The EMBO Journal vol.4 no.7 pp.1701 – 1708, 1985

# Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*

**Tom Strachan[1], David Webb and Gabriel A.Dover**

Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

[1]Present address: Department of Medical Genetics, University of Manchester, UK

Communicated by J.Fincham

**Multigene and non-genic DNA families are in a state of turnover and hence are continually being replaced throughout a population by new variant repeats. To quantify such molecular processes, in the absence of selection, it is necessary to find and compare stages of transition during the homogenization of at least two non-genic families evolving in parallel in a closely related group of species. Detailed sequence analysis of patterns of variation, at each nucleotide position considered independently, amongst repeats of two tandem DNA families from seven related *Drosophila* species, reveals all stages of transition during the spread of randomly produced variant repeats. Variant repeats are found at different stages of homogenization and fixation in a population, irrespective of the loci, chromosomes or individuals from which they were cloned. Differences between the families in the relatively small number of variants at each transition stage and the greater number of fully homogenized and fixed variants between species of greater divergence indicate that the process of spread (molecular drive) is rapid relative to the mutation rate and occurs at seemingly different constant rates for each family. Occasional gene conversions, in addition to unequal exchanges, have contributed to family turnover. The significance of these results to the evolution of functional multigene families and divergence and conservation of sequences is discussed.**

*Key words:* multigene families/molecular drive/concerted evolution/evolution rates/*Drosophila*

## Introduction

Multigene families are influenced by several mechanisms of non-reciprocal exchanges such as gene conversion, unequal exchange, transposition, slippage replication and RNA-mediated transfers (for reviews, see Ohta, 1980; Jones and Kafatos, 1982; Dover, 1982; Dover *et al.*, 1982; Kedes, 1979; Hood *et al.*, 1975; Flavell, 1982; Ohta and Dover, 1984; Long and Dawid, 1980; Fedoroff, 1979; Arnheim, 1983). These mechanisms can gradually spread a variant gene throughout a family within a sexual population. Family homogenization and population fixation are inextricably linked. This population genetics process is called molecular drive (Dover, 1982) and is invoked to explain an observed pattern of variation known as concerted evolution, (that is, high levels of family homogeneity for species-diagnostic mutations). Molecular drive, by taking into account non-reciprocal genetic transfers between homologous (and sometimes non-homologous) chromosomes, attempts to explain how all relevant chromosome lineages (for example, all X and Y rDNA arrays in a *Drosophila* species – Coen and Dover, 1983) have achieved high levels of genetic identity.

If molecular drive is a time-dependent process then it is necessary to reveal the expected stages of transition during the spread of a variant gene, irrespective of the loci, chromosomes and individuals from which genes have been sampled. Secondly, it is important to dissociate the dynamics of spread due to molecular drive from that due to natural selection.

Accordingly, we have made an analysis of sequence variation in family members that were selected at random from two abundant tandem families of non-coding sequences, the 360 and 500, which are found within a group of eight sibling species of the *melanogaster* species subgroup of *Drosophila* (Strachan *et al.*, 1982; Dover *et al.*, 1982; Barnes *et al.*, 1978).

A detailed comparative survey of sequence variation in two families that are evolving in parallel in an extended number of species having varying degrees of genetic relatedness, reveals the dynamics of spread of new mutations in a way not previously accessible from generally more restricted studies of single families taken from one or very few species. Our method of analysis reveals, for the first time, different stages of transition in the fixation of randomly produced variant repeats of each family, by molecular drive. The rate of spread between loci is much faster than the rate of mutation at a locus, and is seemingly different in the two families. There are indications that gene conversion, in addition to unequal exchange, has contributed to the evolution of the families. The significance of the dynamics of change by molecular drive to the function of multigene families and the involvement of molecular drive in both the divergence and conservation of sequences are discussed briefly (see also Dover and Flavell, 1984; Dover and Tautz, 1985).

## Results

### *Intraspecific variation in the '500' and '360' families*

Table I shows the nature and range of intraspecific variation amongst repeat units of each family sampled from several thousand individuals in each species. There are several points of interest relating to the processes of change in the two families.

First, the lengths of the repeats can be different between, and sometimes within, species. Such intraspecific length variation can be used to define subfamilies (see below). Length differences are due to insertions and deletions of sequences, most noticeably in the 360 family of *D. teisseiri* most of whose members are shorter than the interspecific consensus length by 164 bp.

Secondly, >65% of the variation is due to single base substitutions. The majority of addition/deletion events involve single nucleotides, although the proportion of longer addition/deletions is higher in the 500 family. Mutations appear to be evenly distributed throughout the repeat units within family consensus sequences in all species (see Figures 1 and 2), and also in the 'group' consensus of all species (the consensus of consensuses) of each family. (There are no absolute statistical tests for randomness of distribution, for all rely on arbitrarily chosen 'windows' of comparison within a repeat unit.) An absence of selection in the direction of mutation is suggested by the observation of a

**Table I.** Intraspecific variation in the 360 and 500 families

| Family | Species | Repeat unit length (bp) | Number of clones | Data base (kb) | Intraspecific variation method 1 | | | | Intraspecific variation method 2 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Base substit. % | Transversion/ transitions ratio | Insertions/ deletions % | Total % | Variation per position |
| 360 | *mauritiana* | 357 | 6 | 1.9 | 1.65 | 1.46 | 0.92 | 2.57 | 3.40 |
| | *simulans* | 360 | 9 | 2.4 | 1.80 | 2.16 | 1.16 | 2.96 | 4.53 |
| | *orena* | 360 | 8 | 2.1 | 2.55 | 1.41 | 0.19 | 2.74 | 4.82 |
| | *teissieri* | 196 | 4 | 0.8 | 3.83 | 1.99 | 0.89 | 4.72 | 8.51 |
| | *yakuba* | 360 | 9 | 2.4 | 0.33 | 3.12 | 0.25 | 0.58 | 0.98 |
| 500 | *mauritiana* | 507 | 8 | 2.2 | 2.67 | 1.36 | 0.77 | 3.44 | 6.38 |
| | *simulans* | 504 | 15 | 4.3 | 3.67 | 1.50 | 0.84 | 4.51 | 7.76 |
| | *erecta* | 482 (248) | 9 | 2.5 | 0.70 | 1.33 | 0.15 | 0.85 | 3.40 |
| | *teissieri* | 546 (520) | 9 | 2.5 | 2.50 | 2.05 | 0.99 | 3.49 | 14.75 |
| | *yakuba* | 549 (330) | 11 | 3.1 | 1.30 | 1.95 | 0.38 | 1.68 | 3.41 |

