Sequence and Sequence Variation Within the 1·688 g/cm³
Satellite DNA of Drosophila melanogaster

TAO-SHIH HSIEH AND DOUGLAS BRUTLAG

Department of Biochemistry
Stanford University School of Medicine
Stanford, Calif.
94305, U.S.A.

(Received 29 June 1979)

We have determined the complete nucleotide sequence of the monomer repeating unit of the 1·688 g/cm³ satellite DNA from Drosophila melanogaster. This satellite DNA, which makes up 4% of the Drosophila genome and is located primarily on the sex chromosomes, has a repeat unit 359 base-pairs in length. This complex sequence is unrelated to the other three major satellite DNAs present in this species, each of which contains a very short repeated sequence only 5 to 10 base-pairs long. The repeated sequence is more similar to the complex repeating units found in satellites of mammalian origin in that it contains runs of adenylate and thymidylate residues. We have determined the nature of the sequence variations in this DNA by restriction nuclease cleavage and by direct sequence determination of (1) individual monomer units cloned in hybrid plasmids, (2) mixtures of adjacent monomers from a cloned segment of this satellite DNA, (3) mixtures of monomer units isolated by restriction nuclease cleavage of total 1·688 g/cm³ satellite DNA. Both direct sequence determination and restriction nuclease cleavage indicate that certain positions in the repeat can be highly variable with up to 50% of certain restriction sites having altered recognition sequences. Despite the high degree of variation at certain sites, most positions in the sequence are highly conserved. Sequence analysis of a mixture of 15 adjacent monomer units detected only nine variable positions out of 359 base-pairs. Total satellite DNA showed only four additional positions. While some variability would have been missed due to the sequencing methods used, we conclude that the variation from one repeat to the next is not random and that most of the satellite repeat is conserved. This conservation may reflect functional aspects of the repeated DNA, since we have shown earlier that part of this sequence serves as a binding site for a sequence-specific DNA binding protein isolated from Drosophila embryos (Hsieh & Brutlag, 1979).

1. Introduction

The DNA that comprises most of the constitutive heterochromatin of eukaryotic genomes consists of highly repetitive DNA sequences (Pardue & Gall, 1970; Jones, 1970; Peacock et al., 1973). These tandemly repeated DNAs, called satellite DNAs, are of two major types: those consisting of very short nucleotide sequences between 2 and 10 base-pairs in length, such as found in crabs or in Drosophila virilis (Laskowski, 1972; Gall & Atherton, 1974); and those that have longer and more complex repeat units like the satellite DNAs of mammalian origin (Southern, 1970, 1975; Botchan,
Based on partial sequence data, renaturation kinetics, cross hybridization experiments, and restriction nuclease cleavage, several workers proposed that the more complex repeating units had evolved from shorter repeated sequences by a process of mutation followed by amplification of a longer, more complex sequence (Southern, 1970; Sutton & McCallum, 1971; Biro et al., 1975; Altenburger et al., 1977; Streeck & Zachau, 1978). Recent sequence analyses of several mammalian satellites have not shown evidence of short internal repeats within the repeating unit (Rosenberg et al., 1978; Shmookler Reis & Biro, 1978; Manuelidis & Wu, 1978). However, evidence does show that both a minor DNA component of human satellite I and a major satellite DNA of the bonnet monkey consists of two related internal repeats 171 base-pairs in length which are similar to the 172 base-pair sequence of the major satellite DNA from African Green monkeys (Rosenberg et al., 1978; Manuelidis & Wu, 1978; Rubin et al., 1979). A detailed analysis of the African Green monkey satellite DNA has shown it to be very homogeneous with only a limited number of sequence variations.

We have determined the complete nucleotide sequence of a cloned monomer unit from a complex satellite DNA of Drosophila melanogaster. This satellite DNA, with a buoyant density of 1.688 g/cm³, has a repeat length about 356 base-pairs as determined by restriction enzyme digests (Manteuil et al., 1975; Shen et al., 1976; Carlson & Brutlag, 1977). This property makes it quite distinct from the three other major satellite DNAs present in D. melanogaster, each of which contains a very short repeated sequence between 5 and 10 base-pairs long (Peacock et al., 1973; Endow et al., 1975; Brutlag & Peacock, 1975, 1979; Endow, 1977; Brutlag et al., 1978; Fry & Brutlag, 1979). By cloning individual segments of the 1.688 g/cm³ satellite (1-888 satellite) DNA it was shown that there are some regions with a very regular arrangement of restriction cleavage sites while other regions were completely lacking some sites (Carlson & Brutlag, 1977, 1979). Direct sequence analysis of one cloned satellite region containing 15 repeating units described here shows that there is sequence variation in only a few nucleotide positions within the satellite repeat. We have also determined the degree of sequence variation in repeat units isolated by restriction nuclease digestion of total satellite DNA and again found variation in only a limited number of sites within the monomer. From these results we conclude that the sequence variations observed in this satellite DNA are not randomly distributed throughout the repeat unit and that the bulk of the repeat is very homogeneous in sequence. The homogeneity of the satellite sequence and the specificity observed in the variable positions suggests the sequence may be conserved for functional reasons.

