Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 1611–1615, March 1991
Evolution

Concerted evolution of duplicated protein-coding genes in *Drosophila*

(gene conversion/gene duplication/molecular evolution/α-amylase)

DONAL A. HICKEY*, LAURE BALLY-CUIF, SUMAIA ABUKASHAWA, VERONIQUE PAYANT, AND BERNHARD F. BENKEL

Department of Biology, University of Ottawa, Ottawa, ONT, Canada, K1N 6N5

Communicated by Wyatt W. Anderson, December 6, 1990 (received for review October 1, 1989)

ABSTRACT **Very rapid rates of gene conversion were observed between duplicated α-amylase-coding sequences in *Drosophila melanogaster*. This gene conversion process was also seen in the related species *Drosophila erecta*. Specifically, there is virtual sequence identity between the coding regions of the two genes within each species, while the sequence divergence between species is close to that expected based on their phylogenetic relationship. The flanking, noncoding regions are much more highly diverged and do not appear to be subject to gene conversion. Comparison of amylase sequences between the two species provides a clear demonstration that recurrent gene conversion does indeed lead to the concerted evolution of the gene pair.**

There is considerable evidence that the members of gene families evolve in a nonindependent, or concerted, fashion (1–15). Here, we describe a striking example of concerted evolution, resulting from the interconversion of a pair of linked genes. Specifically, we present DNA sequence data from two closely linked enzyme-coding genes, from each of two closely related species of *Drosophila*. The alignment of these four sequences provides a straightforward comparison of within-species and between-species patterns of sequence divergence. The results provide a simple but convincing example of (*i*) rapid gene conversion between the duplicated genes within each species and (*ii*) concerted evolutionary divergence of the gene pairs between species. Moreover, we can identify a distinct boundary between the coding sequences that are undergoing concerted evolution, and the flanking sequences that are evolving independently. This system is exceptional because the conserved gene arrangement is such that it maximizes the rate of gene conversion and, consequently, it enables us to get a relatively complete description of the evolutionary dynamics. Thus, it provides a link between laboratory experiments that show the possibility of gene conversion (16–18), on the one hand, and long-term evolutionary patterns of concerted evolution among gene families, on the other hand. Often, the long-term patterns are difficult to interpret, whereas the evolutionary relevance of the short-term experiments can easily be questioned. The time scale, in the case of the data reported here, ≈15 million years (19, 20), is long enough for the observation of recurring evolutionary events, but it is short enough to facilitate an unambiguous interpretation of the results. Consequently, we can think of these duplicated *Drosophila* genes as a "natural experiment" in molecular evolution.

The genes that we examined encode α-amylase enzymes and they have already been studied extensively at the molecular level in both *Drosophila* (21, 22) and mammals (23). α-Amylase enzymes of *Drosophila melanogaster* are encoded by a pair of closely linked, glucose-repressible genes located on the right arm of the second chromosome (24, 25). These genes are divergently transcribed and the transcriptional start sites are ≈4 kilobases apart (26). In many *Drosophila* strains, the products of the two genes are electrophoretically distinguishable; moreover, there are high levels of allozyme and restriction polymorphism found in natural populations (27–29).

The work described here follows from our earlier observations on the degree of sequence divergence between the duplicated amylase genes (26). In using these sequences to estimate the time since the duplication event, we encountered a paradox—namely, that the duplication appeared to be very recent if one compared only the duplicated coding sequences, whereas the amount of sequence divergence was much greater (and the estimate of the time since duplication was consequently much longer) if we considered only the flanking upstream regions. Subsequent genomic Southern blot analyses of several related *Drosophila* species indicated that the duplication was indeed older than the estimate provided by the comparison of the duplicated coding sequences and that it predated the speciation events within the species subgroup (30). This suggested the possibility of concerted evolution of the duplicated structural genes. The results reported here confirm this suggestion. First, the comparison of duplicated amylase sequences within *D. melanogaster* demonstrates the effects of gene conversion between the two coding regions; second, the comparison of these sequences with the homologous regions in *Drosophila erecta* shows that gene conversion can indeed lead to concerted evolution of the gene pair.†

MATERIALS AND METHODS

The isolation and characterization of the amylase clones discussed here have been described in detail elsewhere (26, 31, 32). DNA sequences were obtained by the dideoxynucleotide chain-termination technique (33) using custom-synthesized oligonucleotide primers and the LKB Macrophor electrophoresis equipment. The sequences were assembled and analyzed with a sonic digitizer and the Microgenie (Beckman) computer programs (34).

