Complete Sequences of the rRNA Genes of *Drosophila melanogaster*[1]

Diethard Tautz,[2] John M. Hancock, David A. Webb, Christiane Tautz,[2] and Gabriel A. Dover

Department of Genetics, University of Cambridge

In this, the first of three papers, we present the sequence of the ribosomal RNA (rRNA) genes of *Drosophila melanogaster.* The gene regions of *D. melanogaster* rDNA encode four individual rRNAs: 18S (1,995 nt), 5.8S (123 nt), 2S (30 nt), and 28S (3,945 nt). The ribosomal DNA (rDNA) repeat of *D. melanogaster* is AT rich (65.9% overall), with the spacers being particularly AT rich. Analysis of DNA simplicity reveals that, in contrast to the intergenic spacer (IGS) and the external transcribed spacer (ETS), most of the rRNA gene regions have been refractory to the action of slippage-like events, with the exception of the 28S rRNA gene expansion segments. It would seem that the 28S rRNA can accommodate the products of slippage-like events without loss of activity. In the following two papers we analyze the effects of sequence divergence on the evolution of (1) the 28S gene "expansion segments" and (2) the 28S and 18S rRNA secondary structures among eukaryotic species, respectively. Our detailed analyses reveal, in addition to unequal crossing-over, (1) the involvement of slippage and biased mutation in the evolution of the rDNA multigene family and (2) the molecular coevolution of both expansion segments and the nucleotides involved with compensatory changes required to maintain secondary structures of RNA.

Introduction

The ribosomal DNA (rDNA) gene family of *Drosophila* species undergoes continual rounds of unequal crossing-over leading to the concomitant spread of mutations (molecular drive) through the family (homogenization) and through the species (fixation) (for reviews, see Coen et al. 1982; Dover 1982; Arnheim 1983; Flavell 1986). Recent detailed examination of the organization and sequences of the intergenic spacers (IGSs) (formerly called the nontranscribed spacer [NTS]) and external transcribed spacers (ETSs) from four species of the genus *Drosophila* (Tautz et al. 1987) shows that both minor point-mutational and major structural alterations have accumulated in each species. Some of the structural alterations are due to the activities of slippage-like mechanisms in different regions of the spacers, mechanisms that result in gains and losses of short sequence motifs. Gains and losses of longer stretches of DNA also occur, frequently in regions that are either of functional significance (sites of transcription promotion and initiation) or positionally conserved over long periods of time (ETS). How are essential functions maintained during continual sequence divergence?

[1] Key words: rDNA sequences, cryptic simplicity, slippage, *Drosophila melanogaster.*

[2] Current address: Max-Planck-Institut für Entwicklungsbiologie, Spemannstrasse 35, D-7400 Tübingen, Federal Republic of Germany.

Address for correspondence and reprints: Gabriel A. Dover, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, United Kingdom.

Mol. Biol. Evol. 5(4):366–376. 1988.

0737-4038/88/0504-0004$02.00

The facility with which the transcription-promoter regions undergo rapid sequence evolution in both minor and major ways, despite their functional importance, is probably due to their relatively simple molecular interactions with RNA polymerase I and its transcription-initiation factors. This has led to a molecular coevolution between the multiple promoters and the genes for the polymerase I complex. Evidence for this is derived from the incompatibility, in several species complexes of animals and plants, between the promoters of one species and the polymerase complex of another (for reviews, see Arnheim 1983; Dover and Flavell 1984; Reeder 1984; Gerbi 1985; Moss et al. 1985). It is probable that this example of molecular coevolution arose out of the gradual and cohesive manner in which variant repeats are spread by molecular drive through the rDNA families of all individuals of a population (Dover 1982; Ohta and Dover 1984; Tautz et al. 1987). The dynamics of rDNA change at the population level would permit natural selection to ensure the coevolution of genes of other cellular components involved in a given function (for detailed arguments, see Dover and Flavell 1984).

In the light of such changes in the spacers and of the emergence of coevolutionary changes elsewhere in the genome, it is important to examine the types and rates of changes occurring within the rDNA gene regions. Such regions, with the exceptions of specific 28S gene regions known as "expansion segments" (Clark et al. 1984; Gerbi 1985), are known to be under more severe constraints than are the various spacers. Are the same mechanisms of mutation and same processes of fixation operating in the ribosomal RNA (rRNA) gene regions? What types of coevolutionary changes are taking place in such constrained regions, and what might be the dynamics of their establishment?

To answer these questions, we present first the complete sequence of the rRNA genes and of the internal transcribed spacers (ITS) of *D. melanogaster* and consider the effects of point mutation and slippage-like mechanisms on their sequence evolution. Only fragments of sequences of the rRNA genes have been previously available (see text). In the two accompanying papers we first show that the set of expansion segments in the 28S gene are coevolving in any one species, despite the high rate of slippage-generated sequence divergence within each segment; second, we consider the effects of sequence evolution on the secondary structure and coevolution of compensatory mutations in *D. melanogaster* rRNAs (Hancock and Dover 1988; Hancock et al. 1988).

Material and Methods

Cloning and Sequencing

For our sequencing studies we used the clone pDm238, which was a gift from D. M. Glover (Imperial College, London) and contains a complete *Drosophila melanogaster* rDNA repeat unit inserted in the *Eco*RI site of pBr322. We sequenced the entire repeat unit, including the spacers. The sequences of the IGS and ETS have been published elsewhere (Tautz et al. 1987). Sequencing was performed by the dideoxy chain-termination method (Sanger et al. 1977) using single-stranded DNA produced via the pEMBL vector system (Dente et al. 1983). Both strands were sequenced by first subcloning several 2–3-kb-large pieces and then producing overlapping deletions from each clone by Bal31 exonuclease digestion (the principle of the method is outlined by Guo and Wu [1982]). To ensure that no small fragments were lost in the subcloning process, we used overlapping fragments to sequence across all the restriction sites utilized for subcloning. All regions could be sequenced except for a small region in the 28S rRNA gene that was not readable from either strand (positions 5052–5056). To fill this gap, we have adopted the sequence of Delanversin and Jacq (1983) in this region (see Discussion). The position of the 5′ end of the 18S rRNA was taken from Simeone and Boncinelli (1984), the 3′ end from Jordan et al. (1980), the ends of the 5.8S and 2S rRNA genes from Pavlakis et al. (1979), and the 3′ end of the 28S rRNA gene from Mandal and Dawid (1981). The 5′ end of the 28S rRNA gene was tentatively assigned by assuming homology with the mouse 28S rRNA gene (Ware et al. 1983).