DNA sequences of all clones were established either fully, by sequencing in both directions (especially in the case of the 360 family), or partially, such that consensus sequences were generally established by data from a minimum of four clones (and frequently more). For methods of analysis of intraspecific variation by method 1 and intraspecific variation by method 2, see Materials and methods. The 500 family complements of some species are characterized by prominent repeat unit length variants (numbers in brackets). These are due to either deletions in the repeating unit or to insertions of other non-homologous sequences. Methods 1 and 2, in estimating the mean variation per nucleotide position, under and over estimate the degree of variation, respectively, depending on the extent of the deletions/insertions (see text ).

```
          10         20         30         40         50         60         70         80         90        100
CON  CNATTTTTTG CAAATTTTGA TGATGNTACC CCTTACAAAA AATGCGAAAA TTTGNCCAAA AATTAATTTN CCAAAATCNG TNAAAAAGTG ANAGGGATN*
mel  -C-CA----- ---------- ---CCCCC-T ---------- ---------- --GAT----- ---------C --T-----CT -C--------A -T------C*
mau  -AG------- -----A---- ----***--- ---------- ---T------ ----G----- ---------T A-------A- -TT------- -A----G-T*
sim  -A-------C ---------- -----G---- ---------- ---T------ ----G----- ---------T A-------A- -TT------- -A----G-T*
ore  -C------G- --------A- -----T---- ---------- ---------- -CGAC----- --CG-----C --------C- -C--------- -T------C*
tei  TA-------C ---T------ -----A-*** ********** ********** ********** ********** ********** ********** **********
yak  -C------C- ---------- -----G---- -----TCG-- ---------- ----T-A--T TT--TT---G -G------** *G-------* **------AG

          110        120        130        140        150        160        170        180        190        200
CON  *****GTTAG CATTGGTTAT TAGCAGCACA AAACAGTNNT TCTTTTAGCT NTNTGACCAT TTTTAGCCAA GTTATAGCCA AAAN*GCCAA TTNNAAATNT
mel  *****----- --C----A-- ----T--T-- ------ATA- --G-AC-T-- A-G------- ---------- ------A-G- ---T*TT-GT --GT----A-
mau  *****----- T-------G- A--G--T--- ---TG--AC- ------*--- C-C------- ------T--- ---------- ---CA----- --TG----T-
sim  *****----- T-------G- A--G--T--- ---TG--AC- C---C-T--- C-C------- ------T--- ---------- ---AA----- --TG----T-
ore  *****----- -------A-- ----T--T-- -------TA- -------TG- T-A----T-- ---------- -----GATT- ---T*----- -AAA----A-
tei  ********** ********** ********** ********** ********** ********** ********** ********** ********** *******_T-
yak  ACATA----- -TA-CT---- ---------- -------CT- -A-------- G-GC-T---- -----A---- -----G---- ---C*---T- --GA----A-

          210        220        230        240        250        260        270        280        290        300
CON  NCNATTTTTT GCCAAAATNT ATTTTTCCA* NATTTTGNNC NNAAAATA*A TCNGTTNTTT TGATNNNAAC ATTNNAAATA ATTGTCCAAA TATGGAATGT
mel  C-AC------ -*--G-G-C- G--------* A----C-GT- ATC-----*- --AT--A--- --CCA*C--- --AAA----- ------TG-- ----------
mau  *-AT------ --------*G- -----C---* G-----TTGT GA----A-T- -TT-G-A--- ----CAA--- ---CG----- ---AC----- ----------
sim  *-AT------ --------*A- -----C---* G-----TTGT GA----A-T- -TT-G-T--- ----CAA--- ---CG----- ---AC----- ----------
ore  TTT-A----- --------CC G-----T--* AA-----AT- AT------*- --A---T--- --CAA*C--- TC-AA----- ------TG-- -TC------G
tei  TGG------- C-G-----GG *------TGT TT-----CGA TTT-----** --A-*-A--- ----CAG--- ---CG----- TGG------- ----------
yak  C-G------- A-A-----T- T-----T-GA TT-----AT- *A------*- --C---T--- -T-AA*T--- CACAA----- G--------- AG-------C

          310        320        330        340        350        360        370
CON  CATACCTCGT TGAGTTCGTA *ATTAAATTN CCAATCGAAC TGTGTT*CAA AATTTGNAAT TCTATTTNTT TN
mel  ----T---AC ----C----- *--A-----T ------A--- ------*--- --A-G-A--- -AA----T-- GG
mau  ---------- ------T--T TC------*C ---------T --C---*--- GT----G--- ----GGAGG- -T
sim  ---------- ------T--T TC------*C ---------T --CA--*--- GT---*G--- ----GGAGG- -T
ore  T--------- --C------- *G--T----T ---------- ------*T-- --G---G--C -------T-- -G
tei  ----T----- ---A------ *--------T ---------- ------TAC- --AA*AA--- GAAT---T-- -T
yak  ---------- ---A------ *-CA-----C --T-----T- CA---A*--T -C----A--- ----A--G-- -G
```

**Fig. 1.** Species consensus of 360 family sequences and a group consensus of all species. Consensus 360 family sequences from individual species (including the *D. melanogaster* sequence established by Hsieh and Brutlag, 1979b) are aligned to maximise interspecific sequence homologies. Dashes signify that the nucleotides in individual species sequences are the same as the overall consensus represented as CON. Asterisks denote deletions as required in the consensus sequence, to accommodate insertion events occurring in a minority of the species. N denotes positions in the overall consensus sequence where there is no strong majority nucleotide shared across the species. Species abbreviations are: mel: *D. melanogaster*; mau: *D. mauritiana*; sim: *D. simulans*; ore: *D. orena*; tei: *D. teissieri*; yak: *D. yakuba*.

1.5- to 2.0-fold excess of transversions over transitions in each family (Table I), (see Discussion). Furthermore, no statistically significant repeating subunits are observed within each family repeat (see Materials and methods).