RESULTS

The arrangement of the duplicated amylase sequences in *D. melanogaster* is shown in Fig. 1*A*, and the pattern of sequence divergence between the duplicated gene copies is illustrated in Fig. 1*B*; the sequences that span the junctions of the coding and flanking regions are aligned in Fig. 1*C*. Essentially, these sequences are very similar over the entire coding sequence (i.e., between the ATG and TAA codons

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.

†The sequences reported in this paper have been deposited in the GenBank data base (accession nos. M55995 and M55996).

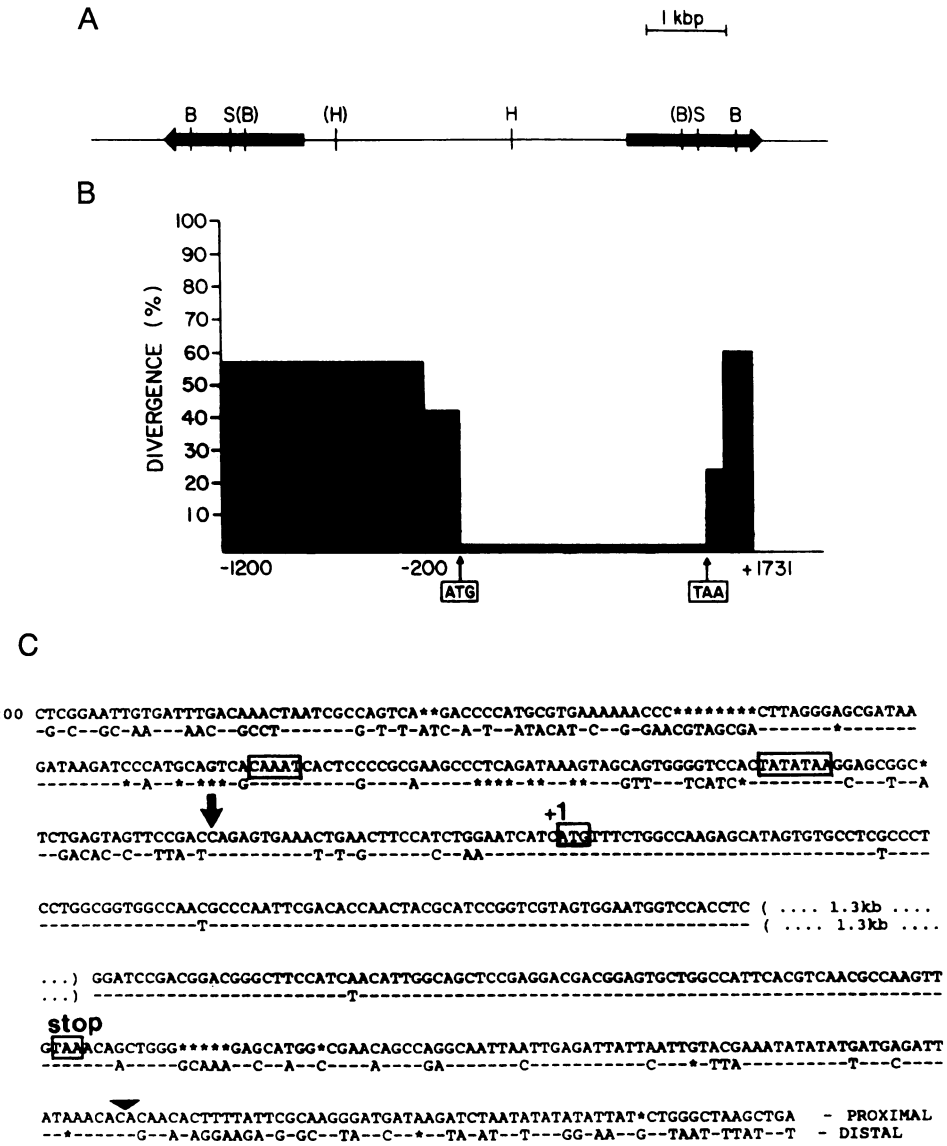

FIG. 1. Comparison of the duplicated amylase-coding genes in *D. melanogaster*. (*A*) Arrangement of the duplicated coding sequences. The two genes (arrows) are ≈4 kilobases apart (24) and are divergently transcribed (26). The proximal and distal gene copies (24) are shown on the left and right, respectively. Restriction enzyme sites for *Bam*HI (B), *Hin*dIII (H), and *Sal* I (S) are shown. Those sites that are not shared with the *D. erecta* sequences are shown in parentheses. kbp, Kilobase pair. (*B*) Percentage sequence divergence between the duplicated genes. Three regions were compared: (*i*) upstream, noncoding; (*ii*) coding; and (*iii*) downstream, noncoding. Positions of the start and stop codons are indicated. (*C*) Alignment of sequences flanking the coding regions of the duplicated amylase genes. This alignment illustrates the relatively sharp boundary between the conserved transcribed regions and the divergent flanking regions. Regulatory motifs (CAAT and TATA) are boxed, along with the start and stop codons. Initiation of transcription (26) is indicated by an arrow. The polyadenylylation site is indicated by an arrowhead. Sequence identities are indicated by dashes; gaps are shown as asterisks.