Sequence Analysis

DNA simplicity analysis was carried out using the SIMPLE program (Tautz et al. 1986) and run in Fortran 77 on the Cambridge University IBM 3081. This program counts the number of repeats of a tri- or tetranucleotide motif surrounding the central nucleotide in a window of 64 bp, as the window is moved one base at a time from one end of a sequence to the other. At the same time it reveals the statistical significance of the repeat numbers by comparison with equivalent numbers derived from 10 times 10 kb of randomized sequences of the same composition as the natural sequence. The randomization procedure consisted of producing random sequences from the percentage occupation of each of the four bases in a given natural sequence. No attempt was made to preserve nearest-neighbor frequencies because these might arise, as with other internal DNA structures, by the very process of slippage under examination. Our null hypothesis is that, for any given composition, no repetition over and above that produced by chance occurs. If a sequence is to be considered as cryptically simple, it should have a relative simplicity factor (RSF) that is significantly >1.0. When RSFs are judged to be statistically significant ($P < 0.003$), that is, if they are greater than the randomized value by more than 3 SDs of the values obtained in the 10 randomized runs, the natural sequence is internally repetitious, often for motifs that are scrambled among themselves (cryptic simplicity) (for details, see Tautz et al. 1986). Since the program is designed to find high relative numbers of direct repeats, RSF values that are significantly <1.0 will occur in regions that have a large number of inverted repeats. In essence the program is designed to reveal scrambled permutations of direct repetitive short motifs (cryptic simplicity), which are not as visible to the eye as tandem runs of a particular motif (pure simplicity).

RSF values are derived from a complete sequence rather than from separate regions within it, except where otherwise stated. In addition to the assessment of RSFs (i.e., the overall level of simplicity of a natural sequence relative to that of the randomized sequences), the program can display graphically the repeat numbers surrounding any given position (see fig. 2; figs. in the accompanying paper by Hancock and Dover 1988). For further details, see text and Tautz et al. (1986).

