Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 1800–1803, March 1993
Genetics

Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*

(crossing-over/DNA sequence analysis/restriction map analysis/hitchhiking effect)

CHARLES H. LANGLEY*†, JENNIFER MACDONALD‡, NAOHIKO MIYASHITA§, AND MONTSERRAT AGUADÉ¶

*Center for Population Biology and the Section of Evolution and Ecology, University of California, Davis, CA 95616; ‡Zoology Department, University of Washington, Seattle, WA 98195-0001; §Laboratory of Genetics, Faculty of Agriculture, Kyoto University, Kyoto 606, Japan; and ¶Departamento de Genètica, Facultat de Biologia, Universitat de Barcelona, 08071 Barcelona, Spain

Communicated by R. W. Allard, November 23, 1992

ABSTRACT **Levels of DNA sequence polymorphism at the suppressor of forked [*su(f)*] region in natural populations of *Drosophila melanogaster* and *Drosophila simulans* are estimated by restriction map analysis. *su(f)* is located at the base of the euchromatic portion of the X chromosome where the level of crossing-over per physical length is extremely low. In a survey of 55 alleles from three natural populations of *D. melanogaster*, only 2 restriction sites of 27 hexanucleotide and 108 tetranucleotide restriction sites scored are polymorphic. Among 103 alleles from three natural populations of *D. simulans*, just one polymorphic restriction site is found in 109 tetranucleotide-recognizing restriction sites scored. The few polymorphisms in these surveys yield estimates of per site heterozygosities (0.00, 0.0002, and 0.0005, respectively) at least a factor of 10 less than the average observed at loci located in regions of the genome with normal levels of crossing-over. Because under a broad category of models of molecular evolution (including the neutral theory) a correlation between levels of polymorphism and interspecific divergence is expected, the DNA sequence divergence is examined for the *su(f)* region. Contrary to the predicted correlation, the estimated divergence (0.12 substitution per silent site) is, in fact, greater than that observed at loci in regions of normal crossing-over. According to an alternative hypothesis (hitchhiking effect model) intraspecific polymorphism is swept out of the population in regions of the genome closely linked to rare but selectively favored variants as they quickly go to fixation; the rate of divergence is, however, unaffected by these rare hitchhiking events. Thus, the observed paucity of polymorphism and lack of correlation with divergence are in accord with the theory of the hitchhiking effect and several recent reports of polymorphism and divergence in other genomic regions with reduced crossing-over per physical length.**

A theory of the molecular genetic basis of evolution will make quantitative and qualitative predictions about divergence among species and about polymorphism among individuals within populations. The simplest theories—e.g., the neutral theory (1)—view polymorphism as a simple transient phase determined quantitatively by a single parameter (the product of the neutral mutation rate and 4 times the effective population size). Other models (having more parameters) view polymorphisms as evolutionary endpoints with much more adaptive significance than a neutral transient phase (2). One approach to the evaluation of these competing theories is to investigate the relationship between divergence and polymorphism in different regions of the genome. Can we find situations under which the relationship varies, and can we infer plausible explanations for the heterogeneity? Recently, the ability to compare polymorphism and divergence in the same units at the same loci came into the grasp of population geneticists. Indeed several studies of gene-sized regions have identified evidence for inconsistency between levels of polymorphism and divergence (3–9). On a much greater scale, that of a chromosome, another pattern may be emerging. Surveys of genes in regions where crossing-over per physical length is extremely low [yellow–achaete–scute complex (*y-ASC*) (8–10) near the telomere of the X chromosome and cubitus interuptus Dominant (ci^D) (7) on the fourth chromosome] have found that the amount of polymorphism is significantly reduced. Surprisingly, there is no evidence for reduced divergence (7–9). These results have been interpreted as the consequence of the "hitchhiking effect" under which rare variants that are selectively favored sweep through the population to fixation. A region around the favored variants will hitchhike along, leading to reduced polymorphism in that part of the chromosome that is tightly linked. The region affected is much larger if crossing-over is reduced. But this hypothesis predicts normal levels of divergence between species (11, 12). Indeed, the studies of the *y-ASC* and ci^D region support this prediction.