marked in Fig. 1*B*), but the similarity drops off very quickly both upstream and downstream of the coding region (see Fig. 1 *B* and *C*). One might argue that the high level of coding sequence conservation is due to selective constraints acting on the coding capacity. Such an explanation, however, would not account for the very low level of synonymous substitutions between the duplicated coding sequences in both *D. melanogaster* (26, 35) and *D. erecta* (ref. 32; see Fig. 2). For instance, the percentage divergence between the two species at silent sites is ≈6% (32), whereas the divergence at silent sites between gene copies within a species is of the order of 0–1%.

A relatively ancient gene duplication event, coupled with concerted evolution of the two coding regions, can explain both the divergence of the flanking sequences and the conservation of the coding sequences, including the silent sites. A comparison of the duplicated amylase genes of *D. melanogaster* with those from a related *Drosophila* species, *D. erecta*, indicates that this is indeed the case. Both *D. melanogaster* and *D. erecta* contain the duplicated amylase gene structure, as do the other six members of the *D. melanogaster* species subgroup (30, 36). The estimated time since the divergence of *D. melanogaster* and *D. erecta* is 15 million years (19, 20). The first piece of evidence for concerted evolution of the duplicated genes came from patterns of restriction map divergence between the two species. For instance, a *Bam*HI restriction endonuclease site that is present in both genes of *D. melanogaster* is absent from both gene copies in *D. erecta* (see Fig. 1*A*). A comparison of the amylase coding region of *D. melanogaster* with the duplicated coding sequences of *D. erecta* provides compelling evidence for the concerted evolution of these duplicated genes (see Fig. 2). For instance, in the coding region there are 39 nucleotide substitutions separating the two species, but 38