The last two columns in Table I show the within-species variation calculated by two different methods (described in detail in Materials and methods). The first method is more flexible but less robust than the second. Method 1 uses a consensus sequence from a single species as a point of reference for assessing a percentage sequence divergence of each clone. Method 2 assesses

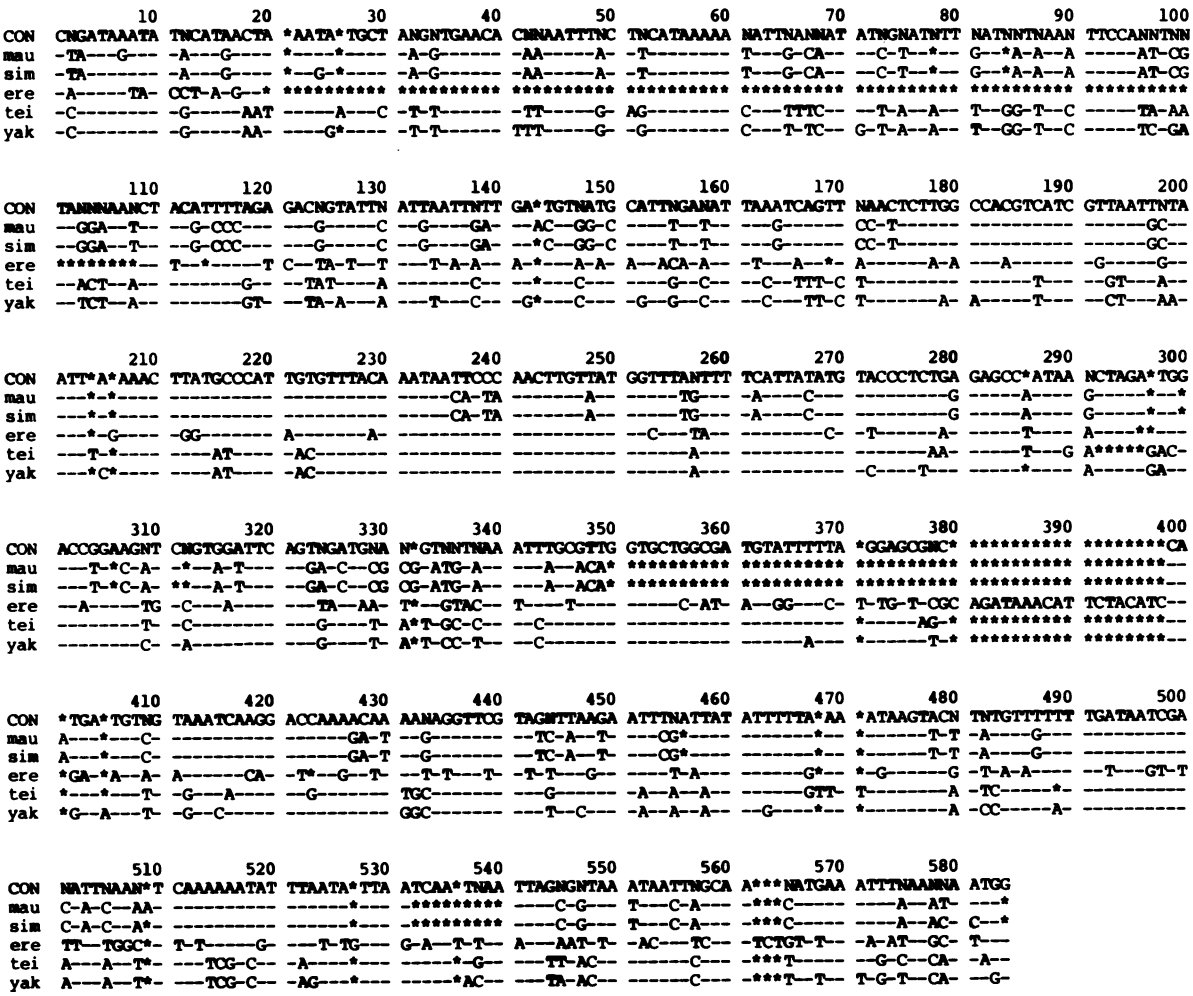

**Fig. 2.** Species consensus of 500 family sequences and a group consensus of all species. Symbols and species abbreviations are as in Figure 1.

the mean variation per nucleotide position using all pair-wise clonal comparisons with a species. For interspecific comparisons method 2 is limited to those closely related species where an unambiguous alignment of sequences is possible. Method 1 can be used for species with large differences. Method 1 counts large deletions/insertions as single events and hence is expected to underestimate the total variation, whereas method 2 counts each base of a deletion/insertion independently and hence overestimates the total variation. The assessments of divergence using both methods for the two closest species, *D. mauritiana versus D. simulans*, are shown in Table II.

In all cases, the consensus sequence of each species family reflects the restriction sites and repeat length of the majority of members of a family derived from whole family analysis (Strachan *et al.*, 1982, and unpublished results). There are some indications of subfamilies, as defined by different repeat unit lengths in the 500 family of *D. erecta, D. teissieri* and *D. yakuba*. However, the within-species differences averaged over both subfamilies is very much less than that between species (see below). The clones appear to be random samples of the available variation in each species notwithstanding potential cloning biases and loss of subfamilies during the two rounds of purification of buoyant density fractions.

### *Concerted evolution in the '500' and '360' families*

Table II shows the matrix of all available pair-wise species comparisons of sequences of each family. All comparisons based on

**Table II.** Interspecific sequence variation

Interspecific variation method 1

| 360 family | Mau | Sim | Ere | Ore | Tei | Yak | Mel |
|---|---|---|---|---|---|---|---|
| 500 family | | | | | | | |
| Mau | | 3.3 | n.d | 31.5 | 33.4 | 34.8 | 33.4 |
| Sim | 1.4 | | n.d | 31.6 | 33.2 | 34.7 | 34.9 |
| Ere | 42.5 | 41.8 | | n.d | n.d | n.d | n.d |
| Ore | – | – | – | | 28.8 | 35.0 | 20.6 |
| Tei | 34.2 | 34.2 | 38.7 | – | | 35.3 | 30.7 |
| Yak | 34.1 | 33.8 | 40.5 | – | 12.0 | | 35.0 |

Interspecific variation method 2

| | |
|---|---|
| Mau/Sim 360 | 5.44 |
| Mau/Sim 500 | 8.10 |

Top. Interspecific variation (method 2) was estimated as described in Materials and methods from the consensus sequence alignments of Figures 1 and 2. Dashes signify the inability to detect the 500 family in *D. orena*. n.d. denotes not determined. Species abbreviations are as in Table I. Bottom. Interspecific variation (mean variation per nucleotide position) assessed by method 2 (see Materials and methods) for both families of *D. mauritiana versus D. simulans*.