2. Materials and Methods
(a) Bacterial strains
E. coli HB101 (recA, hsr, hsm; Boyer & Roulland-Dussoix, 1969) and E. coli RK1 (leu, pro, Bi, r, m, lac y, strt) harboring the plasmid pBR322 (Bolivar et al., 1977) were gifts from D. S. Hogness and H. Boyer, respectively.

(b) Enzymes
EcoRI restriction endonuclease (Modrich & Zabel, 1976), HinfI endonuclease (Carlson & Brutlag, 1977), polynucleotide kinase (Weiss et al., 1968), and terminal deoxynucleotidyl transferase (Chang & Bolton, 1971) were gifts of P. Modrich, M. Carlson, J. Chien and
VARIATION IN SATELLITE DNA SEQUENCES

R. Ratliff, respectively. Bacterial alkaline phosphatase was purchased from Worthington Biochemicals (BAPF) and the restriction endonucleases Atul, HaeIII, MboI were from New England Biolabs.

(c) Isolation of DNAs

The 1.088 g/cm³ satellite from D. melanogaster was isolated from total embryonic DNA according to Brutlag et al. (1977). Covalently closed circular plasmid DNA was isolated from bacterial lysates according to the procedure of Katz et al. (1973), followed by centrifugation twice in CsCl/ethidium bromide gradients (Radloff et al., 1967). After the removal of ethidium by n-butanol extraction, the DNA was dialyzed into TKE buffer (10 mM-Tris-HCl, pH 7.9, 50 mM-KCl, 0.1 mM-EDTA).

Restriction fragments were recovered from polyacrylamide gels by slicing out the appropriate band and extracting the gel slice 3 times with an equal vol. of TKE buffer. The DNA in combined extracts was concentrated either by precipitation with ethanol or by adsorption and elution from hydroxyapatite.

(d) Construction of hybrid plasmids

Hybrid plasmids were constructed carrying only a single repeating unit (HaeIII monomer) of the 1.088 satellite DNA linked to the vector pBR322 through oligo(dC)·oligo(dG) tails. Single repeat units cloned in this way had two HaeIII sites reconstructed at either end and the monomer units could be readily isolated from the vector DNA after cleavage with HaeIII restriction enzyme. The satellite monomers to be cloned were isolated from the plasmid pDM23, a hybrid plasmid of pSC101 and 6x10⁵ base-pairs of 1.088 satellite (Carlson & Brutlag, 1977). DNA of this plasmid was digested with HaeIII endonuclease, the satellite monomers purified by polyacrylamide electrophoresis, and oligo(dC) approx. 20 residues long added to the 3' ends with terminal transferase according to a modification of the method of Roychoudhury et al. (1976). Oligo(dG) about 10 nucleotides in length, was added to 10 µg EcoRI-cleaved pBR322 DNA in 140 µl of 0.1 M-sodium cacodylate (pH 7.0), 1 mM-CoCl₂, 10 mM-dithiothreitol, 0.1 mM-dGTP with 20 µg of terminal transferase at 37°C for 2-3 min. The reaction was terminated by adding 15 µl of 1% sodium dodecyl sulphate and 0.1 M-EDTA (pH 8.1). The reaction mixture was extracted with an equal vol. of buffer-saturated phenol once and then dialyzed against 0.1 M-NaCl in TE buffer (10 mM-Tris-HCl, pH 7.9, 1 mM-EDTA) for 24 h with 3 buffer changes in order to remove both nucleotides and phenol.

Equal volumes of vector DNA carrying oligo(dG) tails (87-5 µg/ml) and satellite monomer DNA with oligo(dC) tails (3-3 µg/ml) were mixed and incubated at 65°C for 10 min followed by 45°C for 1 h and 37°C for 1 h. Transformation of E. coli HB101 to tetrazolium resistance was performed as previously described (Wensink et al., 1974). The transformation efficiency of the annealed DNAs was 10,000 colonies/µg. Cleavage of hybrid plasmid DNA from several independent transformants with HaeIII showed that most of the plasmids contained a satellite monomer unit about 359 base-pairs in length, which had HaeIII sites reconstructed at both ends. All of the work with hybrid plasmids was carried out under the P2 level of physical containment and the EK1 level of biological containment according to the National Institutes of Health Guidelines on Recombinant DNA Research (1978).

