Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*

David J. Begun & Charles F. Aquadro

Section of Genetics and Development, Biotechnology Building, Cornell University, Ithaca, New York 14853-2703, USA

Two genomic regions with unusally low recombination rates in *Drosophila melanogaster* have normal levels of divergence but greatly reduced nucleotide diversity[1,2], apparently resulting from the fixation of advantageous mutations and the associated hitch-hiking effect[3,4]. Here we show that for 20 gene regions from across the genome, the amount of nucleotide diversity in natural populations of *D. melanogaster* is positively correlated with the regional rate of recombination. This cannot be explained by variation in mutation rates and/or functional constraint, because we observe no correlation between recombination rates and DNA sequence divergence between *D. melanogaster* and its sibling species, *D. simulans*. We suggest that the correlation may result from genetic hitch-hiking associated with the fixation of advantageous mutants. Hitch-hiking thus seems to occur over a large fraction of the *Drosophila* genome and may constitute a major constraint on levels of genetic variation in nature.

Table 1 summarizes levels of DNA variation and interspecific divergence (between *D. melanogaster* and *D. simulans*) where available. These estimates of DNA polymorphism are derived from restriction site surveys with one exception (*cubitus interruptus*) and therefore are estimates of average levels of variation over 13 to 65 kilobases (kb) from each gene region. To explore the relationship between levels of DNA sequence variation and recombination rates we compared estimates of nucleotide diversity (π)[5] and the coefficient of exchange[6], a measure of recombination rate per physical distance.

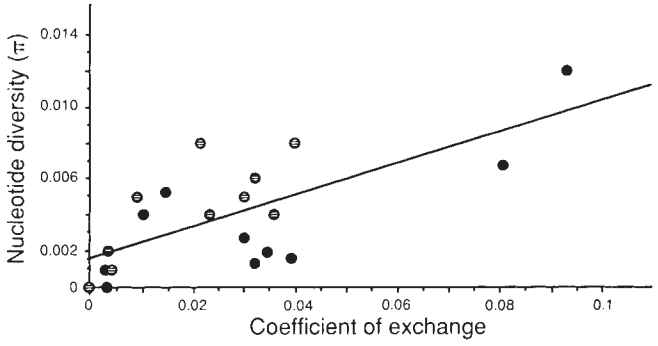

FIG. 1 Scatterplot of nucleotide diversity (π) versus coefficient of exchange in *D. melanogaster*. Autosomal and X-linked genes are represented by hatched and closed circles, respectively. To make autosomal and X-linked genes directly comparable we made the simplifying assumption of equal numbers of males and females. Then, under neutrality, nucleotide heterozygosity is an estimate of $3N\mu$ for X-linked genes and $4N\mu$ for autosomal genes (N is the effective population size and μ is the neutral mutation rate). Therefore, before doing regression or correlation analyses, we multiplied estimates of π from X-linked gene regions by four-thirds. The recombination rates estimated by the coefficient of exchange are for females. An X-linked gene region spends two-thirds of the time in females (where it can recombine) and only one-third of the time in males (where it cannot recombine), whereas an autosome spends half its time in females and half in males. Therefore, we multiplied the coefficient of exchange for autosomal genes and X-linked regions by one-half and two-thirds, respectively. Regression line is indicated by a solid line.

TABLE 1 Coefficients of exchange and nucleotide heterozygosities in *D. melanogaster* and divergence with *D. simulans*

Gene region	Coefficient of exchange	π	Divergence	Reference
Chromosome I (X)				
yellow-achaete (*y, ac*)	0.0045	0.001	0.054	1
phosphogluconate dehydrogenase gene (*Pgd*)	0.0154	0.003	0.029	1
zeste-tko (*z, tko*)	0.0222	0.004	—	12
period (*per*)	0.0520	0.001	0.050	1
white (*w*)	0.1400	0.009	—	13
Notch (*N*)	0.1212	0.005	—	14
vermilion (*v*)	0.0590	0.001	0.047	(D.J.B. and C.F.A., unpublished results)
forked (*f*)	0.0455	0.002	—	15
glucose-6-phosphate dehydrogenase gene (*Zw*)	0.0485	0.001	—	16
suppressor of forked (*su(f)*)	0.0050	0.000	—	15
Chromosome II				
sn-glycerol 3-phosphate dehydogenase gene (*Gpdh*)	0.0800	0.008	—	17
alcohol dehydrogenase gene (*Adh*)	0.0647	0.006	0.045	18, 19
DOPA decarboxylase gene (*Ddc*)	0.0184	0.005	—	(C.F.A. *et al.*, unpublished results)
amylase gene (*Amy*)	0.0435	0.008	—	20
Punch (*Pu*)	0.0718	0.004	—	17
Chromosome III				
esterase-6 gene (*Est-6*)	0.0604	0.005	—	21
metallothionein-A gene (*MtnA*)	0.0083	0.001	0.072	22
heat-shock protein-70A gene (*Hsp70A*)	0.0069	0.002	0.023	23, 24
rosy (*ry*)	0.0471	0.003	0.050	25
Chromosome IV				
cubitus interruptus Dominant (ci^D)	0*	0.000	0.050	2

