were often chosen without heeding possible biases. Indeed, earlier *Drosophila* studies contained a large fraction of rapidly evolving genes (many pertaining to male reproduction), which are more likely to be driven by positive selection than the average loci. To have an unbiased view of the average effect of positive selection on the entire genome, it is necessary to conduct an extensive survey of polymorphism and divergence with the following sampling considerations.

(i) Choice of genes should not have any overt bias. Criteria based on functional category, disease etiology, and prior knowledge of polymorphism or rate of evolution are likely to create biases. A strategy without biases is to choose genes by location and to cover a large part of the genome evenly. Furthermore, because the efficacy of selection depends critically on the rate of recombination (17), the contrast between loci from regions of low and high recombination will inform about the operation of selection.

(ii) The number of loci should be large enough to yield several thousand SNPs.

(iii) The number of individuals has to be large enough to reveal the difference in the frequency of deleterious and neutral polymorphism. It is also desirable to concentrate the sampling in the center of the species' origin for two reasons. First, the diversity of the species can be more accurately captured. Second, the confounding effects of drastic demographic changes associated with colonization can be minimized.

(iv) There should be multiple outgroup sequences to minimize misinference of ancestral states.

In this study, we report a survey of 419 loci on the autosomes of *D. melanogaster*. These loci are mainly from the third chromosome and are evenly distributed on that chromosome, which accounts for 40% of the *Drosophila* genome. We used 21 isogenic lines of *D. melanogaster*, 15 of them of African origin. For outgroups, we

Author contributions: J.A.S., Z.C., and C.-I.W. designed research; J.A.S., W.H., C.Z., and M.J.H. performed research; W.H., C.Z., S.F., and Z.C. contributed new reagents/analytic tools; J.A.S., M.J.H., J.L., D.A.T., S.F., H.-Y.W., R.R.H., R.N., and C.-I.W. analyzed data; and J.A.S. and C.-I.W. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: MK, McDonald–Kreitman; FI, Fixation Index; P/D, polymorphism/divergence; A/S, amino acid/silent changes.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. EF384276–EF392362).

|| To whom correspondence may be addressed. E-mail: zchen@stn.sh.cn or ciwu@uchicago.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0610385104/DC1.

© 2007 by The National Academy of Sciences of the USA

Table 1. Estimates of nucleotide diversity, Tajima's D, and standardized Fay and Wu's H for 419 autosomal loci, separated by recombination rate

Site class	θ_{Π}	θ_W	θ_L	D	H
Normal recombination, $n = 252$					
A	0.0019 (0.0024)	0.0024 (0.0028)	0.0020 (0.0024)	-0.585*	0.320*
S	0.0232 (0.0130)	0.0233 (0.0120)	0.0260 (0.0159)	-0.134	-0.311†
NC	0.0172 (0.0165)	0.0180 (0.0157)	0.0183 (0.0152)	-0.147†	-0.158
Low recombination, $n = 167$					
A	0.0013 (0.0021)	0.0014 (0.0020)	0.0014 (0.0020)	-0.468*	0.188†
S	0.0113 (0.0114)	0.0118 (0.0106)	0.0134 (0.0131)	-0.288†	-0.364
NC	0.0094 (0.0242)	0.0096 (0.0200)	0.0112 (0.0189)	-0.211†	-0.120

Low recombination is defined as <0.002 cM/kB. Unweighted means are shown, with the standard deviation among loci for estimates of θ given in parentheses. P values are for a two-tailed Wilcoxon rank sum test where the true mean is 0 A, amino acid changes; S, silent changes; NC, noncoding.

* $P < 10^{-8}$.

† $P < 10^{-2}$.

sequenced one line from each of the three sibling species of *D. melanogaster*: *Drosophila simulans*, *Drosophila mauritiana*, and *Drosophila sechellia*. For comparison, we also analyzed polymorphisms in the seven sequenced *D. simulans* genomes.

Results

Polymorphism. Summaries of polymorphism and nucleotide diversity for 419 autosomal loci are shown in Table 1. Genes in normal and low recombination regions are separately tallied, and sites are divided by mutation class into synonymous, nonsynonymous, and noncoding sites. The data presented are for all samples, which largely reflect African polymorphism. In general, the cosmopolitan samples showed lower levels of polymorphism than those from Africa, as reported (12, 18–20), with the bulk of cosmopolitan polymorphisms forming a subset of the African polymorphisms. Overall, the full combined sample was not significantly different from the African group alone [see supporting information (SI) Fig. 5]; we therefore chose to use the full set of samples for further analysis.

