Sequence relationships between single repeat units of highly reiterated African Green monkey DNA

Ronald E. Thayer, Maxine F. Singer and Thomas F. McCutchan

Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, USA

Received 10 October 1980

ABSTRACT

Individual monomer and dimer units of the highly repeated α -component DNA of African Green monkeys were isolated and amplified by molecular cloning in pBR322. The purified sequences were characterized by digestion with restriction endonucleases and by primary nucleotide sequence analysis. Comparison of the cloned units with the 172 base pair long sequence representing the most abundant nucleotide at each position in the set of sequences comprising α -component allows the following conclusions. The set of sequences comprising α -component is made up of a very large number of related but slightly divergent sequences. Two neighboring repeats of the monomer unit are not necessarily more similar to one another than are randomly isolated monomers.

One of the cloned segments is an almost perfect inverted repeat of a portion of the α -component monomer. Although cloning artefacts cannot now be excluded, the possibility that such inversions occur in the monkey genome is of interest.

INTRODUCTION

The highly reiterated DNA of the African Green monkey (Cercopithecus aethiops) termed α -component (1) has been characterized in several ways. It was first defined as a rapidly reannealing component of total DNA with a density of 1.699 g/cm³ (1). Subsequent work demonstrated that α -component is composed largely, if not entirely, of tandemly repeated segments (2,3,4) 172 base pairs in length (5). Various estimates indicate that 13 to 20 percent by weight of the African Green Monkey (AGM) DNA is α -component (1,4,6) while hybridization kinetic studies demonstrated approximately 5×10^6 copies of the α -component monomer per haploid genome (6).

The presence of a recognition site for the restriction endonuclease HindIII within most of the 172 base pair long repeat units proved a convenient tool for purification of the sequence. Digestion of total AGM DNA with endo R-HindIII directly yields relatively pure preparations of the

monomeric repeat unit [AGMr(HindIII)-1] as well as lesser amounts of multimers [e.g., AGMr(HindIII)-2, -3, etc]] (2-5). The multimers arise when one or more tandemly repeated units lack the usual HindIII site.

A primary nucleotide sequence has been determined for AGMr(HindIII)-1 isolated from total AGM DNA (5); the sequence is not internally repetitive. A variety of data indicate that AGMr(HindIII)-1 is a heterogeneous set of closely related DNA segments (2,4,5). Thus, the determined primary structure represents the most abundant base pair at each of the 172 positions within the set of segments, but not necessarily the structure of any particular member of the set. We will refer to this sequence (see Figure 3) as the average sequence for AGMr(HindIII)-1. Previous experiments suggested that at least 90 percent of the molecules in AGMr(HindIII)-1 contain the most abundant nucleotide at any particular position (5). Divergence of individual members of the set from the average sequence appeared to occur more frequently at certain positions in the sequence than at others (5) suggesting that divergence may not be completely random.

The number of different sequences in the AGMr(HindIII)-1 set is unknown. Also, very little is known about the arrangement of differing members of the set relative to one another within tandemly repeated regions. The presence of AGMr(HindIII)-2 and higher multimers in limit digests of AGM DNA or α -component suggests that dissimilar members of the set can occur next to one another. On the other hand, experiments with endo R^{*}EcoRI indicate that those members of the set that contain an EcoRI site may tend to be clustered (4). We report here studies designed to begin elucidating the extent and nature of divergence in individual members of the set of repeated segments and the organization of divergent sequences relative to one another. For this purpose, individual segments of AGMr(HindIII)-1 and AGMr(HindIII)-2 were isolated and amplified by molecular cloning in *Escherichia coli* and their structures were analyzed by restriction endonuclease digestion and primary nucleotide sequence determination.

MATERIALS AND METHODS

Enzymes. Restriction endonucleases were purchased from either Boehringer-Mannheim or New England Biolabs; digestions were carried out in enzyme excess using the conditions recommended by the manufacturer. Polynucleotide kinase and calf intestinal alkaline phosphatase (used for 5' end-labeling with [γ -³²P]-ATP) were obtained from Boehringer-Mannheim.

