CONSENSUS SEQUENCE OF MOUSE SATELLITE DNA INDICATES IT IS DERIVED FROM TANDEM 116 BASEPAIR REPEATS

L. MANUELIDIS
Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA

Received 30 March 1981; revised version received 4 May 1981

1. Introduction

Mouse satellite DNA was one of the first satellites observed in eukaryotic cells by isopycnic centrifugation [1] yet its sequence has not been fully characterized. Initial studies using restriction enzymes [2] indicated segments of a repeating ~240 basepair sequence were organized into larger arrays. Fingerprinting studies of purified mouse satellite were considered to contain a simpler consensus sequence of ~9 basepair [3] yet the sequence was more complex than the very simple satellite tandem arrays initially studied in several other organisms such as *Drosophila*. Some centromeric satellite DNAs, such as human 340 basepair DNA repeat have a higher complexity than originally suspected; the sequence of mouse satellite was compared and was found to represent a repeated sequence of intermediate complexity consisting of 4 highly related ~59 basepair units which had no discernible homology with the human repeats by statistical analysis [4]. The unpublished mouse sequence is reported here in full and is based on both 3′- and 5′-end labelling for unambiguous identification of methyl C-residues.

Subsets of other major or common repeated DNAs in the mouse genome that are not homologous to mouse satellite have been identified [5] and repeated DNAs within and around gene coding regions are now commonly observed. The relation of these various subsets of repeated DNAs to satellite DNAs that occupy a much larger proportion of eukaryotic genomes is unknown. However, particular sequence features are likely to be important in understanding the evolution and chromosomal domains of repeated DNA subsets. Statistical analysis of mouse satellite DNA indicates a multistep evolution where the largest tandem arrays form the most significant units. The detailed base sequence of each minimal or prototype unit appears to be a less important feature that is likely to arise more slowly and possibly by a different evolutionary process.

2. Materials and methods

Mouse satellite DNA was purified from isolated nuclei in Hoescht-CsCl gradients as in [6]. Preparations were generally ≥90% pure and further purification prior to sequencing was achieved by digestion of mouse satellite into discrete bands with *Eco*RII (*Bst*NI) or *Ava*II as in [5]. Both of these enzymes cleave the majority of the satellite into a ~240 basepair and 480 basepair fragment, although high multimers may be seen in various preparations. 5′-^{32}P-Labelling on *Ava*II and *Eco*RII fragments was done as in [4]; 3′-^{32}P-labelling was done using a ^{32}P-labelled G and Klenow fragment of DNA polymerase. Strands were separated on 30 cm 5% acrylamide gels after denaturation in alkalai or DMSO (fig.1) [7]; the denatured strands run behind the xylene cyanol dye, which is run to the end of the gel for optimal resolution. Strands were electroeluted from the gels and DNA was sequenced as in [4,7].

Results and discussion

In studies on uncloned purified mouse satellite DNA cleaved with *Ava*II or *Eco*RII, a series of secondary restriction enzyme digests was tried prior to sequencing. Available restriction enzymes such as *Mbo*I, *Hae*III, *Hind*III, *Hin*f, *Alu*I and *Eco*RI did not

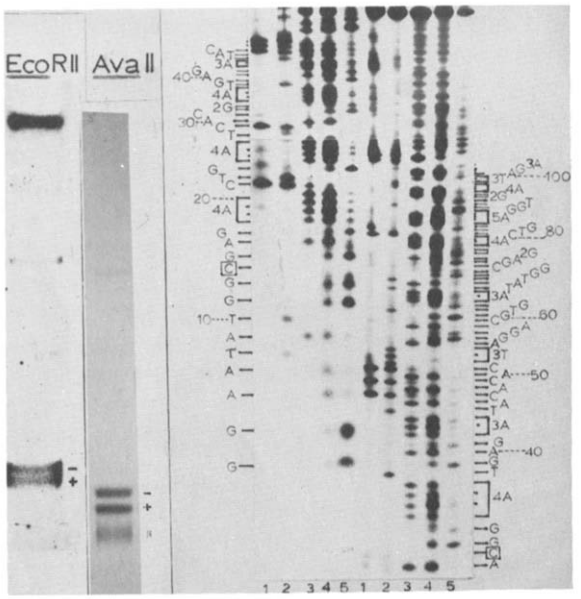

Fig.1. Strand separation prior to sequencing of 5′-labelled mouse satellite cleaved with *Eco*RII and denatured with alkalai, and of 3′-labelled satellite cleaved with *Ava*II and denatured by heating in 33% DMSO. Plus and minus strands, and native undenatured material (N) are noted. Sequence of plus strand from 5′-end cleaved with *Eco*RII shows 2 methylated C residues ([C]). Gel was 80 cm long. Lanes 1–5 are C, C + T, A > C, A > G and G > A, respectively.