Nucleotide diversity (π) is the average pairwise difference for all pairs of sequences drawn at random from a population, and can be thought of as heterozygosity per nucleotide[5]. The coefficient of exchange for a gene region was calculated by selecting two genetically defined loci[26] that flank the region of interest and dividing the distance in map units between the flanking loci by the number of polytene bands between the loci[6]. The number of polytene bands between loci was determined from Bridge's maps[27]. An important assumption underlying the use of this metric as an index of recombination rate is that over large stretches of the genome (for example, 20 to 40 polytene bands), the average amount of DNA per polytene band is roughly similar between regions. Available data suggest that this is a reasonable assumption for at least much of the *Drosophila* genome[6,28,29].

* The recombination rate on the fourth chromosome is effectively zero[30].

Figure 1 is a scatterplot of π versus the coefficient of exchange for 20 genes in *D. melanogaster*. It is apparent that levels of nucleotide diversity increase as rates of recombination increase. Variation in recombination rates explains a large fraction of the variation in nucleotide diversity and the null hypothesis that the slope is zero is rejected with high probability ($F_s = 16.8$, $P = 0.0007$). The non-parametric Spearman and Kendall regression tests are also significantly different from zero (Spearman's $D = 544$, $P < 0.01$; Kendall's $\tau = 0.437$, $P < 0.01$). The same conclusion is reached when data from the *white* region (which has a particularly high level of variation in *D. melanogaster*) is excluded from the analysis ($F_s = 5.8$, $P = 0.03$; Spearman's $D = 544$, $0.02 < P < 0.05$; Kendall's $\tau = 0.374$, $0.02 < P < 0.05$).

One hypothesis to explain this trend is that gene regions in areas of reduced recombination have lower neutral mutation rates. Perhaps recombination itself is mutagenic. If this were true, then under a neutral model these gene regions should also be less diverged between species than gene regions in areas of

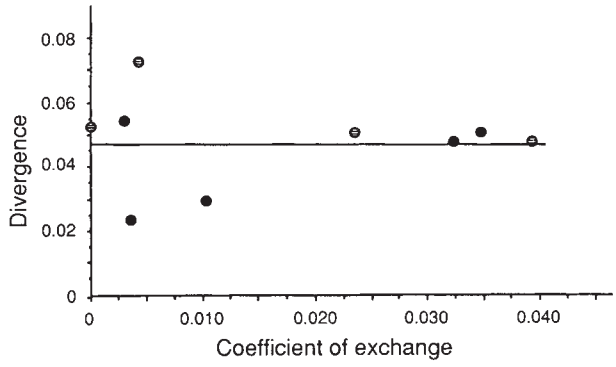

FIG. 2. Scatterplot of sequence divergence between *D. melanogaster* and *D. simulans* versus coefficient of exchange in *D. melanogaster*. Autosomal and X-linked genes are represented by hatched and closed circles, respectively. Coefficients of exchange for X-linked and autosomal regions are modified as described in Fig. 1 legend. Regression line is indicated by a solid line.

greater recombination rates[7]. In Fig. 2 we show a plot of divergence between *D. melanogaster* and *D. simulans* versus the coefficient of exchange, including all gene regions for which we have estimates of divergence. Clearly, estimates of divergence from a larger number of gene regions are desirable. Nevertheless, the lack of a significant positive regression coefficient ($F_s = 0.001$, $P = 0.983$) with the available data argues against the hypothesis that gene regions in areas of low recombination rates have, on average, lower substitution rates.

Theoretical results show that at the time of fixation of a neutral variant, the amount of linked neutral variation is reduced, and that the magnitude of the reduction depends on the recombination rate[8]. But at a random time (which is any time a genomic region in a population is sampled), the average amount of neutral nucleotide polymorphism is unaffected by the recombination rate[8,9]. Therefore, we are unable to arrive at a satisfactory neutral explanation for the patterns seen in Figs 1 and 2.

We propose that the positive correlation between DNA variation and recombination rate results from the selective fixation of advantageous mutants over a significant portion of the genome. This correlation suggests that levels of neutral variation in many of the gene regions for which variation has been measured have been reduced by one or more hitch-hiking events. Provided that a new selectively favoured mutation goes to fixation before another advantageous mutation arises close to it, each fixation will be surrounded by a 'window' of reduced polymorphism, the relative size of which is proportional to the rate of recombination for that region of the genome[4]. Thus, where recombination rates are very low, each fixation will cause a wide window of reduced polymorphism, whereas in regions of higher recombination, the window will be proportionately smaller. Moving along a chromosome towards regions of progressively lower recombination, the windows become closer, and may begin to overlap substantially. Thus, regions of low recombination are 'hit' by selective sweeps more often, keeping polymorphism at a lower average level. Mutations driven to fixation by meiotic drive or biased gene conversion would have similar evolutionary consequences. Hitch-hiking does not affect interspecific divergence[10], consistent with the observed lack of a correlation between DNA sequence divergence and recombination rate.