Our first goal is to examine the level of polymorphism in a third region of extremely reduced crossing-over per physical length—i.e., in the suppressor of forked locus, *su(f)*, at the base of the X chromosome (28). Specifically, the DNA sequence variation in the *su(f)* region is surveyed in samples of alleles from natural populations of *Drosophila melanogaster* and *Drosophila simulans*. We find very little polymorphism in both species. The second goal is determination of the relative (to other loci) divergence of the *su(f)* sequences since the last common ancestor of these two species. Surprisingly, we find the divergence at the upper range of previously reported values for various genes between these two species (9, 13).

MATERIALS AND METHODS

Lines. Sixty-four X chromosome isogenic lines of *D. melanogaster* (20 from North Carolina, 27 from Texas, and 17 from Fukuoka, Japan) as described (14) and 103 independently extracted chromosomes of *D. simulans* (52 from Barcelona; 26 from La Rábida, Huelva, Spain; and 25 from Tenerife, Canary Islands) as described (9) were used in the present study.

Cloning and Sequencing. A random genomic library in λGem11 of a *D. simulans* strain from Putah Creek, CA, was screened using the 6.4-kb *Bam*HI/*Xba* I fragment from *D. melanogaster* as probe (see Fig. 1, ref. 28, and GenBank

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†To whom reprint requests should be addressed.

accession no. X62679). After subcloning two overlapping fragments (3.9 and 4.4 kb) in pBluescript, a set of nested deletions was obtained as described (15) for each strand of each recombinant plasmid. Single-stranded DNA was obtained from each clone and sequenced by the dideoxynucleotide chain-termination method (16) using T7 DNA polymerase and either general primers in the pBluescript polylinker or specially designed primers within the insert. Both strands were completely sequenced.

Restriction Map Analysis. Seven hexanucleotide-recognizing restriction enzymes (*Bam*HI, *Eco*RI, *Hin*dIII, *Pst* I, *Pvu* II, *Sal* I, *Sac* I) and eight tetranucleotide-recognizing restriction enzymes (*Alu* I, *Dde* I, *Hae* III, *Hha* I, *Hpa* II, *Sau*3AI, *Sau*96I, *Taq* I) were used in the *D. melanogaster* survey, and the same eight tetranucleotide-recognizing restriction enzymes were used in the *D. simulans* survey. Procedures for hexanucleotide-recognizing and for tetranucleotide-recognizing analyses were as described (14, 17). For the survey of *su(f)* alleles in *D. melanogaster*, the 6.4-kb fragment was used as a probe in both the hexanucleotide- and tetranucleotide-recognizing restriction enzyme survey. Fifty-five of the 64 *D. melanogaster* alleles were included in the survey with tetranucleotide-recognizing restriction enzymes (18). For the tetranucleotide-recognizing restriction enzyme survey of polymorphism in the homologous region of the alleles of *D. simulans*, two overlapping fragments (described above) were used as probes. The optimal stringency precluded reliable scoring of certain sites (e.g., those in A+T-rich regions) in the *D. simulans* survey. These sites were excluded from the analysis.

Methods. Measures of polymorphism and divergence are as described (19–22). The consistency of polymorphism and divergence across loci with the predictions of the neutral theory of molecular evolution were evaluated by the HKA test (4). The results of this study of *su(f)* were compared to the region 5′ of *Adh* for which there are comparable data: 9 polymorphic of 414 sites scored among 81 *D. melanogaster* alleles and 210 differences in 4052 aligned sites between *D. melanogaster* and *D. simulans* (4).

RESULTS

Hexanucleotide-Recognizing Restriction Enzyme Polymorphism in *D. melanogaster*. Fig. 1 shows the six-cutter restriction sites surveyed in 64 lines. Also shown is the region probed in this survey. Note that many of the restriction sites scored lie outside the probed region. Much of the DNA flanking the probed region has been determined to be repetitive (ref. 28; K. O'Hare, personal communication). Remarkably, no restriction site polymorphisms were detected. This contrasts with typical surveys in regions of normal crossing-over where comparable studies revealed that the great majority of the DNA is unique and >10% of the sites are polymorphic (23). Large insertions were detected in several lines. Lines 26, 27, and 28 appear to have a related sequence inserted in the same region 5′ of *su(f)*. Lines 14, 31, and 43 all have insertions as depicted in Fig. 1. The number and distribution of large insertions are typical of most regions of *D. melanogaster* euchromatic genome (23).