(e) DNA sequence analysis

In order to obtain sufficient resolution of restriction fragments for sequence analysis, restriction digests were separated in a preparative polyacrylamide gel using the Tris-acetate buffer system (Loening, 1967; Maniatis et al., 1975). The desired fragment was isolated from the gel as described in section (c) above. The DNA fragment (0.5 µg in 100 µl of TKE buffer) was treated with alkaline phosphatase at 37°C for 30 min to remove 5' terminal phosphates (Weiss et al., 1968). The DNA was then extracted 3 times with phenol and 3 times with ether, and the ether removed by a jet of anhydrous N₂. Then 200 pmol of [γ-³²P]ATP (1500 Ci/mmol; ICN) was dried in vacuo and resuspended in the solution of dephosphorylated DNA. The solution was then made 10 mM-MgCl₂, 8 mM-dithiothreitol,
and 5% glycerol and the DNA labeled at its 5' ends using 2 units of polynucleotide kinase at 37°C for 30 min. The reaction was terminated by addition of 100 μl of 4 M-ammonium acetate and 1 μl of 24 mg yeast transfer RNA/ml as carrier. The DNA was precipitated with ethanol, the pellet washed once with 80% ethanol, dried in vacuo, and resuspended in 50 μl of TBE buffer with 7 mM-MgCl₂. In order to separate the two 5' termini, the DNA fragments were cleaved with an appropriate restriction enzyme (see Fig. 3) and the resulting fragments separated by polyacrylamide gel electrophoresis. The large AluI fragment can be resolved into 2 halves after Hinfl cleavage despite their similar sizes (92 and 94 base-pairs) if a 30-cm 15% polyacrylamide gel is run at 3 V/cm for 60 h.

The isolated DNA fragments, each labeled at a unique 5' end, were cleaved by base-specific chemical methods (Maxam & Gilbert, 1977) and the products were analyzed on a urea/polyacrylamide gel or on the very thin sequencing gel system (Sanger & Coulson 1979).

(f) Search for internal homologies

The DNA sequence was analyzed for internally repeated regions and for dyad symmetry by use of the computer program described by Korn et al. (1977). This program was translated into SAIL and modified to run in an interactive fashion (Friedland, Kedes & Brutlag, unpublished work). Mismatched nucleotides were allowed in the homology search if they were followed by matches in 2 out of 3 subsequent positions, and loops out of up to 3 nucleotides in length were allowed if they were followed by 3 matched bases. Statistical calculations of expected frequency were made according to the Markov methods suggested by Korn et al. (1977), taking into account the base composition of this A + T-rich sequence and the relaxed conditions used during the homology searching procedure. The homologies found were ranked according to their likelihood and only those homologies of highest statistical significance are reported.

3. Results

(a) Distribution of restriction sites in 1-688 satellite monomers

Previous results had shown that the 1-688 satellite DNA from D. melanogaster has a repeating unit of approximately 365 base-pairs and that there were recognition sites for the restriction endonucleases HaeIII, Hinfl, and AluI in the monomer sequence (Manteuil et al., 1975; Shen et al., 1976; Carlson & Brutlag, 1977). Some of the restriction sites were missing in occasional monomer units resulting in the production of dimers, trimers, etc. upon restriction nuclease cleavage of the satellite DNA. The absence of these restriction sites in this Drosophila satellite DNA is due to an alteration of the recognition sequence (Carlson & Brutlag, 1979), rather than to a base modification as found for satellites from other species (Gautier et al., 1977). To analyze the frequency of alteration of a specific restriction site within monomer units, we have isolated the monomers generated by either HaeIII or Hinfl and then cleaved them with the other restriction enzyme. Figure 1 shows the products of this second cleavage separated by agarose gel electrophoresis. The monomers isolated by Hinfl cleavage of the satellite DNA are denoted by (Hinfl-1-688)359 based on the known length of the fragment (see below and Fig. 3). Cleavage of these Hinfl monomers by HaeIII produced two fragments, one 210 base-pairs and the other 149 base-pairs, as the result of a single HaeIII recognition site (cf. Fig. 3). Those monomers having an altered HaeIII site remain at the position of 359 base-pairs. The densitometric tracing in Figure 1 shows that the material remaining at 359 base-pairs is not symmetric, which is due to contamination by other, more complex DNA sequences in our CCl gradient-purified 1-688 satellite DNA sample (Brutlag et al., 1977). This contaminating DNA was detectable during the original cleavage to produce the Hinfl monomer as a slight
Fig. 1. Restriction nuclease cleavage of satellite monomer units isolated from 1-688 satellite DNA by HinfI or HaeIII endonucleases. Approximately 0-1 μg of satellite DNA monomers, originally isolated by HinfI endonuclease cleavage ((Hinf-1-688)_{359}), were further cleaved by either (a) HaeIII or (b) AluI endonucleases. Monomers originally isolated by HaeIII cleavage of satellite DNA ((HaeIII-1-688)_{359}), were similarly treated with either (c) HinfI or (d) AluI endonucleases. The restriction digest products were analyzed by electrophoresis in a 2% agarose gel at a voltage gradient of 5 V/cm for 3 h. The gel was stained with ethidium bromide, illuminated with short wavelength u.v. light, and the fluorescence of the DNA fragments was photographed through a Wratten no. 22 filter. Densitometer tracings were made directly from the negative. The length of each DNA fragment is given in base-pairs by the number over each peak.