We used three measures for nucleotide diversity, θ_W , θ_{Π} , and θ_L , which estimate the population mutation rate per site, $4N\mu$, by using different weightings of low-, intermediate-, and high-frequency variants. At neutral equilibrium, all three measures should yield comparable values. The normalized difference between θ_W and θ_{Π} is the D statistic of Tajima (21), and the normalized difference between θ_L and θ_{Π} is the H statistic of Fay and Wu, as modified by Zeng *et al.* (13, 22). Although the separate estimates of polymorphism (θ_{Π} , θ_W , and θ_L) are quite comparable over these many loci, the average Tajima's D is somewhat negative for all mutation classes, both in the full dataset and in the African sample alone (data not shown), implying that even within Africa, *D. melanogaster* is not at the neutral equilibrium because of demography and/or selective effects.

Demography alone would be expected to affect patterns of polymorphism in the different classes of sites similarly. The contrast among classes of sites can therefore be informative about the specific effects of selection on each class. For example, the amino acid changing sites are only $\approx 1/10$ as variable as the silent sites, indicating strong selective constraint, as expected. Furthermore, they have significantly negative D and positive H, indicating a deficit of moderate- and high-frequency polymorphisms. Apparently, many nonsynonymous polymorphisms are deleterious and reach only low frequency. We also note that the noncoding regions (which consist mainly of intronic and 5' untranslated regions) are less variable than the synonymous sites of the coding regions (Mann-Whitney U; $P = 3 \times 10^{-9}$). These noncoding regions are likely under moderate selective constraint, as reported (7, 23, 24). Both silent and noncoding sites have negative D and H values, indicating that excess high-frequency variants in each class. (Negative D alone

can mean too many low- and/or high-frequency variants but negative D and H together should mean the latter.) This pattern of too many high-frequency variants may be indicative of either hitchhiking under positive selection (25) or population structure (26). We consider the second explanation less likely, because *D. melanogaster* populations are not in general strongly structured ($F_{ST} = 0.04$ between African and cosmopolitan samples). In addition, when the African samples are considered alone, the same patterns remain (data not shown). Nevertheless, polymorphism data are merely suggestive of positive selection. Other aspects of the data need to be carefully evaluated to form strong conclusions.

The Effect of Recombination on Polymorphism. An important feature of Table 1 is that the level of variation depends on recombination rate (12, 27). The full relationship for silent site variation in *D. melanogaster* is given in Fig. 1a ($P < 10^{-15}$; see figure legends). In contrast, the level of divergence between species (*D. melanogaster* vs. *D. simulans* group) is not influenced by recombination rate ($P = 0.28$; see SI Fig. 6). The cause(s) of this reduced level of variation in regions of low recombination has been subject to debate. On the one hand, hitchhiking with advantageous mutations is expected to reduce the level of variation so-called selective sweep, as recombination decreases (27–29). On the other hand, the background selection model (30) suggests that deleterious mutations, by reducing the effective population size in regions of reduced recombination, may also remove linked neutral variants.

Although background selection is clearly operating, the question is whether background selection can be rejected as the sole contributing force, or whether selective sweeps must be invoked to account for the observations. In Fig. 1b, we show that Tajima's D is also correlated with the recombination rate, indicating that lower recombination regions have an excess of rare variants, confirming earlier observations (27). Although this correlation has been suggested to support the selective sweep hypothesis, the effect is quite weak.

In regions of normal recombination, there should also be pockets of reduced variation created by selective sweeps (13, 28). Under background selection alone, the level of variation should be relatively homogeneous (31). We note in Fig. 1a that a substantial portion of the loci show very low silent-site diversity, even as the recombination rate increases, whereas others appear to have elevated polymorphism. To test the significance of this spread in nucleotide variation among loci with comparable recombination rates, we used the Hudson–Kreitman–Aguade (HKA) test, which compares the relative levels of nucleotide diversity among loci, calibrated against divergence. In this application, we used a variation of the HKA test that incorporates demographic parameters (17).