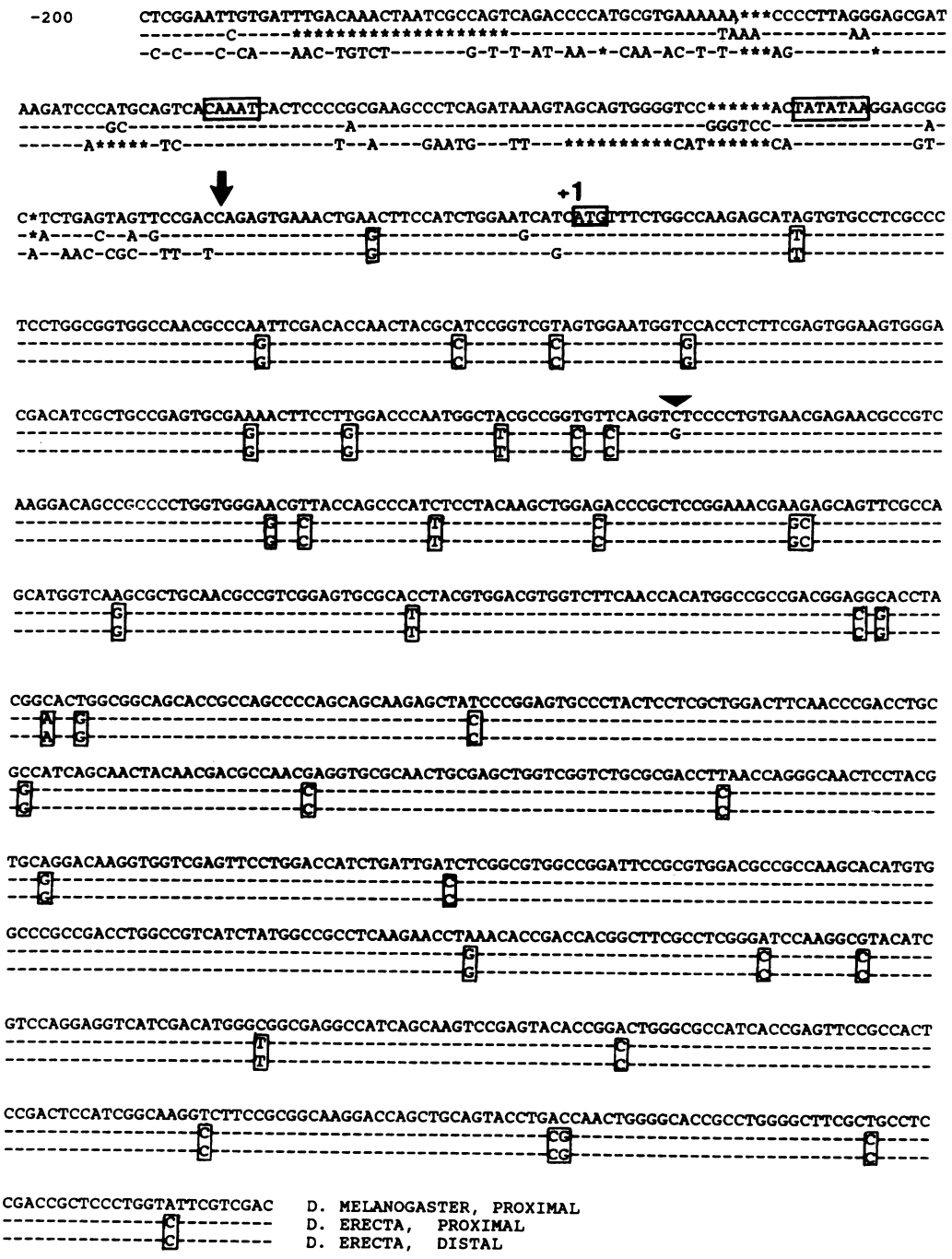

FIG. 2. Comparison of amylase sequences from *D. melanogaster* and *D. erecta*. Sequences from both *D. erecta* genes are aligned with the proximal gene of *D. melanogaster*. Alignment extends from nucleotide position −200 to position +909. Dashes represent sequence identity with the *D. melanogaster* sequence; asterisks denote sequence gaps. Start codon ATG is boxed. Substitutions (with respect to the *D. melanogaster* gene) shared by the two *D. erecta* genes are also boxed. Initiation of transcription is shown by an arrow. Single nucleotide difference between the two *D. erecta* coding sequences is indicated by an arrowhead.

of these are shared by the two *D. erecta* genes; i.e., all but one of the nucleotide substitutions that have occurred since the species divergence have been incorporated into both gene copies of *D. erecta*. In both species, the proximal and distal gene sequences are highly divergent in the upstream nontranscribed region; the level of sequence similarity increases very rapidly, however, downstream of the transcription initiation points (Figs. 1*C* and 2, arrows).