DNA background within the preparative gel. After accounting for these contaminating sequences we estimate that 50% of the (Hinf-1-688)_{359} monomers do not contain HaeIII sites.

The isolated HinfI monomer was also treated with the restriction enzyme AluI which has two recognition sites in the repeat unit generating three fragments, 173, 94, and 92 base-pairs in length (Fig. 1(b) and Fig. 3). An alteration of one of the AluI sites would have resulted in the production of either a 265 or a 267 base-pair fragment. In contrast to the HaeIII digestion, very little material was observed at these positions when the (Hinf-1-688)_{359} monomer was digested with AluI. This indicates that the frequency of alteration of the AluI sites is much lower than that of the HaeIII site in these HinfI monomers. We estimate that only 10% of all (Hinf-1-688)_{359} monomers have one or the other AluI site altered.

We have also isolated monomers from digestion of satellite DNA with HaeIII restriction nuclease ((HaeIII-1-688)_{359}). When the HaeIII monomers were treated with HinfI most of the DNA was cleaved into 210 and 149 base-pair fragments (Fig. 1(c)). A comparison of the amount of uncleaved material present in HaeIII digests of HinfI monomers or of HinfI digests of HaeIII monomers clearly shows that the HinfI sites
are altered much less frequently than the \textit{HaeIII} sites. This conclusion is also true for \textit{HaeIII} and \textit{HinfI} sites in total satellite DNA as well as these isolated monomer units. Digestion of the total satellite with either enzyme shows that the oligomeric pattern produced by \textit{HaeIII} is much more extensive than that produced by \textit{HinfI} in which most of the satellite (>50\%) can be recovered as monomer units (Carlson & Brutlag, 1977). When the \textit{(HaeIII-1-688)}_{359} monomer is cleaved with \textit{AluI}, three fragment 186, 117, and 56 base-pairs in size are generated (Fig. 1(d)). Sequence alteration of one of the \textit{AluI} sites would be expected to yield DNA fragments either 303 or 24 base-pairs in length (Fig. 3). Figure 1(d) shows a prominent peak at 303 base-pair and very little material 242 base-pairs in size indicating that the \textit{AluI} site 117 base pairs from the \textit{HaeIII} site is altered much more frequently than the one 303 base pairs from the \textit{HaeIII} site (cf. Fig. 3).

(b) \textit{Nucleotide sequence of a cloned satellite monomer}

Due to the presence of the sequence variation described above and to the inherent difficulty in analyzing mixtures of closely related sequences, we felt that it was essential to clone single repeating units for nucleotide sequence analysis. We chose to subclone a single monomer from the region of 1-688 satellite originally cloned in \textit{ti} plasmid \textit{pDm23} because this region appeared to be very homogeneous by restrictive analysis. Cleavage of this plasmid with \textit{HaeIII}, \textit{HinfI}, or \textit{AluI} indicated that the restriction sites were present in each of the 15 monomer units of this region (Carlson & Brutlag, 1977). The satellite monomer units were isolated from \textit{pDm23} by cleavage with \textit{HaeIII} and cloned in the plasmid \textit{pBR322} using oligo(dG):oligo(dC) tails described in Materials and Methods. Monomer units cloned in this way could be readily excised from the vector \textit{pBR322} by cleavage at the two \textit{HaeIII} sites which had been reconstructed at either end of the monomer unit due to the G:C tails. Monomer DN from one of these subclones, \textit{aDm23-24}, was isolated and its sequence determined by the procedure of Maxam & Gilbert (1977) (Fig. 2). Nucleotide sequences were determined from the 5' ends of the \textit{HaeIII}, \textit{HinfI}, and \textit{AluI} sites within the monomer unit as shown in Figure 3. This provided sufficient overlap to unambiguously determine the precise length to be 359 base-pairs. In addition to the known restriction sites the monomer unit, the sequence indicates the presence of several other restriction sites: a site for \textit{BbrI} at 119; for \textit{MboI} at 54 and 97; \textit{MulI} at 29; \textit{SatI} at 303; and few \textit{EcoRI*} sites. The base composition of this monomer is 30-9\% G + C, which significantly less than that predicted from the melting temperature for the satellite DNA (40\% with a \textit{t}_{m} value of 68-5°C in 0-1 \times \text{SSC}; Brutlag \textit{et al.}, 1977). This increased thermal stability of the satellite over that predicted from its base composition is probably due to the unusual arrangements of nucleotides in this sequence. The frequency of the dinucleotides ApA and TpT are significantly higher than expected on a random basis. One also notices that a run of adenylates is often followed by a \textit{r} of thymidylates in the sequence.