Results

Sequence

The sequences of the rRNA genes and ITS of *Drosophila melanogaster* are presented in figure 1. The sequence proceeds from the 5′ end of 18S rRNA through ITS1, 5.8S rRNA, ITS2a, 2S rRNA, and ITS2 to the 3′ end of 28S rRNA (see fig. 2). The boundaries of the individual rRNAs are marked. The lengths, positions (5′ end of 18S = 1), and base compositions of the individual regions are listed in table 1. The sequence is numbered 1 to 7232 starting from the 5′ end of the 18S rRNA gene. *Drosophila melanogaster* is the first species for which the sequence of an entire rDNA unit is

```
<18S rRNA----------------------------------------------------------------------------------------------------
ATTCTGGTTG ATCCTGCCAG TAGTTATATG CTTGTCTCAA AGATTAAGCC ATGCATGTCT AAGTACACAC GAATTAAAAG TGAAACCGCA AAAGGCTCAT  100
TATATCAGTT ATGGTTCCTT AGATCGTTAA CAGTTACTTG GATAACTGTG GTAATTCTAG AGCTAATACA TGCAATTAAA ACATGAACCT TATGGGACGT  200
GTGCTTTTAT TAGGCTAAAA CCAAGCGATC GCAAGATCGT TATATTGGTT GAACTCTAGA TAACATGCAG ATCGTATGGT CTTGTACCGA CGACAGATCT  300
TTCAAATGTC TGCCCTATCA ACTTTTGATG GTAGTATCTA GGACTACCAT GGTTGCAACG GGTAACGGGG AATCAGGGTT CGATTCCGGA GAGGGAGCCT  400
GAGAAACGGC TACCACATCT AAGGAAGGCA GCAGGCGCGT AAATTACCCA CTCCCAGCTC GGGGAGGTAG TGACGAAAAA TAACAATACA GGACTCATAT  500
CCGAGGCCCT GTAATTGGAA TGAGTACACT TTAAATCCTT TTACAAGGAC CAATTGGAGG GCAAGTCTGG TGCCAGCAGC CGCGGTAATT CCAGCTCCAA  600
TAGCGTATAT TAAAGTTGTT GCGGTTAAAA CGTTCGTAGT TGAACTTGTG CTTCATACGG GTAGTACAAC TTACAATTGT GGTTAGTACT ATACCTTTAT  700
GTATGTAAGC GTATTACCGG TGGAGTTCTT ATATGTGATT AAATACTTGT ATTTTTTCAT ATGTTCCTCC TATTTAAAAA CCTGCATTAG TGCTCTTAAA  800
CGAGTGTTAT TGTGGGCCGG TACTATTACT TTGAACAAAT TAGAGTGCTT AAAGCAGGCT TCAAATGCCT GAATATTCTG TGCATGGGAT AATGAAATAA  900
GACCTCTGTT CTGCTTTCAT TGGTTTTCAG ATCAAGAGGT AATGATTAAT AGAAGCAGTT TGGGGGCATT AGTATTACGA CGCGAGAGGT GAAATTCTTG 1000
GACCGTCGTA AGACTAACTT AAGCGAAAGC ATTTGCCAAA GATGTTTTCA TTAATCAAGA ACGAAAGTTA GAGGTTCGAA GGCGATCAGA TACCGCCCTA 1100
GTTCTAACCA TAAACGATGC CAGCTAGCAA TTGGGTGTAG CTACTTTTAT GGCTCTCTCA GTCGCTTCCG GGAAACCAAA GCTTTTTGGG CTCCGGGGGA 1200
AGTATGGTTG CAAAGCTGAA ACTTAAAGGA ATTGACGGAA GGGCACCACC AGGAGTGGAG CCTGCGGCTT AATTTGACTC AACACGGGAA AACTTACCAG 1300
GTCGAACATA AGTGTGTAAG ACAGATTGAT AGCTCTTTCT CGAATCTATG GGTGGTGGTG CATGGCCGTT CTTAGTTCGT GGAGTGATTT GTCTGGTTAA 1400
TTCCGATAAC GAACGAGACT CAAATATATT AAATAGATAT CTTCAGGATT ATGGTGCTGA AGCTTATGTA GCCTTCATTC ATGTTGGCAG TAAAATGCTT 1500
ATTGTGTTTG AATGTGTTTA TGTAAGTGGA GCCGTACCTG TTGGTTTGTC CCATTATAAG GACACTAGCT TCTTAAATGG ACAAATTGCG TCTAGCAATA 1600
ATGAGATTGA GCAATAACAG GTCTGTGATG CCCTTAGATG TCCTGGGCTG CACGCGCGCT ACAATGAAAG TATCAACGTG TATTTCCTAG ACCGAGAGGT 1700
CCGGGTAAAC CGCTGAACCA CTTTCATGCT TGGGATTGTG AACTGAAACT GTTCACGATG AACTTGGAAT TCCCAGTAAG TGTGAGTCAT TAACTCGCAT 1800
TGATTACGTC CCTGCCCTTT GTACACACCG CCCGTCGCTA CTACCGATTG AATTATTTAG TGAGGTCTCC GGACGTGATC ACTGTGACGC CTTGCGTGTT 1900

-----------------------------------------------------------------------------------------------18S rRNA><ITS1
ACGGTTGTTT CGCAAAAGTT GACCGAACTT GATTATTTAG AGGAAGTAAA AGTCGTAACA AGGTTTCCGT AGGTGAACCT GCGGAAGGAT CATTATTGTA 2000
TAATATCCTT ACCGTTAATA AATATTTGTA ATTATACAAA TAAAAACAAT TTACCAAAAT AAAAATATAA CAAAATGATT CCATGGAATC AAAAGTTAAA 2100
ATCAAAATAA AACGAAGATG GGTTTTATTT ATATAGTTAG TGTGGGGCTT GGCAACCTCA TAAAAAGATT TTAACATTTC TAATGTATGT TGTGCGTATT 2200
TGTGGCGAGT ACTTACAACA ACGGCGTTTC CTATAAAAAT AATGTTTCGA ACATGAAAAT CGAAGAAACA AAATTCGAAA GTGGAAGTCG AATCAAAATA 2300
AAATAATTTC GAATGTGTGG TAATCATCGA AATAAGTGTT AATATAATTG GTAGATATTA ACTAATTTTT AAAATTTGTG TGTATTTATT ACTATACACG 2400
CGTTGCGAAT ATGTATTGTT CATCTTAGTT ATGGGCATAC GTTGGCTAAT GCAACAACCT GAAATAAACA ATGTTGTACC TGGCATCCAT CAGGTTAATG 2500
TTTTATATAA ATTGCAGTAT GTGTCACCCA AAATAGCAAA CCCCATAACC AACCAGATTA TTATGATACA TAATGCTTAT ATGAAACTAA GACATTTCGC 2600
AACATTTATT TTAGGTATAT AAATACATTT ATTGAAGGAA TTGATATATG CCAGTAAAAT GGTGTATTTT TAATTTCTTT CAATAAAAAC ATAATTGACA 2700

------------------ITS1><5.8S rRNA----------------------------------------------------------------------------
TTATATAAAA ATGAATTATA AAACTCTAAG CGGTGGATCA CTCGGCTCAT GGGTCGATGA AGAACGCAGC AAACTGTGCG TCATCGTGTG AACTGCAGGA 2800

--------------------------------------5.8S rRNA><ITS 2a-----------------ITS 2a><2S rRNA----------------------
CACATGAACA TCGACATTTT GAACGCATAT CGCAGTCCAT GCTGTTATGT ACTTTAATTA ATTTTATAGT GCTGCTTGGA CTACATATGG TTGAGGGTTG 2900

-><ITS 2-----------------------------------------------------------------------------------------------------
TAAGACTATG CTAATTAAGT TGCTTATAAA TTTTTATAAG CATATGGTAT ATTATTGGAT AAATATAATA ATTTTTATTC ATAATATTAA AAAATAAATG 3000
AAAAACATTA TCTCACATTT GAATGTGAAA AACGAAGAGA AATATTTTCT TTTTCAATCA AATAATACTG AGAAATGTCT AGCATAAAAA ATTGAAATAT 3100
TTTTCATCTA GAATTGTCTC TTATTAATGA TTCGGAAATA GAAAAATCTT GGTTATGTTA TTATTCTTCG TTGGTTCGTT AAAAATGGAT AAATAAAAAC 3200

-----------------------------------------------------------------------------------------ITS 2><28S rRNA-----