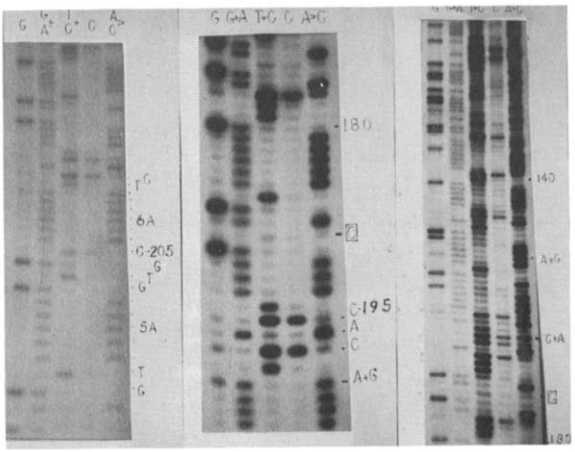

Fig.2. 3′-Labelled plus strand sequenced after cleavage with *Ava*II. Methylated C residues are shown ([C]). Most of the sequence is unambiguous but at several positions minor variants are obtained (e.g., positions 199, 166 and 154). Methyl C residues were confirmed by observation of G in the complementary strand. Note that all methyl C residues are at pCG, however one pCG residue (base 205) is not clearly methylated.

cut the 240 or 480 basepairs fragments to any visible degree and *Ava*II cut satellite could not be further cleaved by *Eco*RII (not shown). Sequencing also confirmed these common restriction sites did not exist in the consensus sequence, and the *Ava*II and *Eco*RII sites overlapped (see fig.3). Although several short 4–6 basepair palindromes were observed in the sequence, restriction enzymes for these palindromes do not yet exist. An *MnlI* site was seen at position 72–75 but it contains a highly methylated C in the minus strand.

Because the sequence is relatively short, methylated C residues in each strand could be confirmed by the sequence obtained in the complementary strand from both 3′- and 5′- end-labelling. Methylated C residues were unambiguous both in the 5′-sequencing studies (fig.1) and in the 3′-sequencing runs (fig.2). For example, even in overexposed autoradiographs (fig.2, center), the C (and C + T) lane at position 189 was extremely faint as compared to all non-methylated C bands in the same gel, as seen at positions 197, 195 and 177. Furthermore the complementary strand registered a G at position 189. Significant methylation of C residues (80–90%) was seen in all pCG sequences with one exception. At position 205 in the plus strand the pCG residue was not clearly methylated (fig.2), and its pCG partner at position 206 in the minus strand showed only partial methylation. From direct sequencing, 7 highly methylated C residues have been identified in each satellite strand, yielding 7 of 234 bases, or 3% total C methylation. Since ~80–90% of each of these 7 C residues was unambiguously methylated a final estimate of 2.4–2.7% mouse satellite was judged to be composed of methyl C residues. It has been noted that >90% of the methyl C residues occur in pCG sequences, and furthermore that mouse satellite is 2.6% methylated as compared to 1% for main band mouse DNA [8]. These methodologically independent results on mouse satellite are thus in good agreement with this number. Here DNA was extracted from the liver, and other tissues could have different levels of satellite methylation.

Most of the bases obtained in this uncloned DNA were unequivocal although at a few positions minor variants were repeatedly obtained in several different satellite preparations and sequencing runs (fig.2).

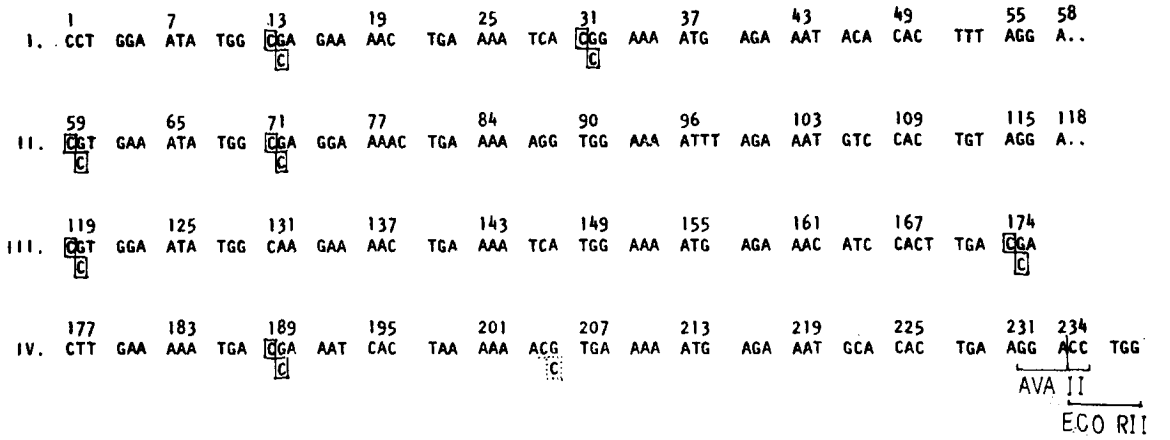

Fig.3. Sequence of mouse satellite from the 5′-end of the plus strand (base 1). Four homologous sequences are seen (I–IV) comprising the 234 basepair repeat. Unambigous methylated C residues are boxed with a continuous line, and methyl Cs observed in the minus strand are similarly noted at their positions on the line beneath.

The consensus sequence in fig.3 indicates the more common base sequence.