Tetranucleotide-Recognizing Restriction Enzyme Polymorphism in *D. melanogaster*. The sample consists of 55 alleles from three populations for which 108 sites are scored. Two polymorphisms are detected. Each is rare and they are independent. Loss of the *Dde* I site at position 2242 occurs in 2 of 20 alleles from North Carolina and in 2 of 27 alleles from Texas. A *Taq* I site near position 974 is found in 2 alleles from the Texas sample. Estimates of the population variation from the pooled samples are quite low: nucleotide diversity (π) = 0.0002 and $3N\mu$ (θ) = 0.0005, where N is the population size and μ is the mutation rate to selectively equivalent nucleotides.

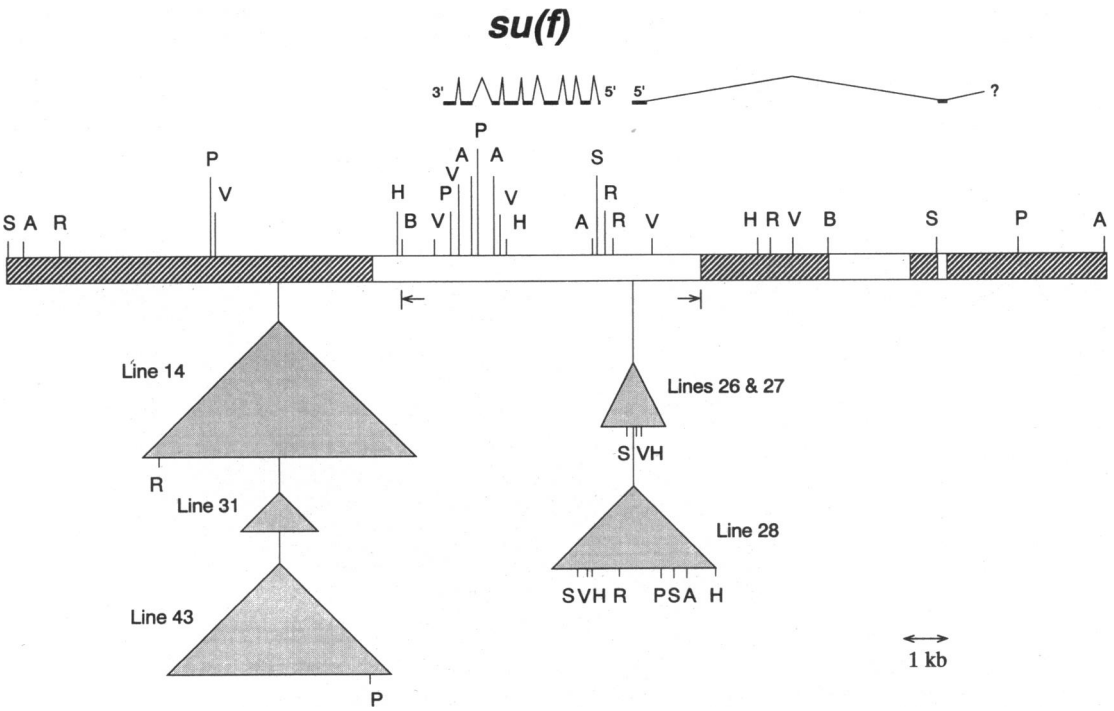

FIG. 1. Summary of results of hexanucleotide-recognizing restriction enzyme map survey of 64 *su(f)* alleles from three natural populations of *D. melanogaster*. (*Top*) *su(f)* transcription unit and part of an unidentified transcription unit (K. O'Hare, personal communication). Bar represents surveyed genomic DNA and positions of the scored sites (all monomorphic) are indicated (B, *Bam*HI; R, *Eco*RI; H, *Hin*dIII; P, *Pst* I; V, *Pvu* II; S, *Sal* I; A, *Sac* I). The 6.4-kb probe is indicated within the arrows. Hatched regions of bar represent segments known to contain repetitive sequences (K. O'Hare, personal communication). (*Bottom*) Insertions (with their approximate sizes and positions) and line numbers in which they were found.