TTTGCATACA AGAATTAATA AAAATGTTAT AACGAATTTA ATTAAATGTT TTATCATTAT ATATAAAGAA TTTATGGCAA GATAAAGTTA TATACAACCT 3300
CAACTCATAT GGGACTACCC CCTGAATTTA AGCATATTAA TTAGGGGAGG AAAAGAAACT AACAAGGATT TTCTTAGTAG CGGCGAGCGA AAAGAAAACA 3400
GTTCAGCACT AAGTCACTTT GTCTATATGG CAAATGTGAG ATGCAGTGTA TGGAGCGTCA ATATTCTAGT ATGAGAAATT AACGATTTAA GTCCTTCTTA 3500
AATGAGGCCC GTATAACGTT AATGATTACT AGATGATGTT TCCAAAGAGT CGTGTTGCTT GATAGTGCAG CACTAAGTGG GTGGTAAACT CCATCTAAAA 3600
CTAAATATAA CCATGAGACC GATAGTAAAC AAGTACCGTG AGGGAAAGTT GAAAAGAACT CTGAATAGAG AGTTAAACAG TACGTGAAAC TGCTTAGAGG 3700
TTAAGCCCGA TGAACCTGAA TATCCGTTAT GGAAAATTCA TCATTAAAAT TGTAATATTT AAATAATATT ATGAGAATAG TGTGCATTTT TTCCATATAA 3800
GGACATTGTA ATCTATTAGC ATATACCAAA TTTATCATAA AATATAACTT ATAGTTTATT CCAATTAAAT TGCTTGCATT TTAACACAGA ATAAATGTTA 3900
TTAATTTGAT AAAGTGCTGA TAGATTTATA TGATTACAGT GCGTTAATTT TTCGGAATTA TATAATGGCA TAATTATCAT TGATTTTTGT GTTTATTATA 4000
TGCACTTGTA TGATTAACAA TGCGAAAGAT TCAGGATACC TTCGGGACCC GTCTTGAAAC ACGGACCAAG GAGTCTAACA TATGTGCAAG TTATTGGGAT 4100
ATAAACCTAA TAGCGTAATT AACTTGACTA ATAATGGGAT TAGTTTTTTA GCTATTTATA GCTAATTAAC ACAATCCCGG GGCGTTCTAT ATAGTTATGT 4200
ATAATGTATA TTTATATTAT TTATGCCTCT AACTGGAACG TACCTTGAGC ATATATGCTG TGACCCGAAA GATGGTGAAC TATACTTGAT CAGGTTGAAG 4300
TCAGGGGAAA CCCTGATGGA AGACCGAAAC AGTTCTGACG TGCAAATCGA TTGTCAGAAT TGAGTATAGG GGCGAAAGAC CAATCGAACC ATCTAGTAGC 4400
TGGTTCCTTC CGAAGTTTCC CTCAGGATAG CTGGTGCATT TTAATATTAT ATAAAATAAT CTTATCTGGT AAAGCGAATG ATTAGAGGCC TTAGGGTCGA 4500
AACGATCTTA ACCTATTCTC AAACTTTAAA TGGGTAAGAA CCTTAACTTT CTTGATATGA AGATCAAGGT TATGATATAA ATGTCCCAGT GGGCCACTTT 4600
TGGTAAGCAG AACTGGCGCT GTGGGATGAA CCAAACGTAA TGTTACGTGC CCAAATTAAC AACTCATGCA GATACCATGA AAGGCGTTGG TTGCTTAAAA 4700
CAGCAGGACG GTGATCATGG AAGTCGAAAT CCGCTAAGGA GTGTGTAACA ACTCACCTGC CGAAGCAACT AGCCCTTAAA ATGGATGGCG CTTAAGTTGT 4800
ATACCTATAC ATTACCGCTA AAGTAGATGA TTTATATTAC TTGTGATATA AATTTTGAAA CTTTAGTGAG TAGGAAGGTA CAATGGTATG CGTAGAAGTG 4900
TTTGGCGTAA GCCTGCATGG AGCTGCCATT GGTACAGATC TTGGTGGATA GTAGCAAATA ATCGAATGAG AGCCTTGGAG GACTGAAGTG GAGAAGGGTT 5000

---------------------------------------------------------------------------------*
TCGTGTGAAC AGTGGTTGAT CACGAGTTAG TCGGTCCTAA GTTCAAGGCG AAAGCGAAAA TTTTCAAGTA AAACAAAAAT GCCTAACTAT ATAAACAAAG 5100

                      *--------------------------------------------------------------------------------------
CGAATTATAA TACACTTGAA TAATTTTGAA CGAAAGGGAA TACGGTTCCA ATTCCGTAAC CTGTTGAGTA TCCGTTTGTT ATTAAATATG GGCCTCGTGC 5200
TCATCCTGGC AACAGGAACG ACCATAAAGA AGCCGTCGAG AGATATCGGA AGAGTTTTCT TTTCTGTTTT ATAGCCGTAC TACCATGGAA GTCTTTCGCA 5300
GAGAGATATG GTAGATGGGC TAGAAGAGCA TGACATATAC TGTTGTGTCG ATATTTTCTC CTCGGACCTT GAAAATTTAT GGTGGGGACA CGCAAACTTC 5400
TCAACAGGCC GTACCAATAT CCGCAGCTGG TCTCCAAGGT GAAGAGTCTC TAGTCGATAG AATAATGTAG GTAAGGGAAG TCGGCAAATT AGATCCGTAA 5500
CTTCGGGATA AGGATTGGCT CTGAAGATTG AGATAGTCGG CCTTGATTGG GAAACAATAA CATGGTTTAT GTGCTCGTTC TGGGTAAATA GAGTTTCTAG 5600
CATTTATGTT AGTTACTTGT TCCCCGGATA GTTTAGTTAC GTAGCCAATT GTGGAACTTT CTTGCTAAAA TTTTTAAGAA TACTATTTGG GTTAAACCAA 5700
TTAGTTCTTA TTAATTATAA CGATTATCAA TTAACAATCA ATTCAGAACT GGCACGGACT TGGGGAATCC GACTGTCTAA TTAAAACAAA GCATTGTGAT 5800
GGCCCTAGCG GGTGTTGACA CAATGTGATT TCTGCCCAGT GCTCTGAATG TCAAAGTGAA GAAATTCAAG TAAGCGCGGG TCAACGGCGG GAGTAACTAT 5900
GACTCTCTTA AGGTAGCCAA ATGCCTCGTC ATCTAATTAG TGACGCGCAT GAATGGATTA ACGAGATTCC TACTGTCCCT ATCTACTATC TAGCGAAACC 6000
ACAGCCAAGG GAACGGGCTT GGAATAATTA GCGGGGAAAG AAGACCCTTT TGAGCTTGAC TCTAATCTGG CAGTGTAAGG AGACATAAGA GGTGTAGAAT 6100
AAGTGGGAGA TATTAGACCT CGGTTTGGTA TCGTCAATGA AATACCACTA CTCTTATTGT TTCCTTACTT ACTTGATTAA ATGGAACGTG TATCATTTCC 6200
TAGCCATTAT ACGGATATAT TTATTATATC TTATGGTATT GGGTTTTGAT GCAAGCTTCT TGATCAAAGT ATCACGAGTT TGTTATATAA TCGCAAACAA 6300
ATTCTTTAAT AAAACGATGC ATTTATGTAT TTTTGATTTG AAAATTTGGT ATAACTCCAA TTACTCAGGT ATGATCCAAT TCAAGGACAT TGCCAGGTAG 6400
GGAGTTTGAC TGGGGCGGTA CATCTCTCAA ATAATAACGG AGGTGTCCCA AGGCCAGCTC AGTGCGGACA GAAACCACAC ATAGAGCAAA AGGGCAAATG 6500
CTGACTTGAT CTCGGTGTTC AGTACACACA GGGACAGCAA AAGCTCGGCC TATCGATCCT TTTGGTTTAA AGAGTTTTTA ACAAGAGGTG TCAGAAAAGT 6600
TACCATAGGG ATAACTGGCT TGTGGCGGCC AAGCGTTCAT AGCGACGTCG CTTTTTGATC CTTCGATGTC GGCTCTTCCT ATCATTGTGA AGCAAAATTC 6700
ACCAAGCGTT GGATTGTTCA CCCATGCAAG GGAACGTGAG CTGGGTTTAG ACCGTCGTGA GACAGGTTAG TTTTACCCTA CTAATGACAA AACGTTGTTG 6800
CGACAGCATT CCTGCGTAGT ACGAGAGGAA CCCGCAGGTA CGGACCAATG GCACAATACT TGTTCGAGCG AACAGTGGTA TGACGCTACG TCCGTTGGAT 6900
TATGCCTGAA CGCCTCTAAG GTCGTATCCG TGCTGGACTG CAATGATAAA TAAGGGGCAA TTTGCATTGT ATGGCTTCTA AACCATTTAA AGTTTATAAT 7000
TTACTTTATA AACGACAATG GATGTGATGC CAATGTAATT TGTAACATAG TAAATTGGGA GGATCTTCGA TCACCTGATG CCGCGCTAGT TACATATAAA 7100
AGCATTATTT AATACAATGA CAAAGCCTAG AATCAATTGT AAACGACTTT TGTAACAGGC AAGGTGTTGT AAGTGGTTGA GCAGCTGCCA TACTGCGATC 7200
CACTGAAGCT TATCCTTTGC TTGATGATTC GA                                                                         7232
```