The interspecific sequence divergence between the coding regions (4.25%) is greater than the divergence between gene copies within a species (1%); it reflects the independent evolution of the two complexes since the speciation event, coupled with the effects of stabilizing selection acting on the coding capacity. The number of silent substitutions between the species (9.3%) and the divergence between homologous flanking sequences (≈15%) give an estimate of the interspecific divergence in the absence of selection. These values are consistent with what one would expect for the neutral rate of substitution. The much higher levels of divergence between the nontranscribed flanking sequences of the two gene copies within a species (40–60%; see Fig. 1*B*) reflects the fact that the duplication of the amylase genes predated the speciation of *D. melanogaster* and *D. erecta*. Finally, the very low levels of sequence divergence between the duplicated coding regions within both species (1% or less) reflects the action of gene conversion; note that the rate of gene conversion is high

enough to maintain near-identity of the duplicated coding sequences, despite the considerable evolutionary age of the duplication. Gene conversion is not, however, so rapid as to homogenize naturally occurring chromosomes carrying distinct proximal and distal coding sequences that are stable for many generations, as is evident from the distinct allozymic patterns of the duplicated genes (37). These data do not allow us to get very exact measurements of the frequency of gene conversion events; there have been either many events involving portions of the coding region or few events involving the entire coding region within both lineages during the period since the species divergence.

DISCUSSION

We propose that the duplicated amylase coding sequences can loop over to facilitate mitotic (or meiotic) recombination and gene conversion through heteroduplex repair. This is essentially the model described for the *Drosophila* heat shock genes (38). Direct evidence for such a mechanism comes from the observation that an amylase mutant isolated from a natural population has an inversion of the intergenic region (39, 40), indicating that recombinational exchanges between the duplicated coding regions do, in fact, occur. The observation that the coding sequences are converted, but not the flanking sequences, suggests that the conservation of these sequences for their coding capacity maintains their similarity at a level sufficient for a state of susceptibility to gene conversion; the flanking sequences have been able to "escape" this effect (41) and so continue to diverge. An alternative possibility, but one we believe to be less likely, is that there is a process of retrotransposition of cDNA copies of one gene that replaces the coding sequence of the second gene. Although the molecular mechanisms underlying these processes in *Drosophila* may be very similar to those involved in the classic examples of gene conversion in fungi, the genetic consequences are not; this is because the interaction here is between two copies of a gene on the same chromosome rather than between single-copy genes on different chromosomes.

Recent experiments have shown that there is a bias in heteroduplex repair that favors the formation of G·C base pairs (18). The results of those experiments suggest that, during the course of evolution, duplicated sequences that undergo repeated rounds of gene conversion would become increasingly G+C-rich. This latter effect would inevitably bias the codon usage within such genes. The genes described here bear out these predictions. The third codon positions within the *Drosophila* amylase genes show an extreme G·C bias (almost 90% G·C), in contrast to the third codon positions of the homologous sequences in other insect species, such as *Tribolium castaneum*, where the third codon position of the homologous gene is ≈50% G·C (22, 42).

Gene conversion is an unbiased process in that it may propagate a new mutation into the second gene copy or, alternatively, it may equally well eliminate a mutation by converting it back to the nonmutant sequence. Thus, it does not affect the average rate of accumulation of mutations in the population, nor will it affect the average rate of neutral substitution. The average genetic distance between amylase variants segregating in a population should also remain unchanged. The most obvious predicted effect of this kind of gene conversion is that the variance in the genetic distance between allelic variants will be doubled—i.e., although intrachromosomal gene conversion tends to homogenize gene copies on the same chromosome in the longer term, it can either increase or decrease the sequence differences between homologous genes on different chromosomes in the shorter term. In practice, this means that a survey of amylase sequences in natural populations should reveal a wide variety of genetic distances between allelic variants, with some pairs of variants showing very few differences while other pairs are relatively widely divergent. Our preliminary data indicate that this is indeed the case (refs. 26 and 31; unpublished observations).