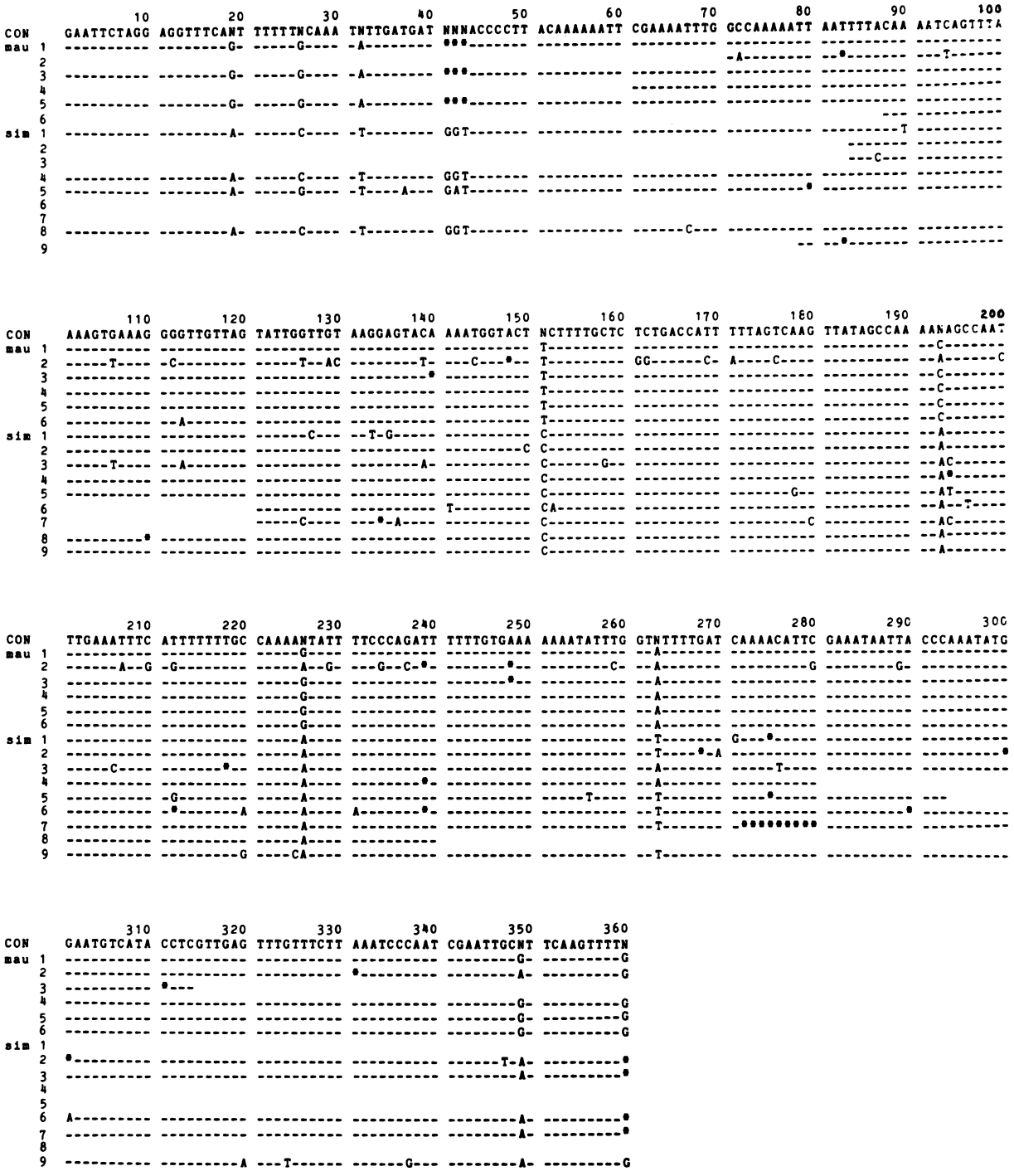

**Fig. 3.** Sequence alignment of clones representing the 360 families of *D. mauritiana* and *D. simulans*. Symbols and species abbreviations are as in Figure 1. The complete sequences of clones 1 and 5 of *D. mauritiana* were obtained by sequencing from both ends.

consensus sequences (method 1), are approximately one order of magnitude greater than the within-species divergences (cf. Table I) excepting between *D. mauritiana* and *D. simulans* (both families) and between *D. yakuba* and *D. teissieri* (the 500 family).

Table II also shows the interspecific variation between *D. mauritiana* and *D. simulans* using the more sensitive analysis of method 2. This method demonstrates more clearly the higher between-species variation in each family even for the two closest species.

### *Transition stages during molecular drive*

We have adopted a new method for analysing the patterns of variation at each nucleotide position considered independently amongst all clones of each family of the two pairs of species, *D. mauritiana versus D. simulans* and *D. yakuba versus D. teissieri*. This method of partitioning of variation (Table III; Figure 4) reveals both the fixation of some variants between these most closely-related pairs and, importantly, the transitional stages in the spread of variants. In Table III the patterns of variation at all individual nucleotide positions considered independently are classified in terms of the six stages (classes 1–6) in the spread of variant repeats through the sequence family and the species. The data include all available clones, irrespective of the loci, chromosomes or individuals from which they may come. We

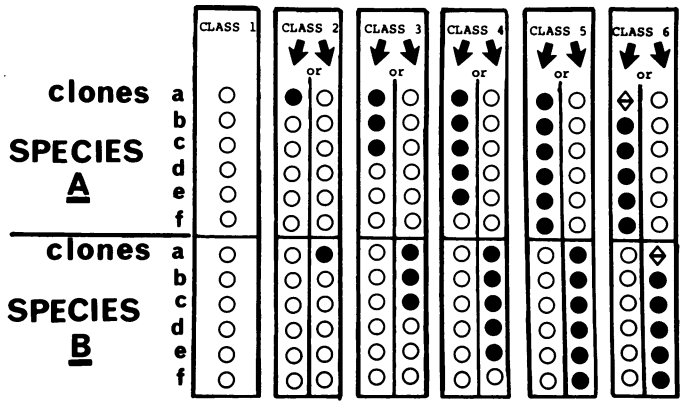


**Fig. 4.** Graphic representation of transition stages during the spread of new mutations. Classes 1–6 represent the patterns of distribution of mutations at individual nucleotide positions across clones a–f in two species A and B. Symbols, ●, ○ and ◇ represent nucleotide differences which can be A, T, G or C. Note that 95% of the positions that differ between the two pairs of species under comparison (see Table III) fall into these six classes. In classes 1–5 only two bases are found at a given position across all clones of a pair of species.

illustrate these classes by reference to the 360 family representatives of *D. mauritiana* and *D. simulans* (Figure 3) whose consensus sequences differ by only 12 positions. Space does not allow us to show the full extent of the variation amongst all clones of the 500 family between *D. mauritiana* and *D. simulans* and amongst clones of the 360 and 500 families of *D. yakuba* and *D. teissieri*; however all the data on the numbers and frequencies of positions that fall into each class are summarised in Table III.