FIG. 1.—DNA sequence of the rRNA genes and IGS spacers of *Drosophila melanogaster.* Boundaries of individual rRNA genes and internal transcribed spacers are marked. The region between the asterisks is processed out of the mature 28S rRNA.

Table 1
Positions (18S 5′ End = 1), Lengths, Base Compositions, and RSFs of rRNA Genes and ITSs in *Drosophila melanogaster* rDNA

Sequence Region	Position	Length	% A-T	RSF[a]
IGS		3,632[b]	71.0	1.27
ETS		864	76.0	1.11
18S rRNA	1–1995	1,995	57.5	0.96
ITS 1	1996–2721	726	73.0	1.07
5.8S rRNA	2722–2844	123	49.6	0.85
ITS 2a	2845–2872	28	82.2	...[c]
2S rRNA	2873–2902	30	56.7	...[c]
ITS 2	2903–3287	385	80.0	0.86
28S rRNA	3288–7232	3,945	60.9	1.14
Coding region		7,232	62.1	
rDNA		11,728[b]	65.9	

[a] For definition, see text and Tautz et al. (1986). See Material and Methods for a note on the significance of values <1.0.

[b] Lengths vary depending on the number of spacer subrepeats (Coen et al. 1982). The values quoted here are determined for the copy of the rDNA repeat cloned in pDm 238.

[c] Sequence too short to analyze.

known (for IGS sequences, see Tautz et al. 1987). This poses a problem as to the best way to proceed with the numbering of nucleotide positions in a way that would be useful for purposes of comparison with other species. This problem is discussed later in more detail. Our sequence for the region encoding 5.8S and 2S rRNA contains three differences (positions 2921, 2931, and 2943–2944) from the previously published DNA sequence (Jordan et al. 1976; Pavlakis et al. 1979), all of them in ITS2. Our 28S rRNA sequence contains six differences (positions 5082, 5864, 5899, 5981, 6046, and 6082) from previously published sequences (Mandal and Dawid 1981; Dawid and Rebbert 1981; Roiha and Glover 1981; Roiha et al. 1981; Delanversin and Jacq 1983). Compared with previous sequences (Jordan et al. 1980; Youvan and Hearst 1981; Simeone et al. 1985), that for 18S rRNA contains one substitution, three insertions, and four deletions.

DNA Simplicity Analysis

The simplicity profile of the entire *D. melanogaster* rDNA repeat is presented in figure 2. This is a graphic display of the weighted number of repeats of all naturally occurring motifs within a window of 64 bp surrounding any given base, averaged over 10 bases. The solid line through the display represents the mean overall simplicity scores for 10 kb of 10 randomized runs (this line is not meant to be a rigid statistical test; for further details, see Tautz et al. 1986). The boundaries of individual sequence domains are indicated in the accompanying map.

RSFs of the individual sequence domains of the rRNA coding region are presented in table 1. RSFs are a measure of the extent to which short direct repeats are represented in a given region over and above that detected in 10 randomized sequences of the same nucleotide composition (for details, see Tautz et al. 1986).

From figure 2 and table 1 it can be seen that the 5.8S and 18S rRNA genes, along with the ITS2 region, are not cryptically simple and that slippage-like mechanisms of variation do not seem to be operating to any great extent within them. By contrast, the IGS and ETS are cryptically simple (see also Tautz et al. 1987 for analysis), and