Isolation of AGMr(HindIII)-1 and AGMr(HindIII)-2. The isolation and puri-

fication of AGMr(HindIII)-1 and -2 were as described (5). Purified AGMr-(HindIII)-2 was resistant to further digestion by endo R-HindIII.

Gel electrophoresis. Analytical and preparative gel electrophoresis (except for sequencing gels) was on 5% polyacrylamide with a ratio of 20 to 1 acrylamide to bis-acrylamide. Staining of gels with ethidium bromide was as described (7).

Molecular Cloning AGMr(HindIII)-1 and AGMr(HindIII)-2. AGMr(HindIII)-1 and -2 were cloned using the host/vector system *E. coli* X1776 (8)/pBR322 (9). The monomer and dimer units were inserted into the HindIII site of pBR322 (10) and transfected into X1776 by standard procedures (11,12). Detection of recombinant plasmids and isolation and purification of plasmid DNA was described previously (13,14); AGMr(HindIII)-1 labeled with ^{32}P by nick translation (15) was used as a probe. [We found tetracycline sensitivity (9) an ineffective screening device for these plasmids. Some colonies were as resistant as cells transfected with pBR322 itself while others were sensitive, with no apparent correlation with known factors.] Recombinant plasmids are named pCa with an identifying number. Those plasmids that were subjected to primary sequence analysis were re-transfected (14) into *E. coli* strain RR1 (PRC #400 from the Plasmid Reference Center, Stanford University) to increase the yield of plasmid DNA. All experiments were carried out under containment conditions specified by the NIH Guidelines for Recombinant DNA Research.

Sequencing. Fragments used for sequence determination (16) were labeled at the 5' end with [$\gamma\text{-P}^{32}$]ATP (ICN). Whenever possible all sequencing was done with fragments several hundred base pairs longer (distal to the 3' end) than the fragment of interest, since this seemed to clarify the banding pattern at distances greater than 80 to 100 base pairs in length.

RESULTS

Characterization of cloned AGMr(HindIII)-1 and -2 units using restriction endonucleases. There are several sites in the average sequence for AGMr(HindIII)-1 (see Figure 1) where a single base pair change would create a new restriction site (5). Similarly, single base pair changes can eliminate restriction sites found in the average sequence. In particular, there are two places where a single change would create an EcoRI site and five places where a single change would create a HaeIII site (Figure 1). In order to determine whether these possible changes occurred in the individual sequences obtained by cloning, twelve plasmids containing AGMr(HindIII)-1 and three

Figure 1. Map of the restriction sites used to analyze AGMr(HindIII)-1 and -2 cloned inserts. Positions 1 to 7 denote sites in the average sequence of AGMr(HindIII)-1 where a single base pair change would create a new restriction endonuclease cleavage site (see Figure 3). A single base pair change at position 1 or 2 would create a new EcoRI site, a single base pair change at 3,4,5,6 or 7 would create a HaeIII site. Positions 8 and 9 denote cleavage sites within the average sequence for HphI and MboII respectively.

containing AGMr(HindIII)-2 were digested with endo R·EcoRI and endo R·HaeIII and analyzed by polyacrylamide gel electrophoresis. The results are summarized in Table 1.

For analysis with endo R·EcoRI the plasmids were first linearized by cleavage at the single BamHI site in pBR322 (see Figure 2) (10), and then treated with endo R·EcoRI. If the AGMr(HindIII)-1 or -2 units contain no EcoRI sites, two fragments are expected (Figure 2). However, if the AGMr(HindIII)-1 or -2 insert contains one or more EcoRI sites three or more fragments will be observed. The size of the fragments will depend on the number of sites, their position, and the orientation of the cloned fragment. Thus, the procedure determines the orientation of the insert as well as the position of any EcoRI site. Two inserts were cleaved by endo R·EcoRI, pCal0 and pCal003. Both cleavages were at EcoRI position 1 as defined in Figure 1. Only one of the two repeat units of the dimeric insert in pCal003 was cleaved.

