

23. Original questionnaire data are available from authors upon request.
24. Although the lowest response category for this question was 0 to 3 km, where respondents marked this category and indicated in other responses that there were major access points to the park, we assumed that there was a major road or river at or within the border (0 km).
25. A. N. James, M. J. B. Green, J. R. Paine, *Global Review of Protected Area Budgets and Staff* (WCMC, Cambridge, UK, 1999).
26. A principle components analysis on the correlation matrix of the factors showed little underlying structure among the attributes tested. The first axis accounted for only 19% of the variation in the data set, and the second axis for only 11%. The following variables were significantly correlated with the first component: total density of people in the park, total funding per hectare, number of guards per hectare, and total economic and development staff. Only one of these variables—number of guards per hectare—correlated with park effectiveness.
27. We thank the many people who contributed the information used in this report, those individuals from both government agencies and NGOs who coordinated responses for entire countries, and the Conservation International staff who managed this project in the field. We also thank P. Benson, T. Brooks, S. Edwards, C. Gascon, J. Ginsberg, J. Hardner, D. Repasky, A. Rylands, C. Short, and J. Waugh for valuable comments and discussions.

28 July 2000; accepted 20 November 2000

Chromosomal Effects of Rapid Gene Evolution in *Drosophila melanogaster*

Dmitry Nurminsky,¹ Daniel De Aguiar,² Carlos D. Bustamante,³ Daniel L. Hartl^{3*}

Rapid adaptive fixation of a new favorable mutation is expected to affect neighboring genes along the chromosome. Evolutionary theory predicts that the chromosomal region would show a reduced level of genetic variation and an excess of rare alleles. We have confirmed these predictions in a region of the X chromosome of *Drosophila melanogaster* that contains a newly evolved gene for a component of the sperm axoneme. In *D. simulans*, where the novel gene does not exist, the pattern of genetic variation is consistent with selection against recurrent deleterious mutations. These findings imply that the pattern of genetic variation along a chromosome may be useful for inferring its evolutionary history and for revealing regions in which recent adaptive fixations have taken place.

We have previously described the de novo evolution of a gene in the lineage of *D. melanogaster* (*1*). This gene, denoted *Sdic*, encodes a novel intermediate chain in a sperm-specific axonemal dynein. Changes that led to the creation of *Sdic* during the short evolutionary history of *D. melanogaster* [about 3 million years (*2*)] exhibit evidence for adaptive evolution. The gene was created from duplicated—and hence dispensable—copies of the genes for annexin X (*AnnX*) and the cytoplasmic dynein intermediate chain (*Cdic*). Three large deletions led to the fusion of the duplicated genes, whereupon a series of smaller deletions and nucleotide substitutions fashioned a new amino end of the *Sdic* polypeptide and created motifs characteristic of known axonemal dynein intermediate chains. The regulatory region of *Sdic*, including a spermatocyte-specific promoter element, also evolved from *AnnX* and *Cdic* sequences (*1*).

In principle, the evolutionary changes in *Sdic* could have taken place relatively rapidly

during and immediately following speciation (*3*). In this case, current selection pressure on *Sdic* should be mainly to eliminate deleterious mutations. However, *Sdic* still appears to be evolving rapidly, as evidenced by the fact that the ratio of replacement to synonymous polymorphisms is in excess of 2:1 [*1*] and additional data shown below].

The evidence for ongoing positive selection of *Sdic* prompted us to examine genetic variation in the surrounding genomic region to determine whether the theoretically predicted consequences of a rapid adaptive fixation (selective sweep) could be detected. The key issue is whether selection has been sufficiently recent and strong enough to yield a statistically significant deviation from the pattern of genetic variation that would be expected from nearly neutral polymorphisms affected only by random genetic drift, as well as selection against linked deleterious mutations [“background selection” (*4, 5*)]. Strong positive selection increases the frequency of a new favorable mutation and displaces linked nucleotide polymorphisms in the process (*6*). Theory predicts that a recent selective sweep should create a characteristic “trough” in the level of polymorphism in a region that includes the selected gene (*7*), as well as an excess of “singleton” polymorphisms (those present in only one sequence in the sample). On the other hand, theory also indicates that levels of polymorphism should be restored rel-

atively rapidly after a selective sweep. The time required for effective recovery of Tajima’s *D* (*8*) is approximately $2N$ generations, where *N* is the effective population number; in *D. melanogaster* $2N$ generations are about 80,000 years. Tajima’s *D* (*9*) is a conventional measure that compares the nucleotide diversity (pairwise differences) in a sample with the proportion of polymorphic sites, and it is negative when there is an excess of low-frequency polymorphisms, such as singletons.