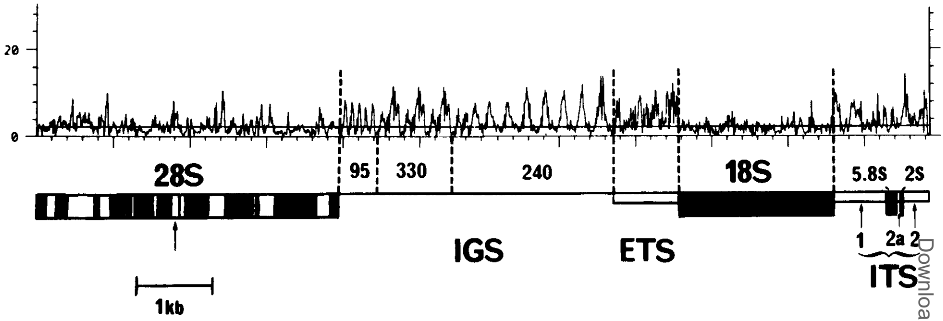

FIG. 2.—Simplicity profile and map of the *Drosophila melanogaster* rDNA repeat. *Top,* Simplicity profile. The solid line represents the mean value of the overall simplicity factors obtained from 10 randomized runs of the same base composition as in the natural sequence. Vertical scale = simplicity; horizontal scale = 200 bp. *Bottom,* Map of the *D. melanogaster* rDNA repeat drawn to the same scale as the simplicity profile above. Positions of genes and spacers are marked. The 95-bp, 330-bp, and 240-bp subrepeat arrays within the IGS are also marked (for details, see Coen et al. 1982; Tautz et al. 1987). The arrow under the 28S rRNA gene indicates the position of the internal RNA processing site yielding the 28Sα and 28Sβ rRNAs. The expansion and core segments of the 28S rRNA gene are indicated as unfilled and filled boxes, respectively. The 18S rRNA also contains expansion segments, but they are not shown here as they are not part of the current analysis.

the regions of high simplicity in the 28S rRNA gene are known to correspond almost exactly to the expansion segments (see Hancock and Dover 1988).

Discussion

Numbering rDNA Sequences

DNA sequences can be numbered in a variety of ways. They may be either simply numbered from the 5′ to the 3′ end or numbered bidirectionally from the start of transcription (see, e.g., Tautz et al. 1987), or individual functional units may be numbered separately. All of these numbering systems suffer from disadvantages. Simply numbering a sequence from the 5′ end may result in the numbering system having no functional relevance. Bidirectional numbering from the start of transcription, although both superficially attractive and adequate for sequences straddling the start of transcription, suffers from a number of disadvantages. It can also produce a numbering system that is functionally irrelevant if applied to sequence regions that do not include the transcription start site, and it cannot be applied if the transcription start site is not known. Using such a numbering system in a tandemly organized multigene family such as rDNA leads to even greater difficulties. As transcription in the rDNA can be initiated at multiple upstream promoters and proceeds through the entire repeat into the IGS, in some cases without termination (Tautz and Dover 1986), and as the "true promoter" cannot be distinguished from the upstream duplications in some species (Cross and Dover 1987), it would seem more useful to number the entire unit without reference to the start of transcription. This would also avoid the problem of numbering the entire IGS negatively over kilobase distances. Numbering individual functional regions separately may be helpful when analyzing these sequence regions in isolation, although it is less helpful in the rDNA, where the lengths of the rRNA genes vary both between (Ware et al. 1983) and within (Maden et al., 1987) species. It can also lead to considerable confusion when one refers to a number of sequence regions at the same time. It may, however, be the method of choice for convenient storage and retrieval of sequence data. In this and the following papers (Hancock and Dover 1988; Hancock et al. 1988), however, we have numbered the sequence region presented here from the 5′ end of the 18S + RNA gene. This has the advantages of simplicity and unambiguity while making no assumptions about the details of rDNA transcription.

Sequence Variation

The complete sequence of the rRNA genes confirms earlier conclusions about the organization of *D. melanogaster* rDNA (reviewed in Gerbi 1985). Our sequence overlaps sufficiently with previous fragmentary sequence information for both *D. melanogaster* and *D. virilis*—in particular with the regions surrounding the types I and II inserts in *Drosophila* 28S rDNA (Rae et al. 1980; Dawid and Rebbert 1981; Mandal and Dawid 1981; Roiha and Glover 1981; Roiha et al. 1981; Delanversin and Jacq 1983)—for significant comparisons to be made. Our sequence shows 18 differences in a total of 2,653 bp from other published sequences for *D. melanogaster.* One region of uncertainty remains between positions 5052 and 5056, which was difficult to read in our gels. The version given here is in accordance with the Dalanversin and Jacq (1983) study and is compatible with our gels—but could be incorrect. If it is assumed that all other positions have been sequenced correctly here—and therefore represent alternative copies of the gene—we detect heterogeneity at 0.68% of positions, although not all positions would be expected to be heterogeneous in all copies of the gene. Consideration of published sequence regions overlapping with our sequence and with one another (Dawid and Rebbert 1981; Roiha and Glover 1981) leads us to conclude that at least four different variant copies of the rRNA genes are represented in the published sequence data. This leads to an estimate of heterogeneity between repeats (i.e., the percentage of variant sites divided by the number of copies) of 0.17%. This figure could be as low as 0.076% if all published sequences derive from different copies. Although it is possible that some of these differences reflect sequencing errors, some variation would be expected given the considerable variation between individual cloned rDNA units detected in human rDNA (Maden et al. 1987), particularly in the 28S rRNA gene. It is interesting that, within the 28S rRNA gene, four differences between our sequence and that of Simeone et al. (1985) reside in double-stranded regions of the rRNA secondary structure (see Hancock et al. 1988). All of these differences represent compensatory mutations, A-U to G-U in three cases and C-G to U-G in one case. In the 18S rRNA the differences between our sequence and previously published sequences (Jordan et al. 1980; Youvan and Hearst 1981) are all located in single-stranded regions.

Yagura et al. (1979) identified two 14-nucleotide-long, 18S rRNA–derived oligonucleotides from the Oregon strain of *D. melanogaster,* one corresponding to an X-linked gene and the other to a Y-linked gene. Our sequence contains only one region that shows >80% sequence similarity to either of these oligonucleotides, the sequence TATTTTTTCATATG (bases 750–763). This sequence is 86% identical to their X-linked sequence (12 matches) and 79% identical to their Y-linked sequence (11 matches) and lies within the region of the 18S rRNA gene encoding expansion segment V4 of the 18S rRNA secondary structure (see Hancock et al. 1988).