For analysis with endo R·HaeIII two separate procedures were used. Plasmids containing AGMr(HindIII)-1 were digested with endo R·HaeIII directly. Regardless of whether the insert contains no HaeIII site or many HaeIII sites, the expected fragments are easily separated from the other

Table 1

Results of restriction endonuclease analysis on AGMr(HindIII)-1 and -2 clones

Plasmids	Insert AGMr(HindIII)-	Variations in restriction endonuclease sites			
		1/ EcoRI	2/ HaeIII	HphI	MboII
pCa9	1				
10	1	+1			
13	1				
15	1				
17	1				
18	1		+7		-9
85	1				
119	1			-8	-9
289	1				
308	1				
312	1				
329	1				
1003a	3/ 2			-8	-9
b		+1			
1004a	2				
b					
1005a	2				
b					

1/ Indicates whether the insert is AGMr(HindIII)-1 or AGMr(HindIII)-2.

2/ A plus (+) under EcoRI or HaeIII denotes a restriction endonuclease site present in the cloned fragment but missing in the average sequence. All the listed clones were tested with both enzymes and thus no entry means no change. A minus (-) under HphI or MboII denotes a restriction site present in the average sequence but missing in the cloned fragment. Only pCa18, pCa85, pCa119, pCa329, pCa1003, pCa1004 and pCa1005 were tested. The number after each change refers to the site shown in Fig. 1.

3/ a or b indicates which of the two repeat units in the dimer is being described; a is the unit closest to the EcoRI site of pBR322 (see Figure 2).

products of digestion of pBR322 (10, and see Figure 1). Only one monomer insert (pCa18) was cleaved by HaeIII (Table 1). From the fragment sizes it was determined that pCa18 was cleaved by HaeIII at position 7 (Fig. 1). AGMr(HindIII)-2 cannot be analyzed by this method since the HaeIII fragment containing the insert is only four base pairs smaller than another HaeIII fragment from pBR322. To analyze the dimers for HaeIII cleavage sites the

Figure 2. The strategy used to sequence cloned AGMr(HindIII)-1 and AGMr(HindIII)-2. All monomers and portions of the dimers were sequenced using Method A. Method B was used to sequence the distal portion of the second repeat unit in the dimers. Method A. The recombinant plasmids were cleaved by combined digestion with endo R·EcoRI and endo R·SalI in EcoRI buffer. The resulting fragments were treated with phosphatase to remove the 5' phosphate and 5'-end labeled with ^{32}P using polynucleotide kinase and $[\gamma^{32}\text{P}]\text{ATP}$ (16). The unincorporated label was removed by chromatography on Sephadex G50 followed by ethanol precipitation. The fragment carrying the insert was cleaved with endo R·BamHI and the fragment to be sequenced was separated from the other labeled fragments on a 5% acrylamide gel and eluted using standard procedures (16). Method B. The plasmids were cleaved with endo R·HindIII and labeled with ^{32}P as described above. Cleavage of the labeled fragments with endo R·MboII yielded the sequence of interest in a fragment that was readily separable from the other labeled fragments by gel electrophoresis. Because the dimeric units were inserted into the plasmid so that position 172 of the repeat was proximal to the EcoRI site of the vector, this procedure afforded good sequence data in the region that was inaccessible by Method A.

insert was first removed with HindIII, isolated by standard procedures (16) and then treated with HaeIII. None of the three dimers screened contained a HaeIII site (Table 1). Several monomer and dimer clones were also analyzed with endo R·HphI and endo R·MboII (Table 1), sites for which are present in the average AGMr(HindIII)-1 sequence (Figure 1). The inserts were isolated

as described above and 5'-end labeled with ^{32}P before digestion. Some of the individual cloned units lacked either one or both of these two sites. Sequence analysis. Six of the individual cloned segments, four monomers and two dimers, were subjected to sequence determination. The sequencing strategy is outlined in Figure 2 and the results are summarized in Figure 3. None of the repeat units had the average AGMr(HindIII)-1 sequence. Except for the insert in pCal19 all the sequenced repeat units contained 172 base pairs and all changes from the average sequence were base pair alterations. The number of base pair changes from the average sequence was four in pCa85, pCa329 and pCal004b (and pCal005b), five in pCal004a (and pCal005a), and seven in pCal18.