The demographic history of populations can have dramatic

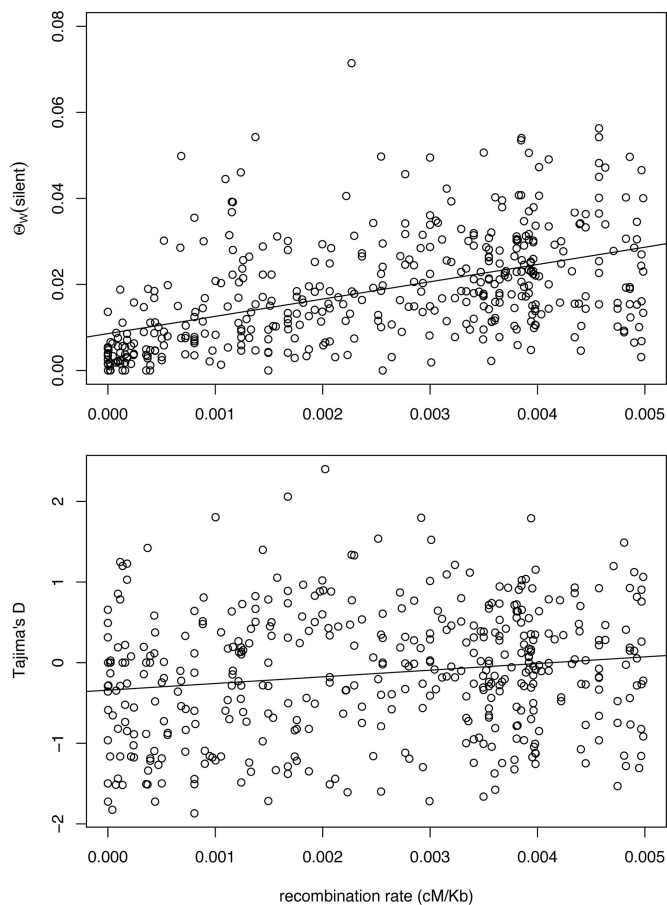


Fig. 1. Relationship between recombination rate and θ_w for silent sites (Spearman rank correlation $R = 0.53$, $P < 10^{-15}$) (Upper) and Tajima's D for silent sites ($R = 0.16$, $P = 8.6 \times 10^{-4}$) (Lower).

effects on the observed pattern of noncoding polymorphism, and recent work has suggested that much of the observed pattern of noncoding polymorphism in *Drosophila* can be explained by purely demographic models (18–20). To address this question, we fit a demographic model to the data under the assumption that the African and cosmopolitan flies represent two diverged populations with migration between them. We then estimated the number of loci with a significant deficit or excess of polymorphism relative to divergence by simulation under the inferred model. Of the 247 loci in the normal recombination region, 26 have significantly lower polymorphism than expected, whereas 31 appear to have an excess of polymorphism (SI Table 3). Although the excess polymorphism may be due to balancing selection, differential selection among populations, or weak background selection, the excess of divergence is far more suggestive of positive selection.

We also attempted to directly identify loci that had undergone a selective sweep by using the method of Nielsen *et al.* (32), which has been shown to be fairly robust to assumptions about demography. This method estimates the location and strength of a sweep at each locus, then uses simulation to determine whether that model fits the data better than a model without selection. Under this set of models, 33 of the 419 autosomal loci show significant evidence of selective sweeps, 23 of which are in the “normal” recombination region (of 252 in that region).

These tests support the idea that there has been a substantial amount of positive selection in the history of *D. melanogaster*. However, they remain potentially susceptible to demographic effects. Because the set of possible demographic models is enormous, we cannot completely exclude the possibility that we have simply

	P	D
A	30	100
S	30	100
ratio	1.0	1.0

+

	P	D
A	30	30
S	150	150
ratio	0.2	0.2

	P	D
A	60	130
S	180	250
ratio	0.33	0.52

Fig. 2. MK tables with $FI = 1$ can combine to produce a table of $FI > 1$, thus appearing nonneutral. Both hypothetical loci appear neutral as the ratio of nonsynonymous (A) to synonymous (S) changes within each table is the same for polymorphism (P) and divergence (D). The level of polymorphism between loci, however, is not the same, with a P:D ratio of 0.3 and 1.0 for the first and second tables, respectively. Although the different levels of polymorphism may be incompatible with strict neutrality, this rejection, if true, applies only to the polymorphic portion of the data. By contrast, the MK test is usually used to infer positive selection in the divergence among species.

not looked at the correct one. A better method is to use tests that have an internal control of sites known to be neutral.