Rae et al. (1980) have sequenced 491 bp of the *D. virilis* 28S rRNA gene that surround a type I insertion. Comparison of their sequence with the corresponding region in *D. melanogaster* shows 14 sequence differences, including three deletions in *D. virilis* relative to *D. melanogaster.* This represents 2.85% difference between the two sequences, although this is necessarily a crude estimate because of the small amount of sequence available. Although approximate times of separation of *D. melanogaster* and *D. virilis* have been assessed (Beverley and Wilson 1984), it is not meaningful to convert percent divergences of arbitrary sections of DNA into rates of divergence (see also Dover 1987; Tautz et al. 1987). This is because different regions of the rDNA unit—for example, the core and expansion segments of the 28S gene—differ in both base composition and the degree to which slippage-generated variation and point-mutational variation have accumulated. These points are discussed below—and in detail in the accompanying papers (Hancock and Dover 1988; Hancock et al. 1988).

Base Composition

The rDNA of *D. melanogaster* is significantly more AT rich than that in other eukaryotic species. Table 1 shows that the spacer regions of the *D. melanogaster* rDNA repeat are particularly AT rich. The IGS, ETS, and ITS are significantly AT richer than the rRNA gene regions, while the 28S rRNA sequence is much more AT rich than those of the 18S, 5.8S, or 2S rRNAs. Dot-matrix analysis of the *D. melanogaster* 28S rDNA sequences against themselves and against those of other eukaryotic species (Hancock and Dover 1988; see also Ware et al. 1983; Clark et al. 1984) shows the AT richness of *D. melanogaster* 28S rDNA to derive largely from blocks of AT-rich sequence in positions corresponding to the expansion segments. In an accompanying paper (Hancock et al. 1988) we consider both the effects of such drastic alterations in base composition on the secondary structures of the *D. melanogaster* rRNAs and the coevolution of compensatory mutations.

DNA Simplicity Analysis

We have shown (Tautz et al. 1987) a strong correlation between the generation of novel sequence regions within the rDNA of *Drosophila* species and the level of cryptic simplicity within that sequence as measured by an algorithm that searches for repeats of a motif in its immediate neighborhood (see Results). Figure 2 shows a simplicity profile of the entire *D. melanogaster* rDNA repeat generated using this algorithm. It is apparent that spacer regions in general have higher overall levels of cryptic simplicity than do the rRNA coding regions. Table 1 shows that the IGS and ETS—but not the ITS—show values of relative simplicity that are significantly >1.0 ($P < 0.003$) when taken in isolation, signifying a higher level of cryptic simplicity than would be expected for a random sequence of the same length and base composition. Such elevated levels of cryptic simplicity can be ascribed to the effects of slippage-like mechanisms (Tautz et al. 1986). This suggests that, in contrast to the IGS and ETS, the ITSs as a whole have been refractory to the action of slippage-like events, although localized regions occasionally display high peaks of simplicity. These ITS peaks are due solely to the high AT content of this region (see table 1) and do not indicate elevated levels of simplicity, as becomes clear when the RSFs of these regions are calculated in comparison with a test sequence having the same AT content. For details of unusual features of ITS regions in *Xenopus* species, see Furlong et al. (1983) and Furlong and Maden (1983). Similarly, the 18S and 5.8S rRNA genes have low levels of cryptic simplicity, as revealed both in the graphic displays and in the low RSF values.

In contrast to the 18S and 5.8S genes, the 28S rRNA gene as a whole shows an RSF that is significantly >1.0. The simplicity profile shows that simple regions are nonrandomly distributed within the gene. Detailed analysis of the 28S rRNA genes from a variety of species reveals that regions of high simplicity correspond to expansion segments rather than to the core segments and that the set of expansion segments is coevolving during interspecific divergence (see an accompanying paper [Hancock and

Dover 1988]). This suggests that the 28S rRNA alone of the rRNAs is able to remain functional in the presence of the repetitive and scrambled products of slippage-like events. The role of such mechanisms in the evolution of 26S/28S rRNAs and the extent to which expansion segments coevolve within a species is discussed in more detail in the accompanying papers (Hancock and Dover 1988; Hancock et al. 1988).

In conclusion, DNA simplicity analysis shows that the slippage-like mechanisms, which previously have been shown to contribute to the rapid divergence of the IGS and ETS regions in *Drosophila* species (Tautz et al. 1987), have affected only the 28S rRNA gene among the rRNA gene regions. Additionally, the observed nonhomogeneous distribution of AT richness throughout the rDNA repeat suggests that other processes, perhaps biased occurrence or fixation of point mutations, have been responsible for the AT richness of the *Drosophila* rDNA repeat. These latter processes might be analogous to those supposedly responsible for the establishment of long stretches of AT- or GC-rich regions in the genomes of warm-blooded animals (Bernardi et al. 1985).

Acknowledgments

We thank the Science and Engineering Research Council for research grants GR/D18677 and GR/E53538.

LITERATURE CITED

ARNHEIM, N. 1983. Concerted evolution of multigene families. Pp. 38–61 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL, and F. RODIER. 1985. The mosaic genome of warm-blooded vertebrates. Science **228**:953–958.

BEVERLEY, S. M., and A. C. WILSON. 1984. Molecular evolution in *Drosophila* and the higher Diptera. II. A time scale for fly evolution. J. Mol. Evol. **21**:1–13.

CLARK, C. G., B. W. TAGUE, V. C. WARE, and S. A. GERBI. 1984. *Xenopus laevis* 28S ribosomal RNA: a secondary structure model and its evolutionary and functional implications. Nucleic Acids Res. **12**:6197–6220.

COEN, E. S., T. STRACHAN, and G. A. DOVER. 1982. The dynamics of concerted evolution of rDNA and histone gene families in the *melanogaster* species subgroup of *Drosophila*. J. Mol. Biol. **158**:17–35.

CROSS, N. C. P., and G. A. DOVER. 1987. A novel arrangement of sequence elements surrounding the rDNA promoter and its spacer duplications in tsetse species. J. Mol. Biol. **195**:63–74.

DAWID, I. B., and M. L. REBBERT. 1981. Nucleotide sequences at the boundaries between gene and insertion regions in the rDNA of *Drosophila melanogaster*. Nucleic Acids Res. **9**:5011–5021.

DELANVERSIN, G., and B. JACQ. 1983. Sequence of the central break region of *Drosophila* 26S precursor ribosomal RNA. CR Seances Acad. Sci. [III] **296**:1041–1044.