Tetranucleotide-Recognizing Restriction Enzyme Polymorphism in *D. simulans*. One hundred and three alleles from three populations are scored for 109 sites. One polymorphism is detected. A *Taq* I site is lost at position 3031 (GenBank accession no. L09193) in 22 of 52 alleles from Barcelona, 7 of 26 alleles from Huelva, and 3 of 25 alleles from Tenerife. Estimates of π and θ from the pooled samples are 0.0005 and 0.0002, respectively.

Divergence Between *D. melanogaster* and *D. simulans*. Fig. 2 compares the sequence of *su(f)* in *D. simulans* (GenBank accession no. L09193) and *D. melanogaster*. Because of the large amounts of divergence in some regions, our alignment involves several rather arbitrary decisions. Our conclusions are not altered if alternative alignments are chosen. This particular alignment is available on request. Two important aspects of the divergence of these sequences since their last common ancestor are apparent. First, a great many insertions and deletions have accumulated in the noncoding regions, especially 5′. Second, the amount of substitutional divergence is not reduced as might be expected from the reduction in polymorphism described above. In the coding region, the average silent divergence is 0.102. This is typical of other genes for which a sequence is available (9, 13). No insertions and/or deletions are seen in the coding segments. Four amino acid replacement differences are found in the comparison of 734 codons (see Fig. 2). Fig. 2 shows that the 5′ flanking region of *su(f)* has diverged greatly. Over 13% of the alignable nucleotide positions are different in this region. Furthermore, our alignment indicates that 45 insertions/deletions distinguish the sequences in this region. This is at the high end of the range of observed divergence in noncoding sequences for these two species and certainly does not correlate with the lack of polymorphism in this region in both species. Finally, it should be noted that the lowest level of (silent) divergence in the entire region is in intron 4.

The application of the HKA test to the *su(f)* and 5′ *Adh* regions was based on 2 polymorphic of 603 sites surveyed among 55 *D. melanogaster su(f)* alleles and 412 differences in 3741 aligned sites between *D. melanogaster* and *D. simulans*. The test statistic X^2 for this test is 9.3 (with 1 degree of freedom); this is highly significant ($P < 0.005$) and largely attributable to the differences in numbers of segregating sites at the two loci.

DISCUSSION

The observations described above address both empirical and theoretical questions about the relationship between DNA sequence variation and chromosomal position. The initial studies of polymorphism in the *y–ASC* region suggested that polymorphism may be reduced near the telomere of the X chromosome in *D. melanogaster*. A subsequent survey of this region in *D. simulans* established a parallel reduction in that species (8–10). In *Drosophila ananassae*, average heterozygosity is reduced at the vermilion and furrowed loci, which are near the centromere of the X chromosome (24, 25). These regions have one property in common—i.e., reduced crossing-over per physical length. It was proposed that the drastic reductions in crossing-over per physical length [characteristic of centromeric and at least some telomeric regions of the chromosomes (26)] may allow the hitchhiking effect of even a few selected substitutions to "sweep away" DNA sequence polymorphisms throughout. A recent survey of the DNA polymorphism at ci^D on the fourth chromosome also found greatly reduced polymorphism in both *D. melanogaster* and *D. simulans* (7). The fourth chromosome is very small (essentially only centromeric–telomeric) and enjoys no crossing-over. The low level of polymorphism in the *su(f)* gene in both *D. melanogaster* and *D. simulans* reported here establishes further what is an empirical relationship of DNA sequence polymorphism and crossing-over.