(c) \textit{Sequence variation among adjacent monomers}

In order to determine the degree of sequence variation within a limited region of the satellite DNA, we isolated a mixture of all 15 monomers from \textit{pDm23} by cleavage with \textit{HinfI} restriction nuclease. Figure 5 shows the sequence analysis of this mixture of monomers around the \textit{HinfI} site. The results indicate that most of the sequence is identical with the single subcloned monomer unit in \textit{aDm23-24}. However, we
Fig. 2. Autoradiogram of a sequencing gel for the determination of the sequence near the *Hinfl* site of the monomer cloned in aDM23-24. Base-specific cleavage conditions are indicated at the top of each track according to Maxam & Gilbert (1977). The position of the nucleotides relative to the 5' end of the *Hinfl* site (as numbered in Fig. 3) is indicated on either side of the autoradiogram. The 4 tracks on the right show the sequence beginning near the *Hinfl* site and proceeding to the right along the top strand of Fig. 3, while the 4 left tracks represent the sequence proceeding to the left from the *Hinfl* site along the lower strand of Fig. 3.
## Table 1

<table>
<thead>
<tr>
<th>Nucleotide position</th>
<th>Nucleotide sequence from pDm23</th>
<th>Nucleotide sequence from aDm23-24</th>
<th>1-688-HaeIII-sensitive</th>
<th>1-688-HaeIII-resistant</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>T</td>
<td>T or G</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>29</td>
<td>C</td>
<td>C or G</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>131</td>
<td>A</td>
<td>A or T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>137</td>
<td>G</td>
<td>G or T</td>
<td>G or T</td>
<td>G or T</td>
</tr>
<tr>
<td>161</td>
<td>C</td>
<td>C or G</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>170</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>195</td>
<td>C</td>
<td>C or A</td>
<td>C or A</td>
<td>A</td>
</tr>
<tr>
<td>198</td>
<td>T</td>
<td>T or A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>199</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>238</td>
<td>T</td>
<td>T or A</td>
<td>T or A</td>
<td>T</td>
</tr>
<tr>
<td>263</td>
<td>A</td>
<td>A or C</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>276</td>
<td>T</td>
<td>T</td>
<td>A or T</td>
<td>A or T</td>
</tr>
<tr>
<td>294</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

This Table presents a summary of the sequence variations detected among the monomers isolated from either the cloned satellite segment in pDm23, or from the HaeIII-sensitive or resistant monomers isolated from bulk satellite DNA, compared with the nucleotide sequence present in the single cloned monomer present in aDm23-24. The nucleotide position is numbered beginning at the 5' terminal C of the HaeIII-cleaved monomer as shown in Fig. 3. N.D. denotes not determined.

able to detect nine positions in which the sequence was ambiguous (Table 1 and Fig. 4). Each of these positions showed a mixture of two nucleotides one of which was the same as the nucleotide appearing in the homologous position in the aDm23-24 sequence. Most of the variations detected, appear to have resulted from transversions and are seen as mixtures of purines and pyrimidines on the sequencing gel. This is due to the Maxam–Gilbert sequencing procedure, which is unable to detect mixtures of pyrimidines (which would only be considered as a cytidylate residue) or mixtures of purines (which would only be considered as an adenylate residue). This limitation means that we would be unable to detect nucleotide mixtures resulting from transitions. The only transition we have detected occurred at position 294 in which a thymidylate is found in the sequence of aDm23-24 whereas there appears to be a cytidylate residue in the sequence of the mixture of monomers from pDm23. The finding of a thymidylate at position 294 of the aDm23-24 monomer suggests that this site actually contains a mixture of both C and T in the monomers from pDm23. This conclusion is supported by digestion of HaeIII monomer units from pDm23 by the restriction enzyme MnlI. Those monomers containing a C at position 294 would contain an MnlI site (5'-C-C-T-C-3') whereas monomers with a T at this position would not. A similar situation exists at position 29 where both C and G residues have been detected. Those monomers containing a C at position 29 would be expected to have an MnlI recognition site there as well. Electrophoresis of the products of MnlI cleavage of HaeIII
Fig. 4. The nucleotide sequence and the location of sequence variations within the 16 cloned monomers present in pDm23. The arrows indicate restriction nucleoscleavage sites in the upper strand of the sequence. The sequence variations were detected by the presence of a mixture of nucleotides on the sequencing gel at a single site and are indicated here by the presence of two base-pairs at a single position (see Fig. 5). The central sequence was arbitrarily chosen to be that of pDm23-24 and the variations were placed on the outside. No attempt to quantitate the relative frequencies of either base-pair at ambiguous sites is possible because of the sequencing procedure employed. Hyphens have been omitted for clarity.
monomers isolated from pDm23 resolved three large fragments: intact 359 base-pair monomers, and two cleavage products about 300 and 330 base-pairs long (data not shown). The fragment 300 base-pairs in length would be expected from those monomers containing an MnII cleavage site at position 294 while the fragment 330 base-pairs long could have resulted from cleavage at position 29. The absence of a fragment 270 base-pairs long suggested that very few monomer units contain C residues (and hence MnII sites) at both positions 29 and 294. The presence of monomers uncleaved by MnII suggests the presence of monomers lacking MnII sites at both position 29 and 294. This suggestion was confirmed by analysis of a second subcloned monomer from pDm23 called aDm23-40. Unlike the monomer cloned in aDm23-24, the monomer in aDm23-40 contains a G in position 29 and there is no MnII recognition sequence in this monomer unit.