pCal004 and pCal005 have identical nucleotide sequences throughout both repeat units. This was surprising since, in the selection of transfected clones, care was taken to isolate well-separated colonies. We do not know if the identity of pCal004 and pCal005 reflects inadvertent selection of the same clone or a relatively high frequency of this sequence in AGMr(HindIII)-2. Therefore in analysis of the sequences (see Discussion) we have considered the two dimer inserts as one sequence. The primary sequence data demonstrate directly that the loss of the HindIII site between the two monomer units is the result of an altered nucleotide sequence rather than any secondary modification of the bases. pCal003 is clearly different from pCal004 and pCal005 (Table 1).

The insert in pCal19 is markedly different from the average sequence and from all the other cloned inserts. First, it is only 150 base pairs long. Second, its primary structure is unique in several ways. The first half of the molecule contains, in addition to two variations in base pairs, three insertions of extra base pairs relative to the average sequence (Figure 3). The homology with the average sequence seems to disappear in the second half of the molecule (Figure 3) but in fact, as shown in Figure 4, the second half of the pCal19 insert is an inverted repeat of the first half of the average sequence. Comparing the inverted sequence of the second half of pCal19 with the first portion of the average sequence (Figure 4), there are three variant positions and remarkably, these are contiguous to one another. Thus, while pCal19 contains two copies of the first half of the average sequence arranged in an inverted repeat, the two copies differ markedly from one another and we conclude that they represent two different members of the α -component set. We do not know if the inverted repeat configuration in pCal19 represents a true AGM genomic segment or whether it is an artefact

Figure 3. Summary of sequencing data. The average sequence for ACMr(HindIII)-1 (5) is shown above and the variations in the cloned monomer and dimer units noted below. Each of the two repeat units of the dimers is listed separately. "a" is the repeat unit closest to the EcoRI site of pBR322. For convenience the presentation of all sequences is in the direction shown in Fig. 1 regardless of the actual orientation within the plasmid. A zero (0) indicates a base pair deletion. A line between two base pairs indicates an insertion.

Figure 4. Complete structure of the insert in pCal19 displayed as an inverted repeat. Those residues representing an insertion in the average sequence are underlined; those representing a base change from the average sequence are overlined; the zero indicates a possible deletion from the average sequence, but as this was at the end region of the sequencing gel we are uncertain. This display numbers 154 residues rather than 150 because both terminal HindIII sites are shown in their entirety.

of the cloning procedures (see Discussion).

DISCUSSION

As summarized in the Introduction earlier work indicated that the many repeats of the α -component monomer within the AGM genome comprise a set of related but not identical sequences. Experiments conducted with mixed populations of molecules could not, however, resolve the question of the relative abundance of different members of the set. The data were consistent with a large preponderance of average sequence segments (and a smaller percent of diverse variants) or a large mixture of different segments in which the complete average sequence was rare, if it occurred at all. Molecular cloning permits the isolation and amplification of individual members of the set for analysis, thereby providing information on the sequence diversity in α -component DNA.

The present experiments indicate that α -component is a large mixture of different segments. The sequence of five different 172 base pair long monomeric units was determined and all of them diverge from the average sequence. Also, both halves of the inverted repeat in pCal19 diverge from the average sequence and from one another. Restriction enzyme analysis indicates that at least 3 more of the cloned segments (pCa10, pCa1003a and pCa1003b) differ from the average sequence. Graf and coworkers (17,18) have also isolated cloned α -component units. Of nine cloned segments analyzed (17) with endo R·EcoRI, three were cleaved. Partial sequence determination indicated a few changes from the average sequence in at least three (18). Also, individual segments of α -component that occur in defective variants of simian virus 40 differ from the average sequence in a few positions (19).