One theoretical interpretation of this pattern of the little polymorphism in regions of reduced crossing-over per physical length is that the mutation rate to "evolutionarily acceptable" differences is lower in these regions because either the mutation rate itself is lower or there is more functional constraint on the DNA sequence. These hypotheses offer a

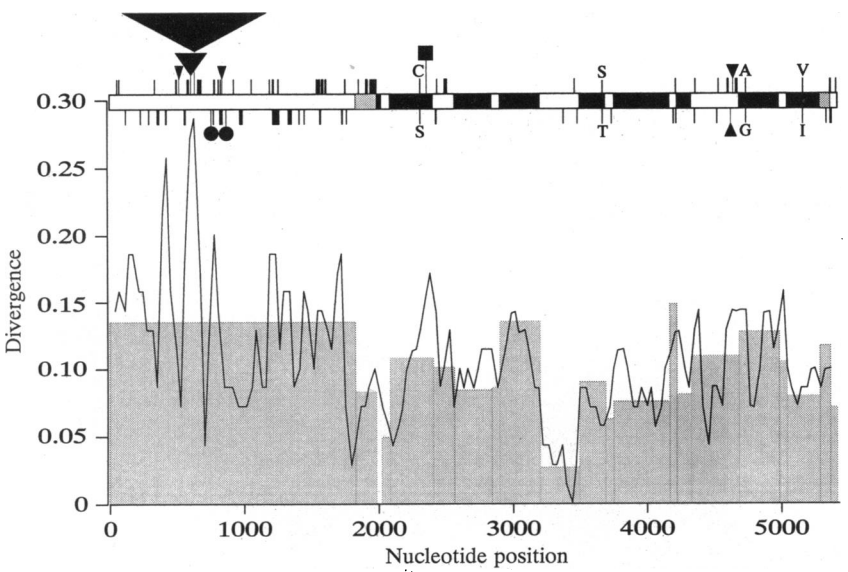

FIG. 2. Summary of interspecific divergence and of tetranucleotide-recognizing restriction enzyme map surveys of 55 *su(f)* alleles from three natural populations of *D. melanogaster* and 103 *su(f)* alleles from three natural populations of *D. simulans*. Bar represents aligned portion of the *su(f)* sequence from *D. melanogaster* and *D. simulans*. (GenBank accession nos. X62679 and L09193, respectively). Solid regions are exons, intervening open regions are introns, and hatched regions are untranslated portions (5′ left and 3′ right) of the *su(f)* transcript (28). Lines and triangles above and below the bar show positions and approximate sizes of unalignable sequences (insertions) in *D. simulans* (above) and *D. melanogaster* (below), respectively. The four amino acid replacements are indicated by the one-letter code (above is *D. simulans* and below is *D. melanogaster*). Positions of the three detected polymorphic restriction sites are indicated by solid circles (*D. melanogaster*) and square (*D. simulans*). Below bar, estimated silent divergence in a 70-bp sliding window (plotted every 35 bp) is plotted against position in the aligned sequence of *su(f)*. Shaded rectangles represent average silent divergence in different functionally distinct segments of the *su(f)* region.

simple and direct test. They predict that the divergence between species should also be reduced. This is clearly not the case. For *y–ASC*, ci^D, and now for *su(f)*, there is no suggestion of any reduced divergence relative to genes in regions of normal crossing-over per physical length. Indeed, the observed silent divergence in and around *su(f)* (0.12) is the highest so far reported between these two species. The proposed alternative hypothesis is that the reduced polymorphism is due to a hitchhiking effect of selected substitutions at linked loci (10–12). The hitchhiking effect model makes a different prediction about the expected divergence; it predicts no particular change in the rate of substitution of unselected mutants. Clearly, this hypothesis is consistent with the data so far gathered.

If this empirical and qualitative relationship between large reductions in polymorphism and a lack of observable crossing-over is accepted and if the hitchhiking effect model is considered the most viable hypothesis, how can we test it more rigorously? One approach is to examine the distribution of the levels of polymorphism in many regions with various amounts of crossing-over per physical length. This way the simple average effect (roughly proportional to the ratio of crossing-over to the intensity of selection and the frequency of hitchhiking events) can be examined. Recently, the first such analysis was attempted (26), which reported a correlation between π and indirect estimates of the levels of crossing-over per kilobase. While this analysis corroborates the general qualitative correlation between crossing-over per physical length and polymorphism, it does not address any quantitative aspect of the hitchhiking effect model. This model is thought to generate linkage disequilibria and to skew the site frequency spectrum (10, 12, 27). To further evaluate the hitchhiking effect model, both mean and higher-order properties of the distribution of molecular polymorphism and divergence must be examined at many additional loci at which the levels of crossing-over per kilobase can be measured directly and shown to vary over several orders of magnitude.