To look for evidence of a selective sweep, we examined the spatial distribution of polymorphisms in the region at the base of the X chromosome that includes *Sdic* in *D. melanogaster*. The same analysis was carried out in the homologous region of the sibling species *D. simulans*, which lacks the *Sdic* gene. The pattern of polymorphism in *D. simulans* serves as a control, since there is no a priori reason to expect a recent selective sweep.

We sampled genes from polytene chromosome bands 18E1 to 20D. Messenger RNAs from 11 genes in *D. melanogaster* and 10 genes in *D. simulans* were reverse-transcribed, and the products were amplified by the polymerase chain reaction (PCR) and sequenced. Our analysis is based on an average of 903 base pairs per gene in each of 15 strains of *D. melanogaster* and 834 base pairs per gene in each of 7 strains of *D. simulans* (*10*). The analysis was confined to synonymous polymorphisms to eliminate possible artifacts due to different selective constraints or rates of amino acid replacement among the proteins.

To analyze the distribution of polymorphism along the chromosome, we used logistic regression. For each gene, let $W(x)$ be the number of segregating synonymous sites and $L(x)$ be the total number of synonymous sites in the sample. In these functions, x corresponds to the relative position of the gene in the chromosome. Under a simple model of background selection, the fraction of segregating sites, $S(x) = W(x)/L(x)$, should decrease monotonically as x moves from the euchromatin of the X chromosome toward and into the pericentromeric heterochromatin, owing to the progressive decrease in the rate of recombination and effective population size (*11*). The logistic regression model is used rather than an ordinary linear regression of S on x , because S is necessarily bounded on (0, 1). This feature favors anal-

¹Department of Anatomy and Cell Biology, Tufts University School of Medicine, 136 Harrison Avenue, Boston, MA 02111, USA. ²U.S. Department of Agriculture–Agriculture Research Service, Center for Medical, Agricultural and Veterinary Entomology, 1700 S.W. 23rd Drive, Gainesville, FL 32604, USA. ³Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02137, USA.

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

REPORTS

ysis of S rather than derivative quantities, such as θ (12), even though, in the present case, the logistic regressions would be equivalent because the sample size is the same for all genes.

Maximum likelihood was used for parameter estimation and hypotheses testing. For each gene, define $\hat{S}(x)$ as the predicted probability of polymorphism, given a logistic model as an n th order polynomial function in x :

$$\log \left[\frac{\hat{S}(x)}{1 - \hat{S}(x)} \right] = \sum_{i=0}^n \beta_i x^i$$

Under a model of background selection, with no effects of selective sweeps, the only parameters that should differ significantly from 0 are β_0 and β_1 , the intercept and coefficient of x , respectively. In contrast, if in the recent past there has been a selective sweep, then we would expect that a significantly better fit would be provided by a model with quadratic and cubic terms. These terms would reflect the expected trough in the level of polymorphism across the region, but higher order terms would not necessarily be expected to be significant.

The data from *D. simulans* are shown in Table 1 and those from *D. melanogaster*, in Table 2. The genes are listed in order along the X chromosome from distal to proximal with respect to the centromeric heterochromatin. The sequenced region of each gene is indicated, along with the number of synonymous nucleotide sites in the region and the observed number of polymorphic synonymous sites. For synonymous nucleotide sites, θ_{syn} is the nucleotide polymorphism, estimated from S and the sample size (θ); the parameter π_{syn} is estimated as the average proportion of pairwise differences per synonymous site. For neutral alleles in mutation-drift equilibrium, $\theta_{syn} = \pi_{syn} = 4N\mu$, where N is the effective population size and μ the nucleotide mutation rate (θ).