Thus the average sequence has never been found. We conclude that α -component contains a large number of different sequences and that the average sequence occurs rarely if at all. Excluding pCall9 the overall extent of divergence from the average sequence is about 3 percent (25 changes out of a total of 860 base pairs sequenced), in agreement with previous estimates (4, 18).

Figure 2 shows various positions on the average sequence where a single base pair change would result in the formation or loss of a particular restriction endonuclease site. Earlier work (2,4,5,17) indicated that divergence to an EcoRI site occurs more frequently at EcoRI site 1 than at EcoRI site 2. Similarly, divergence at HaeIII site 7 was more likely than the other possible changes. These findings lead to the suggestion that divergence at certain nucleotide residues occurs more frequently than divergence at other residues (5). The patterns of restriction site divergence observed here are consistent with the previous observations. Nevertheless, the sequence data in Figure 3 shows little or no clustering of divergences when the different cloned segments are compared [with the possible exception of the two residues (71 and 136) where two changes from the average sequence were observed]. Even with the few sequences reported here no region of the average sequence greater than 18 base pairs in length is without a divergence. While the number of individual repeats that have been analyzed is too small in comparison with the total number of repeats in the monkey genome to allow a conclusion, it is certainly possible that the observed nonrandom divergence reflects a particular subgroup of repeats and is superimposed upon a generalized random variation. Such a result would be expected if certain initially randomly diverged sequences were preferentially amplified and maintained. This interpretation is supported by indications that the repeat units containing EcoRI sites may be clustered (4). It is also interesting that our results are markedly different from data obtained with the complex satellite of Drosophila melanogaster (20). In that instance divergences occur predominantly in only a limited number of the 359 residues in the repeat unit.

Another outstanding question regarding the structure of α -component is the extent to which neighboring repeat units are structurally related. The existence of dimers and higher multimers after complete digestion of AGM DNA or α -component itself with endo R·HindIII or other restriction enzymes (2, 3,4,5) indicated that dissimilar repeat units can exist side by side. The present analysis of AGMr(HindIII)-2 clones confirms this directly. The data show that the absence of the HindIII site between repeat units results from

a change in base sequence rather than a secondary base modification. The two halves of pCal003 (Table 1) and the two halves of pCal004 (and pCal005) (Table 1, Figure 3) differ in other residues as well. The sequences of the two tandem repeats in pCal004 (and pCal005) differ from one another at 8 positions. This is similar to the number (8-11) of different positions between the pairs of sequenced monomers (excluding pCal19). Thus our data provide no indication that neighboring repeat units are more similar to one another than dispersed repeat units.

The sequence analysis of pCal19 gave an unexpected result. The inverted repetition of half of the AGMr(HindIII)-1 sequence, should it prove to be a configuration present in the monkey genome, would be of great significance to the organization of repeated DNA sequences. However additional experiments will be required to confirm this possibility since it is presently impossible to know the origin of the configuration. The fortuitous joining of two AGMr(HindIII)-1 units in a head to head configuration during the cloning procedure would have generated a 344 base pair long almost perfect inverted repeat and elimination of the central portion of such a potential hairpin structure during amplification in *E. coli* might have given rise to the pCal19 molecule. Such an elimination could have been fostered either by hair-pin formation in a single-strand or by recombination involving short direct repeats flanking the deletion (for example, short inverted repeats in the region 68 to 73 and 79 to 84 in the average sequence, Figure 3, will give rise to direct repeats in a putative head to head dimer). The hair-pin mechanism raises the problem of explaining why a 150 base pair long inverted repeat would prove stable while the initial 344 base pair long inverted repeat was unstable. On the other hand, the mechanism involving short direct repeats seems unlikely in view of the stability of the AGMr(HindIII)-2 plasmids and earlier reports of the stability of tandem repeats of complex eukaryote repetitive DNA segments in *E. coli* (21). In any case, any deletion must have occurred during the first round of amplification of the recombinant plasmid because even our first quick isolate of this plasmid from the host X1776 indicated that the insert was shorter than expected. Thereafter the insert remained stable, at least as far as length is concerned. Also, preparations of pCal19 appear homogeneous.