Polymorphism vs. Divergence. Here, we contrast the patterns of amino acid divergence and polymorphism by using the MK test (15). The MK test determines whether there are more amino acid substitutions (DivA) among species than expected as compared with the level of the amino acid polymorphism (PolyA) within species. The ratio of silent site divergence (DivS) to polymorphism (PolyS) changes is used to calibrate the divergence/polymorphism ratio for neutral changes. We summarize the contingency tables of divergent/polymorphic and amino acid/silent changes and define the Fixation Index (FI) as $FI = (DivA/DivS)/(PolyA/PolyS)$. FI is expected to be equal to 1 in the neutral case when all polymorphic sites come from a single genealogy. When the observed FI is larger than the neutral expectation, positive selection is usually invoked (1–4).

Our sampling was designed to detect selection at the genomic level with as little bias as possible. Therefore, each individual gene segment was chosen not to maximize the sensitivity of detection but rather to collectively capture the overall signal of selection across the whole genome. In total, 33 of the 419 loci show a significant excess of amino acid substitutions between species (SI Table 4), but only two of these remain significant when corrected for a false discovery rate of 0.05 (33).

To extract the information from the entire collection of genes, we pooled data across loci and analyze the combined dataset (1, 2). Although combining contingency tables in the MK framework may yield a dataset that is intuitive and efficient, there are potential biases in this practice. The assumption of $FI = 1$ under neutrality, as applied in the MK test, is true only when all of the polymorphisms come from a single genealogy with little or no recombination among sites, or when all polymorphism is unlinked. When the nucleotide sites are derived from regions with different genealogies, caution should be exercised.

A simple example of this effect is shown in Fig. 2. Although the last table is the sum of the two neutral cases, the combination of the two produces an A/S ratio for divergence that is much larger than that for polymorphism. Under the standard interpretation of the

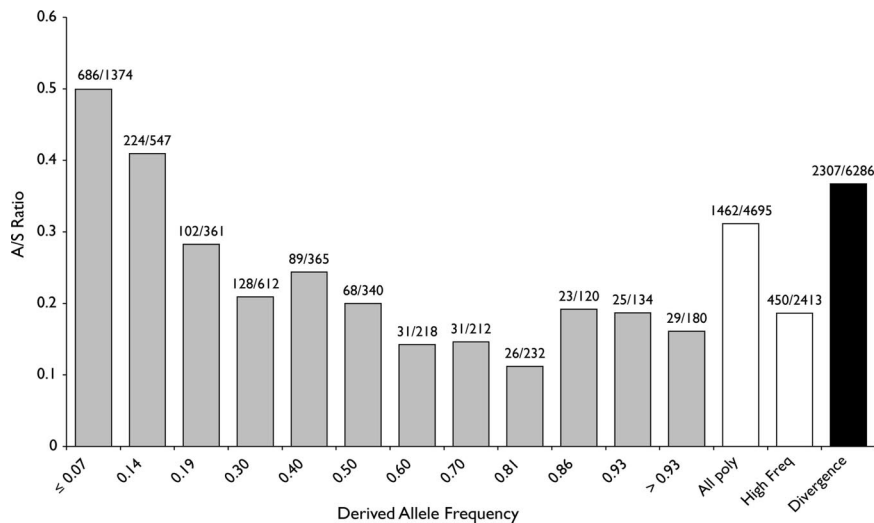


Fig. 4. Polymorphism A/S ratios in *D. melanogaster*, binned by the frequency of the derived variants. The first and last three gray bars represent singletons, doubletons, and tripletons. The middle six gray bars have a frequency increment of 0.1. The nine bars with frequency >0.19 are designated high-frequency classes (see *Results*). Their A/S ratios, being statistically indistinguishable, may be considered the ratio for neutral polymorphic variants. The A/S ratio for high-frequency variants, as well as that for all variants, is shown by the open bar. The black bar gives the divergence A/S ratio between *D. melanogaster* and *D. simulans*.

ratio of 0.186 ($=450/2,413$), whereas the A/S ratio for all sites is 0.311. We estimate that the proportion of nonsynonymous polymorphic sites that are deleterious and destined to be removed is $(0.311 - 0.186)/0.311 = 40\%$.