A graphic representation of the six patterns of variation at individual nucleotide positions is given in Figure 4. For a given nucleotide position, the class 1 pattern is characterised by complete homogeneity across all clones randomly sampled from a pair of species (e.g., the first 18 positions in Figure 3). This class represents an absence of mutations (or their spread) in the supposed progenitor bases which are still shared by the two species. Class 2 represents rare mutations (or low levels of subsequent spread) resulting in the appearance of a minority of clones with a new mutation in a given position in one species whilst the other species remains homogenous for the progenitor base at the corresponding position (e.g., positions 158 and 176 in Figure 3). Class 3 covers those cases where no decision can be made between minority and majority frequencies in that a mutation and the progenitor base are in approximately equal frequencies, whilst the other species is homogeneous for one of the two bases. No examples of this are recorded in Figure 3 although examples can be found in the 500 family between *D. simulans* and *D. mauritiana* and in both families betwen *D. yakuba* and *D. teissieri* (see Table III). Class 4 includes those positions in which a mutation, which is apparently absent in one of the species, has replaced the progenitor base in the majority of members in the other species (e.g., positions 193 and 226, Figure 3). This interpretation assumes that the base which is in the minority in one species and which is homogeneous in the other species, is the progenitor base (see Discussion). Class 5 represents positions where the two species are internally homogeneous for bases that are diagnostically different for each of the species (e.g., position 151, Figure 3). This class represents the classic observation of concerted evolution (Arnheim, 1983; Dover, 1982), that is the final outcome of the dual processes (molecular drive) of intra-family homogenisation and population fixation of mutations that arose independently in one or both species. All subsequent mutations beyond this point are represented by the pattern represented by class 6 (e.g., position 42 in Figure 3).

**Table III.** Partitioning of the distribution of mutations at individual positions across all clones in the 360 and 500 family sequences of two closely related pairs of species (see also Figure 4)

| Classes of mutational distribution | | | 360 family | | 500 family | |
|---|---|---|---|---|---|---|
| Class | Species A | Species B | *mauritiana versus simulans* | *yakuba versus teissieri* | *mauritiana versus simulans* | *yakuba versus teissieri* |
| 1 | $N_1$ only | $N_1$ only | 278 (76.8) | 113 (56.4) | 348 (72.9) | 424 (76.1) |
| 2 | $N_1$ only | $N_1 > N_2$ | 72 (19.9) | 21 (10.3) | 104 (21.8) | 74 (13.3) |
| 3 | $N_1$ only | $N_1 = N_2$ | – | 2 (1.0) | – | 3 (0.6) |
| 4 | $N_1$ only | $N_2 > N_1$ | 6 (1.7) | 4 (2.0) | 2 (0.4) | 10 (1.8) |
| 5 | $N_1$ only | $N_2$ only | 5 (1.4) | 55 (27.5) | 2 (0.4) | 38 (6.8) |
| 6 | $N_1$ only | $N_2 > N_3$ | 1 (0.3) | 9 (4.5) | 1 (0.2) | 9 (1.6) |

$N_1$ represents any nucleotide and species A and B are interchangeable. See text and Figure 4 for an explanation of the classes of mutational distribution that represent transition stages. Figures refer to the absolute number of positions falling into each class. Figures in brackets represent percentages obtained by dividing the absolute numbers of positions within each class by the number of nucleotide positions available for comparison. In the case of the 360 family of *D. teissieri* a large deletion event resulted in there being only ~200 nucleotide positions available for comparison with the 360 family of *D. yakuba*. Those mutation events which could not be assigned an unambiguous location were excluded from the analysis. For example, in the cases of runs of a particular nucleotide, deletions and insertions representing the same nucleotide are topographically ambiguous. Species abbreviations are as in Table I.

More than 95% of the positions are classifiable in this way; one of two species can remain apparently unchanged whilst the other is undergoing replacement, and there are rarely more than two bases at a position across all clones of a pair of species.

From the data represented in Table III which compares similar numbers of repeats in two families evolving in parallel in the same two pairs of species, we draw the following conclusions.

(i) For each of the families a greater preponderance of classes 5 and 6 (representing a more extensive spreading of variant repeats) is found between *D. yakuba* and *D. teissieri* than between *D. mauritiana* and *D. simulans*. For example, in 47 of the 134 nucleotide positions within the 500 family consensus sequences and in 64 of the 91 positions within the 360 family of *D. yakuba* and *D. teissieri*, different bases have been fixed in the two species (classes 5 plus 6). Conversely, for both families classes 5 plus 6 represent only a small proportion of the total amount of variation exhibited between *D. mauritiana* and *D. simulans*.

(ii) There is a suggestion from the data that the rates of spread of mutations are faster in the 360 family than in the 500 family in all four species. For example, it would be expected that if the 360 family was evolving at the same rate as the 500 family, then only 32 of the 91 differences between *D. yakuba* and *D. teissieri* in the 360 family would fall within classes 5 and 6. Instead, the observed number of 64 for classes 5 and 6 signifies a faster rate of 360 family evolution.