Finally, we point out that the function of highly repeated DNAs like α -component remains unknown. As has been widely discussed for some time, it is possible that these sequences are without function at least in the classical genetic sense. With one exception (22), transcription of highly

repeated DNAs has not been detected. On the other hand, the initial impression of millions of redundant repeats of a single DNA sequence is now being replaced by a growing appreciation of the divergence between repeat units and the nonrandom nature of at least a portion of that divergence (this paper and see review in reference 23). This is particularly so for complex repeating units like those of α -component in contrast to the simple satellite DNAs where the repeat unit is less than 10 nucleotide residues in length. Work in the last five years has amply demonstrated that the change of a single base pair in a DNA sequence can have profound biological effects (apart from effects on coding). The emerging sense of the complexity of highly repeated DNA suggests, at least in theory, that the highly repeated DNA fraction of eukaryote genomes could represent an extensive bank of subtle information for use in cell function.

ACKNOWLEDGMENTS

We thank H.G. Zachau and H. Graf for making unpublished data available to us, Martin Rosenberg for helpful discussions, and May Liu for her careful preparation of the manuscript.

REFERENCES

1. Maio, J.J. (1971) *J. Mol. Biol.* 56, 579-595.
2. Brown, F.L., Musich, P.R. and Maio, J.J. (1979). *J. Mol. Biol.* 131, 777-779.
3. Gruss, P. and Sauer, G. (1975) *FEBS Lett.* 60, 85-88.
4. Fittler, F. (1977) *Eur. J. Biochem.* 74, 343-352.
5. Rosenberg, H., Singer, M.F. and Rosenberg, M. (1978) *Science* 200, 394-402.
6. Singer, D.S. (1979) *J. Biol. Chem.* 254, 5506-5514.
7. Sharp, P.A. Sugden, B. and Sambrook J. (1973) *Biochemistry* 12, 3055-3064.
8. Curtiss III, R., Pereira, D.A., Hsu, J.C., Hull, S.C., Clark, J.E., Maturin, L.J., Sr., Goldschmidt, R., Moody, R., Inoue, M. and Alexander, L. (1976) *Recombinant Molecules: Impact on Science and Society*, Beers, R.F. Jr. and Bassett, E.G. Eds., pp45-56, Raven Press, New York.
9. Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.C., Heyneker, H.L. Boyer, H.W., Crosa, J.H. and Falkow, S. (1977) *Gene* 2, 95-113.
10. Sutcliffe, J.G. (1978) *Nucleic Acids Res.* 5, 2721-2728.
11. Cohen, S.N., Chang, A.C.Y. and Hsu, L. (1972) *Proc. Natl. Acad. Sci. USA* 69, 2110-2114.
12. Mandel, M. and Higa, A. (1970) *J. Mol. Biol.* 53, 159-162.
13. Thayer, R.E. (1979) *Anal. Biochem.* 98, 60-63.
14. Papamatheakis, J., Kuff, E., Winocour, E. and Singer, M.F. (1980) *J. Biol. Chem.*, in press.
15. Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) *J. Mol. Biol.* 113, 237-251.
16. Maxam A. and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* 74, 560-564.
17. Graf, H., Fittler, F. and Zachau, H.G. (1979) *Gene* 5, 93-110.
18. Graf, H. (1979) *Hoppe Seyler's Z. für Physiol. Chemie*, 360, 1029.

19. McCutchan, T., Singer, M.F. and Rosenberg, M. (1979) J. Biol. Chem. 254, 3592-3597.
20. Hsieh, T. and Brutlag, D. (1979) J. Mol. Biol. 135, 456-481.
21. Carlson, M. and Brutlag, D. (1977) Cell 11, 371-381.
22. Varley, J.M., Macgregor, H.C. and Erba, H.P. (1980) Nature 283, 686-688.
23. Brutlag, D.L. (1980) Annu. Rev. Genetics, in press.