Several previous examples of concerted evolution are based on a mechanism of unequal crossing-over between tandemly repeated gene clusters (2, 4, 5, 8, 9, 43–46). A major difference between the outcome of unequal crossing-over and the gene conversion mechanism illustrated here is that unequal crossing-over results in the homogenization of the intergenic regions as well as the coding regions (45). Gene conversion tracts may, or may not, include some flanking noncoding sequences (12, 13). Another major difference is that unequal crossing-over can accelerate fixation rates in addition to homogenization rates (7, 47). The gene conversion process described here is very effective in homogenizing the members of a gene family; but other processes, such as biased gene conversion (48, 49) or replicative transposition (50–52), would be required to speed the incorporation of particular mutant types into the population.

In summary, our results clearly demonstrate the homogenization of gene copies on the same chromosome and the consequent concerted evolution of the gene pairs; a question that remains is whether a similar homogenization phenomenon occurs between homologous sequences on separate chromosomes. If interchromosomal gene conversion were also occurring, then mutations that occurred on a given gene copy could be propagated not only into the second gene copy on the same chromosome, but also further propagated into either gene copy on another chromosome. Although some interallelic sequence comparisons are already available, a much larger data set will be necessary to provide conclusive evidence for interchromosomal conversion of these genes, in addition to the high rate of intrachromosomal gene conversion reported here.

We thank Gabby Dover for his comments on the manuscript. This research was supported by an Operating Grant from the Natural Sciences and Engineering Research Council, Canada (D.A.H.), and a Fellowship from the Conseil National de la Recherche Scientifique, France (V.P.). D.A.H. is a Fellow of the Canadian Institute for Advanced Research.

1. Hood, L., Campbell, J. H. & Elgin, S. C. R. (1975) *Annu. Rev. Genet.* **9**, 305–353.
2. Smith, G. P. (1976) *Science* **191**, 528–535.
3. Arnheim, N., Krystal, M., Schnickel, R., Wilson, G., Ryder, O. & Zimmer, E. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7323–7327.
4. Szostak, J. W. & Wu, R. (1980) *Nature (London)* **284**, 426–430.
5. Slightom, J. L., Blechl, A. E. & Smithies, O. (1980) *Cell* **21**, 627–638.
6. Klein, H. L. & Petes, T. D. (1981) *Nature (London)* **289**, 144–148.
7. Dover, G. (1982) *Nature (London)* **299**, 111–117.
8. Arnheim, N. (1983) in *Evolution of Genes and Proteins*, eds. Nei, M. & Koehn, R. K. (Sinauer, Sunderland, MA), pp. 38–61.
9. Ohta, T. (1983) *Theor. Popul. Biol.* **23**, 216–240.
10. Ohta, T. & Dover, G. A. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 4079–4083.
11. Maizels, N. (1987) *Cell* **48**, 359–360.
12. Gumucio, D. L., Wiebauer, K., Caldwell, R. M., Samuelson, L. C. & Meisler, M. H. (1988) *Mol. Cell. Biol.* **8**, 1197–1205.
13. Xiong, Y., Sakaguchi, B. & Eickbush, T. H. (1988) *Genetics* **120**, 221–231.
14. Nagylaki, T. (1988) *Genetics* **120**, 291–301.
15. Matsuo, Y. & Yamazaki, T. (1989) *Genetics* **122**, 87–97.
16. Willis, K. K. & Klein, H. L. (1987) *Genetics* **117**, 633–643.
17. Letsou, A. & Liskay, R. M. (1987) *Genetics* **117**, 759–769.
18. Brown, T. C. & Jiricny, J. (1988) *Cell* **54**, 705–711.
19. Cariou, M. L. (1987) *Genet. Res.* **50**, 181–186.