The results of the logistic regression analy-

ses are shown in the curves in Fig. 1, along with their 50% confidence bands (13). In *D. simulans* (Fig. 1A), we find a monotonic decrease in nucleotide polymorphism as the genetic markers approach the centromeric heterochromatin. This pattern has previously been described in regions of low recombination in *Drosophila* (14), which includes the centromeric heterochromatin, and has been attributed to increased efficiency of both selective sweeps and background selection in such regions (15). In *D. melanogaster* (Fig. 1B), the level of polymorphism shows the depression in a region near *Sdic* that we predict based on the evidence for positive selection of this gene. This trough in the level of polymorphism is consistent with a recent selective sweep in the region. A recent selective sweep is also implied by the frequency spectrum of the polymorphisms. For the 10 *D. melanogaster* genes with one or more polymorphic nucleotides (Table 2), 7 show an excess of singleton polymorphisms, indicated by the negative value of Tajima's D . Although there are so few polymorphisms that none of the individual values of D is significantly different from 0, across the region as a whole, a one-tailed nonparametric Wilcoxon signed-rank test for Tajima's D is significant ($P = 0.04$ for silent sites, $P = 0.01$ for all sites). In contrast, neither test yields a significant value of D for the data from *D. simulans* ($P = 0.44$ and $P = 0.28$, respectively).

Significance tests for the coefficients in the logistic regressions are given in Table 3. The test statistic is the difference in the log-likelihood of the data based on polynomial regressions of different order, which is approximately chi-square distributed with degrees of freedom equal to the difference in the order of the polynomials (16). For the *D. simulans* data, the linear regression is significant, and no higher order terms improve the goodness of fit. On the other hand, in the *D. melanogaster* regression, both linear and cubic terms are significant, and no higher order terms are significant. The cubic term is needed to fit the

trough of polymorphism in the *Sdic* region.

In summary, our prediction that *Sdic* has undergone one or more recent selective sweeps is supported by two independent features of the data. The first is the significant depression in the level of polymorphism near polytene chromosome region 19A (Fig. 1B), and the second is the frequency spectrum of polymorphisms skewed toward rare alleles including singletons (Table 2). Neither of these patterns is observed in the homologous chromosomal region in the sibling species *D. simulans* (Fig. 1A and Table 1). These analyses were based on silent sites

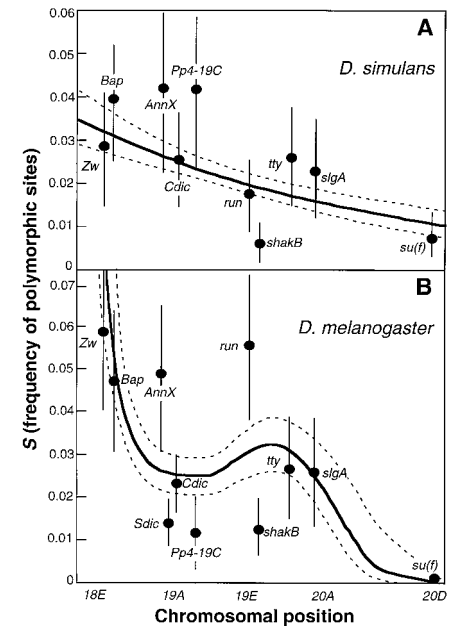


Fig. 1. (A) Results from *D. simulans* showing a monotonic decrease in the proportion of polymorphic sites (S) as a function of gene location at the base of the X chromosome. (B) Results from *D. melanogaster* showing a significant trough in the proportion of polymorphic sites (S) in the region around *Sdic*. The delimiter on each point is the approximate 50% confidence interval.

Table 1. Segregating synonymous sites in *D. simulans*.

Gene	Sequenced region (codons)	Synonymous sites (n)	Segregating synonymous sites			θ_{syn}	π_{syn}	D
			Observed (n)	Proportion of (S)	Expected ($n = 1$)			
<i>Zw</i>	130–354	140	4	0.029	5.41	0.012	0.012	-0.042
<i>Bap</i>	258–557	175	7	0.040	6.68	0.016	0.013	-1.210
<i>AnnX</i>	42–245	142	6	0.042	4.51	0.017	0.020	0.841
<i>Cdic</i>	42–150 and 423–614	228	6	0.026	7.05	0.011	0.011	0.156
<i>Pp4-19C</i>	29–259	168	7	0.042	4.78	0.017	0.019	0.686
<i>run</i>	72–380	226	4	0.018	5.29	0.007	0.007	-0.314
<i>shakB</i>	79–342	169	1	0.006	3.81	0.002	0.002	-1.007
<i>tty</i>	182–508	251	7	0.028	5.02	0.011	0.011	-0.347
<i>slgA</i>	103–394	204	5	0.025	3.67	0.011	0.012	0.707
<i>su(f)</i>	353–679	232	2	0.009	2.78	0.004	0.004	0.212
Total			49		49			
P value in χ^2 test					0.587			

REPORTS

Table 2. Segregating synonymous sites in *D. melanogaster*.