Similarly, we can calculate the proportion of amino acid fixations driven by positive selection. As shown in Table 2, the A:S ratio of the high-frequency polymorphic sites (A/S = 0.186) is significantly lower than the divergent A/S ratio (A/S = 0.367), even after correcting for biases resulting from combining tables as described [FI = 1.97 vs. E(FI) = 1.40; $P < 10^{-4}$]. The proportion of amino acid fixations driven by positive selection is therefore approximately $(1.97 - 1.40)/1.97 = 28.9\%$.

Effect of Recombination on Adaptive Divergence. The reduced level of nucleotide diversity in regions of low recombination (Table 1) has two very different explanations, background selection and selective sweep. Analyses of polymorphism data alone have not been able to delineate the two effects unambiguously. In Table 2, genes from regions of low and normal recombination are grouped separately. Interestingly, for low recombination regions, the A:S ratio for divergence is not much different from polymorphism. In contrast, the differential in the A/S ratio in regions of normal recombination is quite pronounced. The same pattern has been observed in previous studies (34, 35), and the implication is that positive selection is associated with genes in normal but not in low recombination regions. We shall return to this topic in *Discussion*.

Alternative Explanations for the MK Test Results. Positive results from the MK test do not necessarily indicate positive selection. If the effect of selective constraint is constant in divergence and polymorphism, positive selection is indeed a plausible explanation. However, there are at least two scenarios under which selective constraints may not be constant.

First, it is possible that the higher A/S ratio in divergent sites is the result of a reduced population size at some point in the divergence of *D. melanogaster* from the *D. simulans* clade. This would increase the number of effectively neutral mutations ($4N_e s < 1$), allowing more of the slightly deleterious amino acid changes to become fixed in the past. To address this concern, we estimated the ancestral effective population size. Takahata and Satta (36) and many others (37–39) showed that neutral divergence across loci between two closely related species can be informative about the level of polymorphism at the time of speciation. The maximum likelihood estimate of θ at silent sites in the common ancestor of *D. melanogaster* and *D. simulans* is 0.0396 per base pair (K. Zeng, personal communication), which

is larger than the extant level of polymorphism. Moreover, the cosmopolitan sister species, *D. simulans*, is comparably polymorphic with *D. melanogaster*. The two species of *Drosophila* therefore do not appear to show evidence of a population size smaller than the current size since the time of speciation.

Second, it is possible that selection intensity fluctuates for reasons unrelated to population-size changes. For example, when a species colonizes a new territory, selective constraint may change for some, but not all, loci. As a result, the reduced A/S ratios for some loci in one species, relative to the A/S ratio between species, may reflect changes in the strength of selective constraints, rather than the operation of positive selection. Comparing divergence with polymorphism in both species may alleviate, albeit not eliminate, that concern. This was an element of the original proposal of the MK test (15), but the practice has rarely been followed.

We therefore examined polymorphism data available from the *D. simulans* sequencing project. Because the sequencing was performed on seven separate lines of *D. simulans*, we can identify polymorphic sites from a homologous set of genes. We find the same patterns as in the *D. melanogaster* data. The A/S ratio for divergence is higher than for common polymorphism, suggesting that A/S ratio in the polymorphism of *D. melanogaster* is not at a historical low; thus, the level for divergence indeed appears high.

Discussion

Many studies have attempted to determine the extent of adaptive evolution at the genomic level (1–7) by contrasting divergence with polymorphism. There have been two major concerns about these studies: random gene sampling from the genome and the analysis of multiple loci simultaneously. We presented the criteria for random sampling in the Introduction and addressed the second issue extensively in *Results*.

The difficulty with the analysis of multiple genes in the MK test framework is that variants should either be on the same genealogy or completely unlinked, as with the Poisson Random Field model (40, 41). When there is partial linkage among sites, variation in P/D and A/S across linkage units (or genes) should be incorporated into the analysis. The negative correlation between P/D and A/S observed here is intriguing and deserves further investigation. It is plausible that the faster evolving genes (higher A/S) may have experienced more recent sweeps and hence are less variable (lower P/D). If true, previous conclusions about the presence of adaptive evolution in regions will likely stand, but the proof should be of a different kind.