(d) Sequence of monomers isolated from satellite DNA

Since the nucleotide sequence analysis of the 15 adjacent monomer units present in pDm23 detected nine distinct variable sites we decided to investigate the degree of sequence variation present in the bulk of the satellite DNA. We originally expected the sequence variation in the total satellite to be randomly distributed, and hence not detectable by our sequencing procedure since variations are present only at a low level (7%; Brutlag et al., 1977). Thus we expected to see only the predominant nucleotide at each position. We isolated Hinfl monomers from the total satellite DNA ((Hinf-l-688)$_{359}$), since the Hinfl restriction site is less frequently altered than the HaeIII site (Fig. 1(a) and (c)) and therefore monomers generated by Hinfl cleavage would be more representative of the whole satellite. The (Hinf-l-688)$_{359}$ monomers were labeled at their 5' ends and then cleaved with the restriction enzyme HaeIII. About half of these monomers were cleaved and they are referred to as the HaeIII-sensitive monomers. The two fragments produced by HaeIII cleavage can be purified by preparative gel electrophoresis. The HaeIII-resistant monomers were also isolated and it was found that most of them were sensitive to cleavage by the enzyme MboI. This allowed us to determine the sequence of both the monomers containing and lacking HaeIII sites independently. We were able to obtain sequence information from both fractions for 150 nucleotides on either side of the Hinfl site. We were surprised to find that the sequence of both the HaeIII-sensitive monomers and the HaeIII-resistant monomers were both very similar to that from pDm23. Moreover, sequence variations were detectable in these bulk monomers, and again they only occurred at a limited number of sites (Table 1). Indeed, three of the variable positions are the same as those found within pDm23 (nucleotide positions 137, 193, and 238). Only four new variable sites were detected from the analysis of the bulk satellite monomers.

While the sequence of the HaeIII-sensitive monomers was very similar to that of the HaeIII-resistant fraction, there were certain sites where the nucleotide sequence was clearly different between these two fractions (positions 161, 170, 195, and 238). This indicates that the alteration of the HaeIII recognition site is closely coupled with sequence variation in other parts of the repeating unit. The alteration of the HaeIII recognition sites in the I-688 satellite DNA has previously been shown to be the result of sequence variation rather than base modification (Carlson & Brutlag, 1979). Modification of cytidines in the satellite monomers would have protected the base from reacting with hydrazine during the base-specific cleavage reaction used in the sequence
Fig. 5. Autoradiograms of sequencing gels for mixtures of monomers isolated by either Hinfl cleavage of total satellite DNA ((Hinfl-1-688)_{359}) or a cloned segment of the satellite from pDm23 ((Hinfl-pDm23)_{359}). Base-specific cleavage conditions are indicated at the top of each track and position within the nucleotide sequence according to Fig. 3 is given along the sides of the autoradiograms. The right 4 tracks in the analysis of each mixture represents the sequence to the right of the Hinfl site along the top strand of Fig. 3, while the left 4 tracks of each analysis gives the sequence to the left of the Hinfl site along the lower strand shown in Fig. 3. Ambiguous sequences should be noted at position 195 i.e. both the (Hinfl-1-688)_{359} sequence and in the (Hinfl-pDm23)_{359} sequence.

analysis (Maxam & Gilbert, 1977; Miller et al., 1978). Our sequencing experiments therefore indicate that most of the cytidines are not methylated, including the CpG dinucleotides which are often modified in other satellite DNAs. Preliminary sequencing results also indicate that the lack of HaeIII recognition sites in the bulk of the HaeIII-resistant monomers is due to sequence variation.
(e) Presence of internally repeated subunits

We have examined the monomer sequence for evidence of internal repeating subunits. It was immediately evident that there was no relationship between the 1-688 sequence and any of the repeat units of the simple sequence satellites of *D. melanogaster*. These other satellites have at most two adenylate residues in a row, whereas the 1-688 repeat has long runs of adenylate often followed by long runs of thymidylate. The most significant internal homologies present within the aDm23-24 repeating sequence are shown in Table 2. Four of these internal repeats have 11 out of 12 bases matched. A similar pattern of internal homologies is also found in the variant sequence.