Inspection of the sequence showed many GAA... GAAAAA sequences, which is consistent with the original partial prototype sequence in [3]. Additionally certain 'endings' such as CACT were obtained in 4 positions. By aligning GAA... residues and 'endings' it was found that the mouse satellite sequence contained 4 basic repeating units of 58–60 basepairs superimposed on the highly repeated simpler prototype sequence (fig.3 (I–IV)). Furthermore, the base mismatches or insertions between strands I and III was ≤60% less (8 mismatches or 13.8% variants) than between all other strands (e.g., I and II, II and III, I and IV, and III and IV contained 13–18 mismatches or 22–31% variants). Computation of proba bilities [4] indicated the I–III homology had 10^5–10^{10} less random probability of occurring than the other strand matches. From statistical considerations [4] it is obvious that the longer the sequence, the more significant the homology. Indeed, there are 22 base variants between the aligned I–II and III–IV 116 base strands which is $<10^{14}$ less likely to occur than the best matched 58 base subunit ($P = 1.37 \times 10^{-36}$ *vs* $P = 1.6 \times 10^{-22}$). Thus we consider it is likely that the evolution of the mouse 234 basepair satellite DNA is based on an 'amplification' of a 2 unit 116–118 basepair segment rather than a 4 unit 58–60 basepair piece. The basic repeat of 116–118 basepairs is also in keeping with the observation of 0.5–mers in restriction digests [2]. Furthermore, since the consensus sequence is so unequivocal in most positions, 'amplification' into longer tandem arrays of 234 basepair multimers is even more convincing. This combination of 2 or more basic tandem units of >200 basepairs, prior to amplification into enormous tandem arrays with general base fidelity, is similar to that obtained in the human complex repeats which are otherwise unrelated to mouse satellite DNA in detailed base sequence [4]. A basic tandem length of >200 basepairs may be a prerequisite for later amplification into the enormous arrays. Although the mouse satellite sequence may be based on very simple ~9 basepair repeats, possibly generated initially by unequal crossing over [9], and even the 4 more complex 58–60 basepair arrays may be slowly evolving by a similar mechanism, we suggest that the evolution of longer tandem arrays is the most biologically significant step. Indeed homologous simpler sequences of short length are likely to be found in widely divergent organisms albeit in reduced amounts; hybridization experiments in this laboratory indicate these homologies do occur in widely divergent species (unpublished). Such simpler sequence 'libraries' may be continuously created de novo rather than conserved in an evolutionary sense. It has been argued that sequence 'conservation' is less compelling in satellite DNAs than in sequences coding for major cellular proteins [10].

We have stressed that the overall length and tandem arrangement are the most important features of satellite (centromeric) sequences. No convincing

long dyads or other sequence homologies are seen between mouse satellite and human long tandem arrays, and even the methylation is not an obvious common feature. For example all potential pCG residues in the human sequence [4] compose <50% the fraction actually obtained in mouse satellite. In this context it is of interest that a satellite not ordinarily transcribed, in lampbrush chromosomes is transcribed [11], suggesting that long range chromosomal arrangement or three-dimensional structure may be a critical feature in the transcription or 'function' of these sequences.

We have shown that different long repeated DNA sequences in the human genome occupy distinct or different centromeric domains [12]. Mouse satellite is unusual in that all autosomes contain this sequence at their centromeres [13]. However, variants of the long tandem satellites have been noted to occupy distinct mouse chromosomes [14]. The question of how these sequences become so widespread, yet are different among centromeres, or groups of centromeres needs to be resolved. Similarly, the comparison of features of longer intercalary repeated DNA sequence blocks, will help define the evolution and constraints that operate at centromeric domains.

Acknowledgement

This work was supported by NIH grant CA15044

References

[1] Kit, S. (1961) J. Mol. Biol. 3, 711–716.
[2] Southern, E. M. (1975) J. Mol. Biol. 94, 51–69.
[3] Biro, P. A., Carr-Brown, A., Southern, E. M. and Walker, P. M. B. (1975) J. Mol. Biol. 94, 71–86.
[4] Wu, J. C. and Manuelidis, L. (1980) J. Mol. Biol. 142, 363–386.
[5] Manuelidis, L. (1980) Nucleic Acids Res. 8, 3247–3258.
[6] Manuelidis, L. (1977) Anal. Biochem. 78, 561–568.
[7] Maxam, A. and Gilbert, W. (1979) Methods Enzymol. 65, 504–561.
[8] Gruenbaum, Y., Stein, R., Cedar, H. and Razin, A. (1981) FEBS Lett. 124, 67–71.
[9] Smith, G. P. (1976) Science 191, 528–535.
[10] Miklos, G. L. G. and Gill, A. C. (1981) Chromosoma in press.
[11] Varley, J. M., Macgregor, H. C. and Erba, H. P. (1980) Nature 283, 686–688.
[12] Manuelidis, L. (1978) Chromosoma (Berlin) 66, 23–32.
[13] Pardue, M. and Gall, J. (1970) Science 168, 1356–1358.
[14] Brown, S. D. M. and Dover, G. A. (1980) Nucleic Acids Res. 8, 781–792.