DENTE, L., G. CESARENI, and R. CORTESE. 1983. pEMBL: a new family of single-stranded plasmids. Nucleic Acids Res. **11**:1645–1655.

DOVER, G. A. 1982. Molecular drive: a cohesive mode of species evolution. Nature **299**:111–117.

———. 1987. DNA turnover and the molecular clock. J. Mol. Evol. **26**:47–58.

DOVER, G. A., and R. B. FLAVELL. 1984. Molecular coevolution: DNA divergence and the maintenance of function. Cell **38**:622–623.

FLAVELL, R. B. 1986. Structure and control of expression of ribosomal RNA genes. Oxf. Surv. Plant Mol. Cell. Biol. **3**:252–274.

FURLONG, J. C., J. FORBES, M. ROBERTSON, and B. E. H. MADEN. 1983. The external transcribed spacer and preceding region of *X. borealis* rDNA. Nucleic Acids Res. **11**:8183–8196.

Furlong, J. C., and B. E. H. Maden. 1983. Patterns of major divergences between the ITS of the rDNA of *Xenopus borealis* and *X. laevis* and of minimal divergence within ribosomal coding regions. EMBO J. **2**:443–448.

Gerbi, S. A. 1985. Evolution of ribosomal DNA. Pp. 419–517 *in* R. J. MacIntyre, ed. Molecular evolutionary genetics. Plenum, New York.

Guo, L.-H., and R. Wu. 1982. New rapid methods for DNA sequencing based on exonuclease III digestion followed by repair synthesis. Nucleic Acids Res. **10**:2065–2084.

Hancock, J. M., and G. A. Dover. 1988. Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. Mol. Biol. Evol. **5**:377–392.

Hancock, J. M., D. Tautz, and G. A. Dover. 1988. Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. Mol. Biol. Evol. **5**:393–414.

Jordan, B. R., R. Jourdan, and B. Jacq. 1976. Late steps in the maturation of *Drosophila* 26S ribosomal RNA: generation of 5.8S and 2S RNAs by cleavages occurring in the cytoplasm. J. Mol. Biol. **101**:85–105.

Jordan, B. R., M. Latil-Damotte, and R. Jourdan. 1980. Sequence of the 3′-terminal portion of *Drosophila melanogaster* 18S rRNA and of the adjoining spacer: comparison with corresponding prokaryotic and eukaryotic sequences. FEBS Lett. **117**:227–231.

Maden, B. E. H., C. L. Dent, T. E. Farrell, J. Garde, F. S. McCallum, and J. A. Wakeman. 1987. Clones of human ribosomal DNA containing the complete 18S-rRNA and 28S-rRNA genes. Biochem. J. **246**:519–527.

Mandal, R. K., and I. B. Dawid. 1981. The nucleotide sequence at the transcription termination site of ribosomal RNA in *Drosophila melanogaster*. Nucleic Acids Res. **9**:1801–1811.

Moss, T., K. Mitchelson, and R. de Winter. 1985. The promotion of ribosomal transcription in eukaryotes. Oxf. Surv. Eukaryotic Genes **2**:207–250.

Ohta, T. and G. A. Dover. 1984. The cohesive population dynamics of molecular drive. Genetics **108**:501–521.

Pavlakis, G. N., B. R. Jordan, R. M. Wurst, and J. N. Vournakis. 1979. Sequence and secondary structure of *Drosophila melanogaster* 5.8S and 2S rRNAs and of the processing site between them. Nucleic Acids Res. **7**:2213–2238.

Rae, P. M. M., B. D. Kohorn, and R. P. Wade. 1980. The 10 kb *Drosophila virilis* 28S rDNA intervening sequence is flanked by a direct repeat of 14 base pairs of coding sequence. Nucleic Acids Res. **8**:3491–3504.

Reeder, R. H. 1984. Enhancers and ribosomal gene spacers. Cell **38**:349–351.

Roiha, H., and D. M. Glover. 1981. Duplicated rDNA sequences of variable lengths flanking the short type 1 insertions in the rDNA of *Drosophila melanogaster*. Nucleic Acids Res. **9**: 5521–5532.

Roiha, H., J. R. Miller, L. C. Woods, and D. M. Glover. 1981. Arrangements and rearrangements of sequences flanking the two types of rDNA insertion in *D. melanogaster*. Nature **290**:749–753.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.

Simeone, A., and E. Boncinelli. 1984. 5′-Cleavage site of *D. melanogaster* 18S rRNA. FEBS Lett. **167**:249–253.

Simeone, A., A. LaVolpe, and A. Boncinelli. 1985. Nucleotide sequence of a complete ribosomal spacer of *D. melanogaster*. Nucleic Acids Res. **13**:1089–1101.

Tautz, D., and G. A. Dover. 1986. Transcription of the tandem array of ribosomal DNA in *Drosophila* does not terminate at any fixed point. EMBO J. **5**:1267–1273.

Tautz, D., C. Tautz, D. Webb, and G. A. Dover. 1987. Evolutionary divergence of promoters and spacers in the rDNA family of four *Drosophila* species: implications for molecular coevolution in multigene families. J. Mol. Biol. **195**:525–542.

Tautz, D., M. Trick, and G. A. Dover. 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature **322**:652–656.

WARE, V. C., B. W. TAGUE, C. G. CLARK, R. L. GOURSE, R. C. BRAND, and S. A. GERBI. 1983. Sequence analysis of 28S ribosomal DNA from the amphibian *Xenopus laevis.* Nucleic Acids Res. **11:**7795–7817.

YAGURA, T., M. YAGURA, and M. MURAMATSU. 1979. *Drosophila melanogaster* has different ribosomal RNA sequences on the X and Y chromosomes. J. Mol. Biol. **133:**533–547.

YOUVAN, D. C., and J. E. HEARST. 1981. A sequence from *Drosophila melanogaster* 18S rRNA bearing the conserved hypermodified nucleoside am-pseudouridine: analysis by reverse transcription and high-performance liquid chromatography. Nucleic Acids Res. **9:**1723–1741.

RICHARD B. FLAVELL, reviewing editor

Received September 6, 1987; revision received January 20, 1988