A further aspect of the hitchhiking effect model is the position at which the selected variants occur in the genome. While the theory is formulated in terms of the favored mutant arising at random loci throughout the chromosome, it is possible that the selected sites are concentrated in the centromeres and telomeres. Is the observed distribution of polymorphism consistent with the random distribution of selected substitutions, or does it fit better a model in which more selected substitutions occur primarily outside the euchromatin? Independently of how this question is answered, it seems clear now that we can expect polymorphism to be reduced in regions of extremely low crossing-over per physical length—e.g., near the centromeres and some telomeres of *Drosophila*. Yet we cannot expect interspecific divergence to be correlated with polymorphism in these regions of the genome. As more data become available on the distribution of crossing-over per kilobase in humans and mice, it may be possible to examine the hitchhiking hypothesis in these systems also.

We thank Kevin O'Hare and his colleagues for sharing their results from the study of *su(f)* and for many helpful discussions of the structure of the gene. We also thank J. M. Martín-Campos for sharing his lines of *D. simulans*. We also acknowledge support from National Science Foundation Grant BSR-9117222 and North Atlantic Treaty Organization Grant CRG-900651.

1. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).
2. Gillespie, J. H. (1992) *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford).
3. Kreitman, M. & Aguadé, M. (1986) *Genetics* **114**, 93–110.
4. Hudson, R. R., Kreitman, M. & Aguadé, M. (1986) *Genetics* **116**, 153–159.
5. Kreitman, M. & Hudson, R. R. (1991) *Genetics* **127**, 565–583.
6. McDonald, J. H. & Kreitman, M. (1991) *Nature (London)* **351**, 652–654.
7. Berry, A. J., Ajioka, J. W. & Kreitman, M. (1991) *Genetics* **129**, 1111–1117.
8. Begun, D. J. & Aquadro, C. F. (1991) *Genetics* **129**, 1147–1158.
9. Martín-Campos, J. M., Comerón, J. M., Miyashita, N. & Aguadé, M. (1992) *Genetics* **130**, 805–816.
10. Aguadé, M., Miyashita, M. & Langley, C. H. (1989) *Genetics* **122**, 607–615.
11. Maynard-Smith, J. & Haigh, J. (1974) *Genet. Res.* **23**, 23–35.
12. Kaplan, N., Hudson, R. R. & Langley, C. H. (1989) *Genetics* **123**, 887–899.
13. Moriyama, E. N. & Gojobori, T. (1992) *Genetics* **130**, 855–864.
14. Miyashita, N. & Langley, C. H. (1988) *Genetics* **120**, 199–212.
15. Henikoff, S. (1984) *Gene* **28**, 351–359.
16. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
17. Kreitman, M. & Aguadé, M. (1986) *Proc. Natl. Acad. Sci. USA* **86**, 3562–3566.
18. Miyashita, N. (1990) *Genetics* **125**, 407–419.
19. Nei, M. & Tajima, F. (1981) *Genetics* **97**, 145–163.
20. Hudson, R. R. (1982) *Genetics* **100**, 711–719.
21. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3**, 418–426.
22. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Proteins Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 21–132.
23. Langley, C. H. (1990) in *Population Biology of Genes and Molecules*, eds. Takahata, N. & Crow, J. F. (Baifukan, Tokyo), pp. 75–91.
24. Stephan, W. & Langley, C. H. (1989) *Genetics* **121**, 89–99.
25. Stephan, W. & Mitchell, J. (1992) *Genetics* **132**, 1039–1045.
26. Begun, D. J. & Aquadro, C. F. (1992) *Nature (London)* **356**, 519–520.
27. Macpherson, J. N., Weir, B. & Leigh-Brown, A. J. (1990) *Genetics* **126**, 121–129.
28. Mitchelson, A., Simonelig, M., Williams, C. & O'Hare, K. (1993) *Genes Dev.*, in press.