By and large, previous conclusions of extensive adaptive evolution in the coding regions of *Drosophila* (1, 3–5) do turn out to be

correct. Recently, the MK test has been applied to noncoding (N) sites, pooling variants from unlinked genomic segments (7). Because the observed FI is only 1.2 in that study, we caution the reliance on an expected FI of 1 for interpreting deviation from neutrality. Even a weak correlation of P/D with N/S can generate a neutral expectation of FI > 1.2, leaving no excess of noncoding divergence.

Another noteworthy observation of this report is the contrasting MK test results between regions of low and normal recombination. Loci in regions of normal recombination bear the signature of adaptive evolution, whereas loci in regions of reduced recombination do not. In the discussions on background selection vs. selective sweep as an explanation for the effect of recombination on nucleotide diversity (Fig. 1*a*), the focus is generally on the influence of selection on the linked neutral variants. Below, we shall further consider the mutual influences among sites under selection.

Under the so-called Hill–Robertson (HR) effect (17), a reduction in recombination would tend to reduce the efficacy of selection. In regions of low recombination, deleterious mutations may thus linger longer and have a greater effect in removing polymorphism. In contrast, positively selected sites have a reduced probability of becoming fixed. As a result, regions of low recombination may be subjected to even stronger background selection but weaker hitchhiking, when the HR effect is in operation. This interference between hitchhiking and background selection has been partially analyzed (42) and deserves to be modeled more realistically in the future.

Our analysis indicates that adaptive positive selection on amino acid changes is prevalent in *D. melanogaster*. A large unbiased sample of genes and unbiased statistical analyses are both indispensable for reaching such a conclusion.

Materials and Methods

We sequenced 419 loci in 21 lines of *D. melanogaster* and one line each of *D. simulans*, *D. mauritiana*, and *D. sechellia*. A complete list of the lines sampled and their geographic origin is shown in SI Table 5. The sequenced segments consisted of 367 third-chromosome loci and 54 second-chromosome loci. Approximately one-third of the genes were selected from a set of genes broadly annotated as being involved in behavior, reproduction, and response to stimuli (43). Most of the remainder were chosen randomly from a list of all functionally annotated genes on the third chromosome, weighted by recombination rate so that coverage was denser in regions of higher recombination. The

final set ($\approx 10\%$) consisted of genes with no functional annotation, selected by the same random method. A complete list of loci sequenced appears in SI Table 6. Although these selection criteria do not conform to our stated ideal sampling strategy, differences between the genes selected by ontology and those selected randomly were small. In general, the ontology set showed less evidence of selection in our analyses than the remainder, making many of our conclusions conservative.

Population genetic statistics (θ_w , θ_{Π} , and θ_L) were calculated by using custom C++ software based on the libsequence package (44). Numbers of silent and replacement sites in an alignment and the values K_a and K_s were calculated according to the methods of Comeron (45). Recombination rates for each chromosomal band were assigned according to genome-wide estimates (46). Polymorphisms were assigned as ancestral or derived based on the most conservative outgroup (i.e., if a polymorphic site shared an allele with one outgroup and another with a different outgroup, the higher-frequency allele was assigned as ancestral).

Demographic parameters were estimated under a composite maximum-likelihood model. The model assumes that the African and cosmopolitan populations diverged T generations ago, with migration between the populations at rate m and a ratio of population sizes (Afr:Cos) r . Likelihood of the observed silent-site frequency spectrum was estimated for a grid of parameters, with one million coalescent trees simulated for each grid point (47, 48).

For more detailed methods, please see SI Text.

We thank D. J. Begun and C. H. Langley (Center for Population Biology, University of California, Davis) for permission to use the *D. simulans* polymorphism data and K. Thornton (Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY) and A. Eyre-Walker (Centre for the Study of Evolution and School of Life Sciences, University of Sussex, Brighton, U.K.) for providing software, code, and assistance with usage. M.-L. Wu assisted with fly stock maintenance and sample preparation. We also thank the associates from the Department of Genetics, Chinese National Human Genome Center at Shanghai, and Shanghai South Gene Technology Co., Ltd., for technical support. A. Eyre-Walker, P. Andolfatto, J. Fay, M. Przeworski, C. Bustamante, J. Zhang, M. Kreitman, and members of the C.-I.W. laboratory provided critical commentary and helpful discussions. This work was supported by National Institutes of Health grants and an OOCs grant from Chinese Academy of Sciences (to C.-I.W.) and by a predoctoral fellowship from the Howard Hughes Medical Institute (to J.A.S.). W.H. is supported by Chinese High-Tech Program (no. 863; Grants 2002BA711A10 and 2004CB518605) and by Shanghai Municipal Foundation for Sciences and Technologies (Grants 03DJ14008 and 04DJ14003).