Gene	Sequenced region (codons)	Synonymous sites (n)	Segregating synonymous sites			θ_{syn}	π_{syn}	D
			Observed (n)	Proportion of (S)	Expected (n = 3)			
Zw	30–356	153	9	0.059	8.61	0.019	0.019	–0.131
Bap	249–568	149	7	0.047	7.62	0.015	0.013	–0.568
AnnX	44–245	142	7	0.049	3.51	0.015	0.010	–1.259
Sdic	9–201 and 302–497	286	4	0.014	6.89	0.004	0.004	0.062
Cdic	14–150 and 450–600	215	5	0.023	5.18	0.007	0.008	0.433
Pp4-19C	37–257	84	1	0.012	2.04	0.004	0.003	–0.330
run	36–380	159	9	0.057	4.89	0.018	0.016	–0.426
shakB	79–328	160	2	0.013	5.08	0.004	0.002	–1.470
tty	184–491	150	4	0.027	4.59	0.008	0.007	–0.372
slgA	69–394	115	3	0.026	2.57	0.008	0.008	0.147
su(f)	344–679	232	0	0	0.02	0	0	0
Total			51		51			
P value in χ^2 test					0.179			

Table 3. Results of log-likelihood ratio tests.

Comparison	Degrees of freedom	<i>D. simulans</i>	<i>D. melanogaster</i>
Linear fit versus constant	1	5.40 ($P < 0.02$)	10.55 ($P < 0.01$)
Cubic fit versus linear fit	2	0.20 ($P \approx 0.90$)	8.98 ($P < 0.02$)
Quadratic fit versus linear fit	1	0.18 ($P \approx 0.67$)	0.75 ($P \approx 0.39$)
Cubic fit versus quadratic fit	1	0.02 ($P \approx 0.89$)	8.23 ($P < 0.01$)
Higher order fit versus cubic fit	3	2.52 ($P \approx 0.48$)	0.87 ($P \approx 0.84$)

alone. Yet another indication of ongoing selection for *Sdic* is evident in the fact that *Sdic* accounts for 46% of the nonsynonymous polymorphisms but only 15% of the nonsynonymous sites ($P \approx 0.001$), whereas the level of synonymous polymorphism in *Sdic* is one of the lowest of the genes examined. Furthermore, nonsynonymous changes account for 70% of the *Sdic* polymorphisms, which is much higher than the average of 26% for other genes in *D. melanogaster* (17).

These findings confirm our prediction that the newly evolved *Sdic* gene has undergone one or more recent selective sweeps. The more general significance of the findings is the demonstration that natural selection for improved gene function may often be of sufficient magnitude to cause the level of polymorphism to be markedly reduced in or near the target of selection and to generate a distinctive frequency spectrum skewed toward rare alleles including singletons. Analysis of genetic variation across contiguous regions of the genome may therefore be a promising approach for identifying the locations of recently selected genes in *Drosophila* and other organisms.

References and Notes

1. D. I. Nurminsky, M. V. Nurminskaya, D. De Aguiar, D. L. Hartl, *Nature* **396**, 572 (1998).
2. A. Caccone, G. D. Amato, J. R. Powell, *Genetics* **118**, 671 (1988).
3. B. Charlesworth, D. Charlesworth, *Nature* **400**, 519 (1999).