### Table 2

*Repeated sequences within the aDm23-24 monomer unit*

<table>
<thead>
<tr>
<th>Sequence start position</th>
<th>Repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>243</td>
<td>A-T-T-T-A-T-T-T-G-C-C-A-C-A</td>
</tr>
<tr>
<td>331</td>
<td>A-T-T</td>
</tr>
</tbody>
</table>

These are the 6 most highly significant internal homologies found within the monomer sequence obtained from aDm23-24 (Fig. 3).

of the 1-688 satellite which lacks HaeIII restriction sites (Carlson & Brutlag, 1979). While these repeating sequences are suggestive of a simpler evolutionary precursor, these homologies are barely significant statistically. Given the A + T-rich base composition of the 1-688 and the relaxed homology conditions used during the sequence comparison (see Materials and Methods) one would expect between one and ten homologies of this degree of matching or better to be present in a random sequence 359 base-pairs long. The sequence A-A-A-T-T-T- is repeated five times which is more often than expected on a random basis; however, there is no clear pattern to the repeats (Brutlag *et al.*, 1978). If these internal homologies did in fact arise from a more primitive, shorter repeating sequence, the current 1-688 sequence must be highly diverged from it. Therefore these sequences, while suggestive, do not provide strong evidence for the evolution of the 1-688 sequence from simpler repeating units.
4. Discussion

(a) Sequence variation occurs primarily at a few sites

Earlier studies with restriction nucleases have shown that the complex satellite DNA from D. melanogaster, like complex satellites of mammalian origin, have numerous sequence variations (Manteuil et al., 1975; Shen et al., 1976; Carlson & Brutlag, 1977). In the accompanying paper, Carlson & Brutlag have shown by molecular cloning that sequence alteration, rather than base modification, is responsible for the occasional missing restriction sites in the 1-688 satellite. We have shown here that methylation of cytosine is probably very rare in this satellite by direct nucleotide sequence analysis. Furthermore, the MnlI site at position 29 in the monomer unit cloned in aDm23-24 is missing in other monomer units due to a C→G transversion in the MnlI recognition site. We have therefore been able to use restriction nucleases to probe for specific sequence variations in the 1-688 satellite DNA. These experiments showed that the frequency of sequence variation could be very high at certain sites. For example, among all the monomers produced by Hinfl cleavage of satellite DNA, about 50% were missing the HaeIII site. Furthermore, the sequence variations detected by direct sequence analysis of monomers isolated from either the satellite DNA or from the 15 cloned monomers in pDm23, must represent variation in excess of 30% at each position. Variants resulting in intensities of less than 30% of normal cleavage cannot be identified with confidence, despite the low background present in the autoradiograms of the sequencing gels. In addition, the Maxam–Gilbert sequencing procedure is not capable of detecting variations resulting from transition substitutions further suggesting that we are detecting only a fraction of the total variation present in the satellite DNA. Nevertheless, the 13 variable positions out of 359 which we have found can explain over half of the total amount of sequence variation detected by the thermal stability of renatured satellite DNA (3-6% out of 7%; Brutlag et al., 1977). An equal number of variable positions of the transition variety have probably gone undetected in our analysis, since a comparison of this repeat sequence with a closely related one shows a nearly equal number of transition and transversion changes have occurred since these two sequences diverged (12 transitions and 15 transversions; Carlson & Brutlag, 1979). Thus our analysis indicates that the sequence variations within the 1-688 satellite DNA are due to a high degree of variation at a limited number of sites, rather than a low degree of variation distributed randomly throughout the sequence.

We were surprised to find that some of same sequence variations present among the 15 monomer units of pDm23 were also present as major variable sites in the HaeIII monomer units isolated from total satellite DNA. The fact that there are very few sites of variation among the 359 base pairs of this sequence again indicates that these variations are unlikely to have a purely random origin. They are probably limited to specific sites for functional reasons.

Other than the limited number of variable sites, the bulk of the monomer sequence is well conserved within the satellite DNA. Smith (1976) suggested that homogeneity in tandemly repeated regions can be maintained by unequal sister chromatid exchange or by unequal recombination within the repeated region. If the exchange frequency is greater than the mutation frequency then a large degree of homogeneity can be obtained. The homogeneity of the 1-688 satellite sequence cannot be simply explained by this mechanism, since the bulk of the satellite DNA resides on two different chromo-
VARIATION IN SATELLITE DNA SEQUENCES

Most of the 1-688 satellite is distributed equally between the X and Y chromosomes (Peacock et al., 1977), two chromosomes which rarely undergo exchange. Furthermore, these two chromosomes are together only in male *Drosophila*, which lack meiotic recombination. Thus one would expect the 1-688 satellite regions on these two chromosomes to diverge unless they were under an active selective pressure.