- Fay JC, Wyckoff GJ, Wu C-I (2002) *Nature* 415:1024–1026.
- Fay JC, Wyckoff GJ, Wu C-I (2001) *Genetics* 158:1227–1234.
- Bierne N, Eyre-Walker A (2004) *Mol Biol Evol* 21:1350–1360.
- Smith NG, Eyre-Walker A (2002) *Nature* 415:1022–1024.
- Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL (2003) *J Mol Evol* 57 Suppl 1:S154–S164.
- Kohn MH, Fang S, Wu C-I (2004) *Mol Biol Evol* 21:374–383.
- Andolfatto P (2005) *Nature* 437:1149–1152.
- Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
- Yang Z (1998) *Mol Biol Evol* 15:568–573.
- Wyckoff GJ, Wang W, Wu C-I (2000) *Nature* 403:304–309.
- Tang H, Wyckoff GJ, Lu J, Wu C-I (2004) *Mol Biol Evol* 21:1548–1556.
- Begun DJ, Aquadro CF (1992) *Nature* 356:519–520.
- Fay JC, Wu C-I (2000) *Genetics* 155:1405–1413.
- Bamshad M, Wooding SP (2003) *Nat Rev Genet* 4:99–111.
- McDonald JH, Kreitman M (1991) *Nature* 351:652–654.
- Hudson RR, Kreitman M, Aguade M (1987) *Genetics* 116:153–159.
- Hill WG, Robertson A (1966) *Genet Res* 8:269–294.
- Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D (2003) *Genetics* 165:1269–1278.
- Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) *Mol Biol Evol* 22:2119–2130.
- Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005) *Genome Res* 15:790–799.
- Tajima F (1989) *Genetics* 123:585–595.
- Zeng K, Shi S, Fu Y-X, Wu C-I (2006) *Genetics* 174:1431–1439.
- Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD (2004) *Genome Res* 14:273–279.
- Bergman CM, Kreitman M (2001) *Genome Res* 11:1335–1345.
- Fay JC, Wu C-I (2001) *Curr Opin Genet Dev* 11:642–646.
- Przeworski M (2002) *Genetics* 160:1179–1189.
- Andolfatto P, Przeworski M (2001) *Genetics* 158:657–665.
- Maynard Smith J, Haigh J (1974) *Genet Res* 23:23–35.
- Begun DJ, Aquadro CF (1991) *Genetics* 129:1147–1158.
- Charlesworth B, Morgan MT, Charlesworth D (1993) *Genetics* 134:1289–1303.
- Hudson RR, Kaplan NL (1995) *Genetics* 141:1605–1617.
- Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) *Genome Res* 15:1566–1575.
- Storey JD (2002) *J R Stat Soc Ser B* 64:479–498.
- Sheldahl LA, Weinreich DM, Rand DM (2003) *Genetics* 165:1195–1208.
- Presgraves DC (2005) *Curr Biol* 15:1651–1656.
- Takahata N, Satta Y (1997) *Proc Natl Acad Sci USA* 94:4811–4815.
- Yang Z (2002) *Genetics* 162:1811–1823.
- Wall JD (2003) *Genetics* 163:395–404.
- Osada N, Wu C-I (2005) *Genetics* 169:259–264.
- Sawyer SA, Hartl DL (1992) *Genetics* 132:1161–1176.
- Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL (2002) *Nature* 416:531–534.
- Kim Y, Stephan W (2000) *Genetics* 155:1415–1427.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. (2000) *Nat Genet* 25:25–29.
- Thornton K (2003) *Bioinformatics* 19:2325–2327.
- Comeron JM (1995) *J Mol Evol* 41:1152–1159.
- Comeron JM, Kreitman M, Aguade M (1999) *Genetics* 151:239–249.
- Adams AM, Hudson RR (2004) *Genetics* 168:1699–1712.
- Nielsen R (2000) *Genetics* 154:931–942.