4. D. Charlesworth, B. Charlesworth, M. T. Morgan, *Genetics* **141**, 1619 (1995).
5. R. R. Hudson, N. L. Kaplan, *Genetics* **141**, 1605 (1995).
6. J. Maynard Smith, J. Haigh, *Genet. Res.* **23**, 23 (1974).
7. W. Stephan, T. H. E. Wiehe, M. W. Lenz, *Theor. Pop. Biol.* **41**, 237 (1992).
8. M. Perlit, W. Stephan, *J. Math. Biol.* **36**, 1 (1997).
9. F. Tajima, *Genetics* **123**, 585 (1989).
10. The *D. melanogaster* sample included 15 isofemale lines isolated from Europe ($n = 2$), Japan ($n = 2$), North America ($n = 3$), South Africa ($n = 2$), and Zimbabwe ($n = 6$); the *D. simulans* sample included 7 isofemale lines isolated from Japan ($n = 1$), North America ($n = 2$), Africa ($n = 3$), and the Seychelles ($n = 1$). For preparing total RNA, individual males were homogenized in Trizol reagent (GIBCO BRL/Life Technologies, Rockville, MD). First-strand cDNA synthesis was carried out with an oligo(dT) primer using Superscript reverse transcriptase (Alkarni Biosystems, Berkeley, CA), and the RNA template was degraded with RNase H (GIBCO-BRL). Products of the reverse-transcription reaction were used directly as a template for PCR amplification in the presence of a 100:1 mixture of *Taq* polymerase (Boehringer Mannheim, Germany) and *Pfu* polymerase (Stratagene, La Jolla, CA) (18). Amplification was performed in Promega (Madison, WI) PCR buffer in the presence of 2 mM Mg^{2+} by using annealing temperatures calculated with Oligo 4.1 for each primer pair. The primer oligonucleotides were synthetic 20-nucleotide oligomers matching regions flanking the codons specified in Tables 1 and 2, and were chosen from the GenBank entries Zw (M26673, M26674), *Bap* (X75910), *AnnX* (M34069), *Sdic* (AF070688), *Cdic* (AF070699), *Pp4-19C* (Y14213), *run* (X56432), *shakB* (U17330), *tty* (AF184227), *slgA* (L07330), and *su(f)* (X62679). Except for *Sdic*, the amplification products were purified with Qiaquick columns (Qiagen, Valencia, CA) and sequenced directly with a 373A DNA sequencer using the Terminator Ready Reaction mix (PerkinElmer, Norwalk, CT). Because a number of easily recognizable variants of the *Sdic* gene are present in the tandem repeat and several of

these are expressed, the amplified *Sdic* cDNA was first cloned into pCR2 plasmid (Invitrogen, Carlsbad, CA), and individual clones were sequenced. Our analysis is based on the sequence of a single *Sdic* repeat that is expressed in all of the *D. melanogaster* stocks tested. Each sequence difference was verified from at least two independent clones.

11. B. Charlesworth, M. T. Morgan, D. Charlesworth, *Genetics* **134**, 1289 (1993).
12. G. A. Watterson, *Theor. Pop. Biol.* **7**, 256 (1975).
13. To generate 50% confidence intervals for $\hat{S}(x)$, we generated 10,000 parametric-bootstrap samples of the data and report the 2500th and 7500th ranked estimates of $\hat{S}(x)$ for each site in the sample.
14. D. J. Begun, C. F. Aquadro, *Nature* **356**, 519 (1992).
15. C. F. Aquadro, *Curr. Opin. Genet. Dev.* **7**, 835 (1997).
16. For the statistical analysis, we used Newton-Raphson iteration to find values that maximize the likelihood of the data under models either excluding or including a selective sweep, and used the maximum likelihood ratio test to ascertain whether the difference in fit is statistically significant. Under the logistic regression model, the likelihood function is

$$L(\beta_0, \beta_1, \dots, \beta_n | \text{data}) \\ \propto \prod_{j=1}^m S(x_j)^{W(x_j)} [1 - S(x_j)]^{L(x_j) - W(x_j)}$$

where m is the number of gene loci analyzed. To test if a p -order polynomial fits the data better than a q -order polynomial ($p > q$), we compare the likelihood functions under the two models. Let β_p be the vector of parameters in the full model, and let β_q and β_{p-q} be the vectors of estimates of the parameters under the full and restricted models, respectively. The statistic for the maximum likelihood ratio test, Λ , is given by

$$\Lambda = \log \frac{L(\beta_q = \hat{\beta}_q | \text{data}, \beta_{p-q} = \mathbf{0})}{L(\beta_p = \hat{\beta}_p | \text{data})}$$

where β_q represents the first q entries and β_{p-q} the last $p - q$ entries in the β_p vector, and $\mathbf{0}$ is a zero vector of dimension $p - q$. For data sets the size of ours, -2Λ is distributed approximately as chi-square with $p - q$ degrees of freedom.

17. E. N. Moriyama, J. R. Powell, *Mol. Biol. Evol.* **13**, 261 (1996).
18. W. M. Barnes, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 2216 (1994).
19. Supported by NIH grant GM 60035 to D.L.H. We thank R. Nielsen, S. Sawyer, and J. Wakeley for their ideas concerning the time to recover from a selective sweep, and J. Townsend and other members of the D.L.H. laboratory for helpful discussions.

26 September 2000; 20 November 2000