(b) Possible functions of the 1-688 satellite DNA

Several possible functions for these repeated regions are immediately suggested by their chromosomal location. Their presence in the X heterochromatin and on the short arm of the Y chromosome suggests that these chromosomal regions may be involved in pairing between the X and Y chromosomes during meiosis. Both genetic (Peacock & Miklos, 1973) and cytogenetic (Cooper, 1964) evidence implicates these chromosomal regions as important for proper chromosome segregation and spermatogenesis in the male germ line. Deletion mutations which remove heterochromatin from the Y, especially those near the nucleolus organizer, reduce meiotic pairing between the X and the Y chromosome resulting in abnormal segregation of the sex chromosomes and in abnormal spermatogenesis which leads to meiotic drive. On the other hand, these same deletions can have beneficial effects in certain other situations. Deletions of X heterochromatin are known to phenotypically revert the mutations known as *abo* (abnormal oocyte) and *dau* (daughterness) and other closely linked maternal mutants that affect early embryonic development (Sandler, 1977). Based on these results, Sandler has suggested that these autosomal loci produce gene products that interact with sex heterochromatin and that a proper interaction is necessary for normal development. Were the 1-688 satellite DNA to play a role in such an interaction, a clear selective pressure could be imposed on the sequence. In this regard, it is important to note that we have recently purified a protein that preferentially binds to 1-688 satellite DNA and binds in particular to the region of the satellite repeat near the *HaeIII* site (Hsieh & Brutlag, 1979). The presence of a protein-binding site within the 1-688 sequence suggests direct molecular mechanisms for sequence conservation.

(c) Evolutionary implications of the monomer sequence

Several workers have suggested, based on partial sequences or hybridization analysis, that complex mammalian satellites have evolved from simpler tandemly repeated sequences by a process of mutation followed by selective amplification of longer, more complex sequences (Southern, 1970; Biro et al., 1975; Sutton & McCallum, 1971). The sequence of complex satellites from both humans and other primates have not shown evidence of a simple sequence ancestor (Rosenberg et al., 1978; Manuelidis & Wu, 1978). Our analysis of the 1-688 satellite of *D. melanogaster* has indicated that this sequence did not arise from one of the simple repeating satellites prevalent in that species nor from any related simple sequence. The simple satellites have short repeat units of the composition \((A-A-X)_m(A-Y)_n\) where X and Y represent non-adenylate residues (Endow et al., 1975; Brutlag et al., 1978; Brutlag & Peacock, 1979; Fry & Brutlag, 1979). In contrast, the 1-688 repeat unit contains long runs of adenylate and thymydilate as well as one run of six G-C base pairs. While there are several internally repeated sequences within both the 359 base-pair repeat described here, there is no obvious internal pattern of repeats (Brutlag et al., 1978).

On the other hand, a comparison of the 359 base-pair repeat presented here with
the closely related 254 base-pair sequence, which is present in separate tandemly repeated regions, strongly suggests that two sequences with a common ancestor can be amplified independently to form new tandemly repeated regions (Carlson & Brutlag, 1979). These two sequences differ principally in a large 100 base-pair deletion between positions 91 to 190 in the sequence of aDm23-24 as well as other smaller insertions, deletions and single base substitutions. This large deletion (or insertion) indicates that tandemly repeated sequences can vary by many processes other than simple base substitutions. Such larger changes as well as insertion and deletion of single bases must be much rarer than base substitutions. Were they as common as the base substitutions that we have observed here, then we would have been unable to determine a unique nucleotide sequence from the mixture of monomers isolated from the total satellite DNA.

5. Summary

In summary, we conclude that the 1.688 satellite DNA of D. melanogaster contains a complex repeating unit 359 base-pairs in length, which is unrelated to any of the simpler repeating units of the other major satellites of this species. The sequence variations detected within the satellite DNA and within small cloned regions are primarily single base substitutions. These variations occur in only a limited number of sites within the repeat unit, while the rest of the satellite sequence appears to be very highly conserved. The monomer sequence shows no strong pattern of internal repetition, suggesting that either it did not evolve from a simpler repeating sequence or that it has diverged so much as to obscure any simpler precursor. By comparison with closely related sequences from the same species it does appear that this satellite evolved by either a large deletion or insertion in an ancestral sequence and that two forms of this more primitive complex sequence were then amplified into separate tandemly repeated regions.

We would like to acknowledge the advice and help provided by Marian Carlson during the course of this work. One of us (T. H.) is a postdoctoral fellow of the Tumor Biology Training program. This work was supported by grants from the National Institute for General Medical Sciences (GM21498) and from the National Foundation March of Dimes (1-605). This manuscript was composed with the aid of the text-editors and formatting facilities of the SUMEX-AIM computer resource which is supported by the National Institutes of Health.

REFERENCES

VARIATION IN SATELLITE DNA SEQUENCES

Cell, 13, 717–725.
University Press, New York.