*Mechanisms of turnover*

We assume that unequal exchange is the general mechanism of turnover in the tandem arrays of repeats which comprise each family. This is in keeping with the experimental proof of unequal exchanges within tandem arrays of rDNA which are embedded between some of the arrays of the 360 and 500 families

```
                          262                                          311
Teissieri 550  1   cattATatgtacTctcaaagagcctataAaCTAGAgacgaccggaagttc
               2   cattATatgtaccctcaaagagcctataGa*****gacgaccggaagttc
               3   cattATatgtaccctcaa*gagcctataGa*****gacgaccggaagttc
               4   cattATatgtaccctcaagagacctataGa*****gacgaccggaagttc
Teissieri 520  1   cattGGatgtaccctcaaagagcctataAaCTAGAgacgaccggaagttc
               2   cattGGatgtaccctcaaagagcctataAaCTAGAgacgaccggaagttc
               3   cattGGatgtaccctcaaagagcctTtaAaCTAGAgacgac*ggaagttc
               4   cattGGatgtaccct*aaagagcctataAaCTAGAgacgaccggaagttc

                          312                                          361
Teissieri 550  1   cgtggaTTCAGTGGATGTAATTGCTtAAATCTGCGTTGGt*gctggcgat
               2   cgtggaTTCAGTGGATGTAATTGCTCAAATCTGCGTTGGtAgctggcgat
               3   cgtggaTTCAGTGGATGTAATTGCTCAAATCTGCGTTGGt*gctAgcgat
               4   cgtggaTTCAGTGGATGTtATTGCTCAAATCTGCGTTGGt*gctggcgat
Teissieri 520  1   cgtggaCATACA**************************t*gctggcgat
               2   cgtggaCATACA**************************t*gctggcgat
               3   cgtggaCATACA**************************t*gctggcgat
               4   cgtggaCATACA**************************t*gctggcgat
```

**Fig. 5.** Sequence comparisons suggesting the operation of gene conversion in the *D. teissieri* 500 family. Compared sequences represent a 100 base long region of clones representing the 550 and 520 subfamilies of *D. teissieri* which correspond to positions 262–361 in the 500 family consensus sequence of Figure 2. Lower case letters illustrate sequence identity between the clones of the two subfamilies, minority mutations being denoted by upper case letters. Underlined capitals signify general sequence non-identity between the two subfamilies. Note that the first clone representing the 550 subfamily is unique in exhibiting the sequence AACTAGA at positions 290–296 which normally characterises the 520 subfamily at these positions. The comparison shows that sequences flanking this region conform to unambiguous segregation of the subfamilies. The domain of gene conversion (or double unequal exchange) could be either 50 bp in length (position 268–317) or the minimum 7 bp (AACTAGA).

on the X and Y chromosomes (Peacock *et al.*, 1977; Tartof, 1974; Coen *et al.*, 1982a, 1982b; Coen and Dover, 1983).

There are indications, however, that gene conversions have contributed to the evolution of the families. Figure 5 shows a comparison of sequences between the two subfamilies (550 and 520) of *D. teissieri* from positions 262 to 361. Each subfamily can be defined at the left end of either AT or GG (positions 266 and 267) and at the right end by the presence or absence of two non-homologous regions from 318 to 351. A cluster of seven bases AACTAGA 290–296, which is diagnostic for the 520 subfamily, appears in clone 1 of the 550 subfamily. Transfer between subfamilies implies either gene conversion or double unequal exchange of a domain with a maximum length of 50 bp (268–317) or a minimum length of 7 bp (290–296). Gene conversion of 290–296 would have favoured the addition rather than the deletion during mismatch repair: a bias for which precedent exists (Whitehouse, 1982). Similarly, a section of the 360 family of *D. orena* shows the transfer of a region of DNA, lying approximately between positions 328 and 363 (Figure 1), between clones which have dissimilar flanking sequences. Gene conversion domains of the order of tens of nucleotides have been implicated in other multigene families, such as the class I HLA genes (Strachan *et al.*, 1984; N,Guyen *et al.*, 1985).

## Discussion

In order to quantify the contribution of molecular drive to the biology of gene families, it is necessary (i) to dissociate the spreading effects of the internal dynamics of molecular turnover from external selection, and (ii) to find and analyse transition stages during the spread of variant members. These two aims can be achieved through the exploitation of non-coding families which are known to be subject to the same mechanisms of turnover as multigene families, but possibly in the absence of selection (for references, see Flavell, 1982; Singer, 1982; Brown and Dover, 1981; Miklos, 1984; Dover, 1982; Spradling and Rubin, 1981).

The 360 and 500 non-coding families exist in several thousand copies in each individual of species of the *melanogaster* species subgroup (Strachan *et al.*, 1982; Barnes *et al.*, 1978; Brutlag, 1980). The approximate order of magnitude difference between the within-species and between-species variation in most pair-wise species comparisons is indicative of the continual spread of new variants in each family in all species. Relative high intraspecific variation has been observed in other insects (Miklos and Gill, 1982; Miklos, 1984; Trick and Dover, 1984). Observed levels of sequence identity between repeats would depend on many parameters in each species, such as the rates, biases and units of turnover, generation time, family size, effective population size, and any physical constraints within the genome responsible for subfamily differentiation (Dover, 1982; Ohta and Dover, 1983, 1984). A blanket homogeneity in each family is neither expected nor observed.

The apparently even distribution of mutations within the 360 and 500 consensus sequences; the general 2:1 ratio of trans-

versions to transitions, and the general absence of substructure within the repeats of the 360 and 500 families, in each of the seven *Drosophila* species, suggests that each nucleotide position is free to mutate, and that the variants are free to spread as a consequence of family turnover, without strong selective or genomic constraints.

In some cases where subfamilies can be distinguished, rare exchanges by gene conversion have occurred between them. This situation is different from that of some genic and non-genic families which contain subfamilies that are differentiating almost independently (Kedes, 1979; Hood *et al.*, 1975; Anderson *et al.*, 1981; Jones and Kafatos, 1982; Brown and Dover, 1981; Brown, 1984; Miklos, 1984; Singer, 1982).

*Transition stages and rates of turnover*

Our method of analysis of the pattern of differences at any given nucleotide position across all sampled clones from a pair of species (class 1 – 6 Figure 4; Table III) reveals all expected transitions during the course of spread of variant repeats throughout the family and through the population. Classes 2, 3 and 4 represent intermediate levels between the two extremes of no replacement (class 1) and full replacement (classes 5 and 6).

It is significant that in classes 3 and 4 no more than two nucleotides show polymorphism in one species, when the other species is invariant in these positions. If mutation and spreading were operating at similar rates we would expect a consistently high level of within-species variation, possibly for variant repeats representing all four bases, at any one position, each of which might have spread to varying extents. The observation that the overwhelming number of base positions (~95%) fall into the six classes as found indicate that the rate of production of new variant repeats is a slower process than their rate of spread. The general paucity of transition stages (1% for classes 3 plus 4 representing between 20% and 80% of family replacement) indicates also that replacement is relatively fast. Calculations based on estimates of average sizes of populations (Ne = $10^5$, derived from levels of protein polymorphisms) and on average copy-numbers of the 360 and 500 families (between 3000 and 10 000 members), indicate that the rate of unequal exchange would have to be between $10^{-2}$ and $10^{-4}$ per generation to explain the data in Table III (Dover and Strachan, in preparation).