McDonald and Kreitman[11] proposed that patterns of synonymous and non-synonymous variation at the *Adh* locus in and between three *Drosophila* species are incompatible with neutrality. They suggested that selective fixation of amino-acid polymorphisms at this locus is the best explanation for their data. Furthermore, they speculate that selective fixations occur at a large number of loci. Our results are consistent with this view. However, the number of selectively favoured nucleotides relative to the size of the genome could still be quite small[4].

Clearly, hitch-hiking and recombination rates do not explain all of the heterogeneity in levels of variation across the *D. melanogaster* genome. Variation in mutation rate and functional constraint, as well as several different forms of selection, have roles in shaping local levels of DNA sequence variation. But the analysis presented here presents the first evidence that hitch-hiking driven by selective fixation of new mutations may constrain levels of nucleotide polymorphism over large portions of the *D. melanogaster* genome. Inference of effective population size from levels of DNA variation may be compromised by this phenomenon.

Much effort is being expended to assemble physical and genetic maps in several species. An unexpected benefit of these genome mapping projects is that it will be possible to examine whether correlations between recombination rates and levels of DNA variation are a general phenomenon in natural populations of other taxa, including humans. □

Received 19 December 1991; accepted 7 February 1992.

1. Begun, D. J. & Aquadro, C. F. *Genetics* **129,** 1147–1158 (1991).
2. Berry, A. J., Ajioka, J. W. & Kreitman, M. *Genetics* **129,** 1111–1117 (1991).
3. Maynard Smith, J. & Haigh, J. *Genet. Res.* **23,** 23–35 (1974).
4. Kaplan, N. L., Hudson, R. R. & Langley, C. H. *Genetics* **123,** 887–899 (1989).
5. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, 1987).
6. Lindsley, D. L. & Sandler, L. *Phil. Trans. R. Soc* B. **277,** 295–312 (1977).
7. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
8. Tajima, F. *Genetics* **125,** 447–454 (1990).
9. Hudson, R. R. *Theor. Populat. Biol.* **23,** 183–201 (1983).
10. Birky, C. W. Jr & Walsh, J. B. *Proc. natn. Acad. Sci. U.S.A.* **85,** 6414–6418 (1988).
11. McDonald, J. H. & Kreitman, M. *Nature* **351,** 652–654 (1991).
12. Aguadé, M., Miyashita, N. & Langley, C. H. *Molec. Biol. Evol.* **6,** 123–130 (1988).
13. Miyashita, N. & Langley, C. H. *Genetics* **120,** 199–212 (1988).
14. Schaeffer, S. W., Aquadro, C. F. & Langley, C. H. *Molec. Biol. Evol.* **5,** 30–40 (1988).
15. Langley, C. H. in *Population Biology of Genes and Molecules* (eds Takahata, N. & Crow, J. F.) 75–91 (Baifukan, Japan).
16. Eanes, W. F., Ajioka, J. W., Hey, J. & Wesley, C. *Molec. Biol. Evol.* **6,** 384–397 (1989).
17. Takano, T. S., Kusakabe, S. & Mukai, T. *Genetics* **129,** 753–761 (1991).
18. Langley, C. H., Montgomery, E. & Quattlebaum, W. F. *Proc. natn. Acad. Sci. U.S.A.* **79,** 5631–5635 (1982).
19. Aquadro, C. F., Deese, S. F., Bland, M. M., Langley, C. H. & Laurie-Ahlberg, C. C. *Genetics* **114,** 1165–1190 (1986).
20. Langley, C. H. *et al. Genetics* **119,** 619–629 (1988).
21. Game, A. Y. & Oakeshott, J. G. *Genetics* **126,** 1021–1031 (1990).
22. Lange, B. W., Langley, C. H. & Stephan, W. *Genetics* **126,** 921–932 (1990).
23. Leigh Brown, A. J. *Proc. natn. Acad. Sci. U.S.A.* **80,** 5350–5354 (1983).
24. Leigh Brown, A. J. & Ish-Horowitz, D. *Nature* **290,** 677–682 (1981).
25. Aquadro, C. F., Lado, K. M. & Noon, W. A. *Genetics* **119,** 875–888 (1988).
26. Ashburner, M. Drosophila *Genetic Maps* (*Drosophila* Information Service **69**, 1991).
27. Lindsley, D. L. & Grell, E. H. *Genetic Variations of* Drosophila melanogaster (Carnegie Institute, Washington DC, 1967).
28. Sousa, V. *Chromosome Maps of* Drosophila (CRC, Boca Raton, Florida, 1988).
29. Merriam, J., Ashburner, M., Hartl, D. L. & Kafatos, F. C. *Science* **254,** 221–225 (1991).
30. Hochman, B. in *Genetics and Biology of* Drosophila Vol. 1b (eds Ashburner, M. & Novitski, E.) 903–928 (Academic, New York, 1976).

ACKNOWLEDGEMENTS. We thank M. Nachman for comments and all members of our laboratory for discussion. This work was supported by the NIH and NSF.