20. Lachaise, D., Cariou, M. L., David, J. R., Lemeunier, F., Tsacas, L. & Ashburner, M. (1988) *Evol. Biol.* **22,** 159–225.
21. Doane, W. W., Gemmill, R. M., Schwartz, P. E., Hawley, S. A. & Norman, R. A. (1987) in *Isozymes: Current Topics in Biological and Medical Research* (Liss, New York), Vol. 14, pp. 229–266.
22. Hickey, D. A., Benkel, B. F. & Magoulas, C. (1989) *Genome* **31,** 272–283.
23. Meisler, M. H. & Gumucio, D. L. (1986) in *Molecular and Cellular Basis of Digestion*, eds. Desnuelle, P., Sjostrom, H. & Noren, O. (Elsevier, Amsterdam), pp. 249–263.
24. Gemmill, R. M., Schwartz, P. E. & Doane, W. W. (1986) *Nucleic Acids Res.* **14,** 5337–5352.
25. Benkel, B. F. & Hickey, D. A. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 1337–1339.
26. Boer, P. H. & Hickey, D. A. (1986) *Nucleic Acids Res.* **14,** 8399–8411.
27. Hickey, D. A. (1979) *Genetica (The Hague)* **51,** 1–4.
28. Singh, R. S., Hickey, D. A. & David, J. A. (1982) *Genetics* **101,** 235–256.
29. Langley, C. H., Shrimpton, A. E., Yamazaki, T., Miyashita, N., Matsuo, Y. & Aquadro, C. F. (1988) *Genetics* **119,** 619–629.
30. Payant, V., Abukashawa, S., Sasseville, M., Benkel, B. F., Hickey, D. A. & David, J. (1988) *J. Mol. Biol. Evol.* **5,** 560–567.
31. Benkel, B. F., Abukashawa, S., Boer, P. H. & Hickey, D. A. (1987) *Genome* **29,** 510–515.
32. Bally-Cuif, L., Payant, V., Abukashawa, S., Benkel, B. F. & Hickey, D. A. (1989) *Genet. Sel. Evol.* **22,** 57–64.
33. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
34. Queen, C. & Korn, L. J. (1984) *Nucleic Acids Res.* **12,** 581–599.
35. Okuyama, E. & Yamazaki, T. (1988) *Proc. Jpn. Acad. Ser. B* **64,** 274–277.
36. Dainou, O., Cariou, M.-L., David, J. R. & Hickey, D. A. (1987) *Heredity* **59,** 245–251.
37. Doane, W. W. (1969) *J. Exp. Zool.* **171,** 321–342.
38. Leigh-Brown, A. J. & Ish-Horowicz, D. (1981) *Nature (London)* **290,** 677–682.
39. Hickey, D. A., Benkel, B. F., Abukashawa, S. & Haus, S. (1988) *Biochem. Genet.* **26,** 757–768.
40. Schwartz, P. E. & Doane, W. W. (1989) *Biochem. Genet.* **27,** 31–45.
41. Walsh, J. B. (1987) *Genetics* **117,** 543–557.
42. Hickey, D. A., Benkel, B. F., Boer, P. H., Genest, Y., Abukashawa, S. & Ben-David, G. (1987) *J. Mol. Evol.* **26,** 252–256.
43. Petes, T. & Fink, G. R. (1982) *Nature (London)* **300,** 216–217.
44. Coen, E. S., Strachan, T. & Dover, G. A. (1982) *J. Mol. Biol.* **158,** 17–35.
45. Brown, D. D. & Sugimoto, K. (1974) *Cold Spring Harbor Symp. Quant. Biol.* **38,** 501–505.
46. Tautz, D., Tautz, C., Webb, D. & Dover, G. A. (1987) *J. Mol. Biol.* **195,** 525–542.
47. Dover, G. A. (1986) *Trends Genet.* **2,** 159–165.
48. Slatkin, M. (1986) *Genetics* **112,** 681–698.
49. Walsh, J. B. (1986) *Genetics* **112,** 699–716.
50. Doolittle, W. F. & Sapienza, C. (1980) *Nature (London)* **284,** 601–603.
51. Orgel, L. E. & Crick, F. H. C. (1980) *Nature (London)* **284,** 604–607.
52. Hickey, D. A. (1982) *Genetics* **101,** 519–531.