In both families, classes 5 and 6 are more abundant in the *D. yakuba – D. teissieri* comparison than in the *D. mauritiana – D. simulans* comparison. This is to be expected, assuming that the rate of spread of variant repeats is relatively constant, since *D. mauritiana* and *D. simulans* appear to be the most recently diverged pair of species within the *melanogaster* subgroup (for reviews, see Dover *et al.*, 1982; Tsacas *et al.*, 1984).

The numbers of nucleotide positions that fall within the higher classes in the 360 family are relatively more numerous than those of the 500 family in both pairs of species, suggesting that the 360 family is evolving at a faster rate *per se*. It is known that the 360 family is confined to the X chromosome (Brutlag, 1980; Peacock *et al.*, 1977) and that the 500 family is distributed on all four pairs of chromosomes (Strachan *et al.*, 1982). It is probable that a family on a single chromosome tends to homogeneity and fixation for a new variant more rapidly than a family dispersed on several chromosomes (Dover, 1982; Coen and Dover, 1983; Ohta and Dover, 1983, 1984).

A phylogeny based on the 360 and 500 families is concordant with the accepted phylogeny based on a variety of biological criteria (Dover *et al.*, 1982; Tsacas *et al.*, 1984). This concordance is consistent with the idea that each family has its own relatively constant rate of fixation of variants by molecular drive.

*Sequence divergence and function*

Several regions of high similarity have been located between sequences within the 360 family in *D. melanogaster* and sequences found in diverse genera, for example yeast centromeric sequences (Brutlag, 1980; Miklos and Gill, 1982; Fitzgerald-Hayes *et al.*, 1982). Furthermore, a sequence of dyad symmetry in the *melanogaster* 360 family, which appears to be bound tightly to embryo-specific proteins, has been described by Hsieh and Brutlag (1979a). Our data reveal that divergence between the sibling species, with respect to this sequence and also the sequence similar to yeast centromeres, is not obviously less than in the rest of the 360 repeat unit.

Relatively rapid divergence in sequence in multigene and non-genic DNA families does not mean, however, that such sequences lack biological significance, in that the co-evolution of other genes might be taking place whose products are involved with the function of a given family. Both our experimental and theoretical analysis of the dynamic spread of new variants in a family by molecular drive show that a sexual population evolves gradually and cohesively. That is, there are no large differences between individuals at any given generation, in the ratio of old to new variants per individual, during the long period it takes to completely replace a family of genes (Dover, 1982; Ohta and Dover, 1984). Hence, molecular drive provides the relaxed conditions for the natural selection of alleles of other genes that are more efficient at recognizing the newly emerging family of sequences. By this means biological function, dependent on several interacting genes or their products, can be maintained despite continual rounds of sequence divergence in the gene family. Conversely the mechanics and small biases of some turnover mechanisms that favour ancestral sequences can dramatically conserve sequences over long perids of time, in the absence of function and selection.

The detailed population dynamics of co-evolution at the molecular level, which is essentially an interaction between molecular drive and natural selection, and the extent to which molecular drive can affect the divergence and conservation of sequences are described elsewhere (Dover and Flavell, 1984; Ohta and Dover, 1984; Dover and Tautz, 1985).

## Materials and methods

*DNA cloning*

DNA was prepared from adult flies as described by Barnes *et al.* (1978). Repetitive DNA families were purified by two or three rounds of Hoechst 33258-CsCl equilibrium density gradient centrifugation according to Brown and Dover (1979). Highly purified family DNA was digested to completion by *Eco*RI or *Hae*III. In the case of the 360 family sequences, up to 40% of the total family complement was susceptible to cleavage by *Eco*RI while 70 – 80% of the total 500 family complement of each species was sensitive to digestion by the restriction enzyme used to isolate fragments for cloning. Digestion products were separated on a 1.6% agarose gel in Tris-borate buffer (89 mM Tris, 2.5 mM $Na_2$ EDTA, 89 mM boric acid) for 3 h at 10 V/cm. Appropriate restriction fragments representing the basic DNA repeat unit lengths were excised from the gel prior to electroelution of the DNA (McDonnell *et al.*, 1977), and purification of the DNA by phenol extraction.

Purified DNA was ligated into the *Eco*RI or *Sma*I site of M13mp8. Subsequent transformation of $CaCl_2$-treated JM101 cells was assayed using conventional IPTG/BCIG indicator plates and candidate positive clones were verified following transfer to a nitrocellulose filter, denaturation, then hybridisation with suitable nick-translated probes.

*DNA sequencing*

DNA sequencing was accomplished using the chain terminator method developed by Sanger *et al.* (1977) and detailed by Smith (1980). Under the conditions employed, generally 250 – 300 bases could be read unambiguously starting from the cloning site. The sequences of eight or more clones including four each in the two possible orientations was established for each DNA family in the species indicated in Table I.

*Analysis of DNA sequences*

Computer analysis was conducted on the Cambridge University IBM 3081 computer. Alignment of homologous sequences between species was achieved using a FORTRAN dot matrix program which sequentially matched all 5 bp units between the two sequences (Coen and Dover, 1982) and by use of a FORTRAN-IV version of the SEQA program of Goad and Kanehisa (1982). The presence of internal homologies including dyadic components within a repeat unit was checked using modifiied versions of the dot matrix program which entailed sequence matching of component units at a variety of different component unit lengths and different degrees of matching. Internal substructure was also investigated by recourse to a FORTRAN-IV overlap program which plotted the degree of matching between a sequence and derivatives obtained from it by sequentially staggering the sequence by one base at a time. A FORTRAN-IV DICT program constructs a dictionary of all the unique oligonucleotides present in the DNA sequence and a FORTRAN-IV ANALYSE program estimates the various nucleotide, dinucleotide and trinucleotide compositions of input sequence.

*Analysis of sequence variation*

Sequence variation was analysed by two methods. In the first of these, intraspecific variation was calculated by summing all nucleotide differences shown in the clones with reference to an intraspecific consensus sequence, then dividing this result by the total number of nucleotides available for comparison (INTRASPECIFIC VARIATION method 1). In this calculation all large deletions and insertions were treated as single events. A similar calculation was extended to obtain values for the interspecific variation by comparing individual intraspecific consensus sequences from pairs of species (INTERSPECIFIC VARIATION method 1). The second method considered the mean variation per nucleotide position. In this calculation all clones were compared against each other in pair-wise comparisons, for each nucleotide position considered independently. Variation was defined as the ratio of the number of pair-wise comparisons showing a clonal difference at one nucleotide position and the total number of pair-wise comparisons. The mean variation per nucleotide was then obtained by summing the individual variations at each nucleotide position and dividing by the total number of nucleotide positions (INTRASPECIFIC VARIATION method 2). In this calculation large deletions were considered to contribute to the sequence variation at all nucleotide positions encompassed by the deletion. Extension of this type of calculation to obtain values for interspecific sequence variation was limited to those species where the degree of sequence divergence between clones from the two species was very small, thereby permitting confident alignment of the clonal complements of both species and calculation of heterozygosity as defined above (INTERSPECIFIC VARIATION method 2).

## Acknowledgements

We owe our thanks to Nigel Harris and Meryl Lusher for the analysis and computation of the sequences and to Enrico Coen, Martin Trick, Gerald Franz and Steve D.M.Brown for their help and thoughtful discussion. We are grateful to June Hunt for patient typing of the drafts.

## References

Anderson,D.M., Scheller,R.H., Posakony,J.W., McAllister,L.B., Trabert,S.G., Beall,C., Britten,R.J. and Davidson,E.H. (1981) *J. Mol. Biol.*, **145**, 5-28.
Arnheim,N. (1983) in Nei,M. and Koehn,R.K. (eds.), *Evolution of Genes and Proteins*, Sinauer, pp. 38-61.
Barnes,S.R., Webb,D.A. and Dover,G.A. (1978) *Chromosoma*, **67**, 341-363.
Brown,S.D.M. (1984) in *Genetics: New Frontiers, Proceeding of the 15th International Congress of Genetics*, Oxford and IBH Publishing Co., pp. 221-234.
Brown,S.D.M. and Dover,G.A. (1979) *Nucleic Acids Res.*, **6**, 2423-2434.
Brown,S.D.M. and Dover,G.A. (1981) *J. Mol. Biol.*, **150**, 441-466.
Brutlag,D.L. (1980) *Annu. Rev. Genet.*, **14**, 121-144.
Coen,E.S. and Dover,G.A. (1982) *Nucleic Acids Res.*, **10**, 7017-7026.
Coen,E.S. and Dover,G.A. (1983) *Cell*, **33**, 849-855.
Coen,E.S., Strachan,T. and Dover,G.A. (1982a), *J. Mol. Biol.*, **158**, 17-35.
Coen,E.S., Thoday,J.M. and Dover,G.A. (1982b) *Nature*, **295**, 564-568.
Dover,G.A. (1982) *Nature*, **299**, 111-117.
Dover,G.A. and Flavell,R.B. (1984) *Cell*, **38**, 622-623.
Dover,G.A. and Tautz,D. (1985) in Clarke,B.C., Robertson,A. and Jeffreys,A.J. (eds.), *The Evolution of DNA Sequences, Phil. Trans. Roy. Soc.*, in press.
Dover,G.A., Brown,S.D.M., Coen,E.S., Dallas,J., Strachan,T. and Trick,M. (1982) in Dover,G.A. and Flavell,R.B. (eds.), *Genome Evolution*, Academic Press, London, pp. 343-372.
Fedoroff,N.V. (1979) *Cell*, **16**, 697-710.
Fitzgerald-Hayes,M., Clarke,L. and Carbon,J. (1982) *Cell*, **29**, 235-244.
Flavell,R.B. (1982) in Dover,G.A. and Flavell,R.B. (eds.), *Genome Evolution*, Academic Press, London, pp. 301-323.
Goad,W.B. and Kanehisa,X. (1982) *Nucleic Acids Res.*, **10**, 247-264.
Hood,L., Campbell,J.H. and Elgin,S.C.R. (1975) *Annu. Rev. Genet.*, **9**, 305-353.
Hsieh,T.H. and Brutlag,D. (1979a) *Proc. Natl. Acad. Sci. USA*, **76**, 726-730.
Hsieh,T.S. and Brutlag,D.L. (1979b) *J. Mol. Biol.*, **135**, 465-481.
Jones,W.C. and Kafatos,F.C. (1982) *J. Mol. Evol.*, **19**, 87-103.
Kedes,L.H. (1979) *Annu. Rev. Biochem.*, **48**, 837-870.
Long,E.H. and Dawid,I.B. (1980) *Annu. Rev. Biochem.*, **49**, 727-764.
McDonnell,M.W., Simon,M.N. and Studier,F.W. (1977) *J. Mol. Biol.*, **110**, 119-130.
Miklos,G.L.G. (1984) *Evol. Biol.*, in press.
Miklos,G.L.G. and Gill,A.C. (1982) *Genet. Res. Camb.*, **39**, 1-30.
N,Guyen,C., Sodoyer,R., Trucy,J., Strachan,T. and Jordan,B.R. (1985) *Immunogenetics*, in press.
Ohta,T. (1980) *Lecture Notes in Biomathematics*, Vol. **37**, published by Springer-Verlag, Berlin.
Ohta,T. and Dover,G.A. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 4079-4083.
Ohta,T. and Dover,G.A. (1984) *Genetics*, **108**, 501-521.
Peacock,W.J., Lohe,A.R., Gerlach,W.L., Dunsmuir,P., Dennis,E.S. and Appels,R. (1977) *Cold Spring Harbor Symp. Quant. Biol.*, **43**, 1121-1135.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
Singer,M.F. (1982) *Int. Rev. Cytol.*, **76**, 67-112.
Smith,A.J.H. (1980) *Methods Enzymol.*, **65**, 560-580.
Spradling,A.C. and Rubin,G.M. (1981) *Annu. Rev. Genet.*, **15**, 219-264.
Strachan,T., Coen,E.S., Webb,D.A. and Dover,G.A. (1982) *J. Mol. Biol.*, **158**, 37-54.
Strachan,T., Sodoyer,R., Damotte,M. and Jordan,B.R. (1984) *EMBO J.*, **3**, 887-894.
Tartof,K. (1974) *Cold Spring Harbor Symp. Quant. Biol.*, **38**, 491-500.
Trick,M. and Dover,G.A. (1984) *J. Mol. Evol.*, **20**, 322-329.
Tsacas,L., David,J., Lachaise,D., Lemeunier,F. and Ashburner,M. (1984) *The Genetics and Biology of Drosophila*, Vol. **3e**, in press.
Whitehouse,H.L.K. (1982) *Genetic Recombination*, published by Wiley.

*Received on 8 March 1985; revised on 30 April 1985*