Sequence Organization of the Rat Genome by Electron Microscopy†

Mahlon M. Wilkes,‡ William R. Pearson, Jung-Rung Wu, and James Bonner*

ABSTRACT: The size and arrangement of repetitive and inverted repeat (foldback) sequences in rat DNA were studied by visualization of hybrid and heteroduplex structures in the electron microscope. The self-reassociation of repetitive sequence-bearing DNA strands often results in the formation of four-ended "H" structures, whose duplex regions equal the repetitive sequence length and can be measured in the electron microscope. In this way, it was determined that the average size of the class of numerous short repetitive sequences is 0.40 ± 0.15 kbp. Heteroduplex structures were prepared between long whole DNA single strands and short repeat-sequence-bearing strands. The analysis of these structures confirms that the size of the repetitive sequences is 0.4 kbp on average. Length measurements between adjacent duplexes show that the average spacing between two interspersed repeats is at least 1.5–1.8 kbp. By examining 29.4-kbp single strands after brief renaturation, the size and distribution of foldback sequences were determined. There are 1.9×10^5 foldback pairs per rat genome, spaced an average of 9.7 kbp apart according to our measurement. Repetitive, inverted repeat and unique sequences are interspersed with each other in at least half the genome.

A desire to understand the structural basis of eucaryotic gene regulation has stimulated studies of the physical organization of DNA sequences. Detailed models of the *Xenopus* (Davidson et al., 1973) and sea urchin genomes (Graham et al., 1974) have been derived from hydroxylapatite binding data by computational methods. In *Xenopus* half the DNA consists of repetitive sequences about 0.3 kbp[1] long interspersed with unique sequences of 0.7–0.9 kbp average in length. One-quarter of the DNA is similarly arranged but with the average unique sequence length in excess of 4.0 kbp. Regions of exclusively unique DNA and of clustered repetitive sequences comprise the remainder of the genome. In sea urchin the average unique sequence length in the short period interspersed portion of the genome is about 1 kbp though otherwise this genome is organized much like that of *Xenopus*. *Aplysia* DNA also generally corresponds to this pattern (Angerer et al., 1975).

An extension of the techniques of Wu et al. (1972) can in principle provide direct measurements of repetitive and unique sequence length distributions. We chemically isolate moderately short DNA strands, which are nonetheless larger than the average repetitive sequence, each containing repetitive DNA. The repetitive sequences embedded within the strands will be terminated with either one or two unique DNA flanking sequences. When this DNA is renatured to a repetitive C_0t, it will form H structures: a duplex flanked in general by four single-stranded, unique sequence tails. In this way we can measure the length of the repetitive sequences.

When short DNA fragments bearing repetitive sequences are renatured to a repetitive C_0t and in vast excess to long, unsheared total DNA single strands, heteroduplex structures form. These structures consist of long DNA strands, whose repetitive sequences have renatured to the complementary sequences among the driver DNA population. The unique DNA sequences flanking the driver strand repeats will be noncomplementary to the sequences flanking the repeats to which they have renatured, as in the formation of H structures. Therefore, these unique DNA tails will project from the long strands and be identifiable in the electron microscope. Measurement of the duplex regions terminated by two tails yields a second measure of the length of repetitive sequences. Measurement of the single-stranded regions bounded by repetitive duplex structures gives the unique DNA sequence length. The present study extends previous work on repetitive sequence interspersion in rat DNA (Bonner et al., 1973).

An additional feature of this technique is that it identifies inverted repeat, or foldback, sequences in the DNA. Foldbacks have been studied previously in the electron microscope (Wilson and Thomas, 1974; Schmid et al., 1975; Cech and Hearst, 1975), because they form hairpin or looped hairpin structures easily distinguishable under formamide spreading conditions. It is shown below that the rat genome contains 1.9×10^5 inverted duplex repeat sequences of 0.71 kbp number average in length spaced an average of 10 kbp apart. Foldback duplexes form in a rapid, concentration-independent reaction, when DNA single strands are renatured under appropriate conditions. The long DNA strands used in this study of repetitive sequences arrangement can be anticipated to form hairpin structures in addition to heteroduplexes. Since these two types of structures can occur on the same long DNA strand, an assessment of the relative spatial arrangement of repeat and inverted repeat sequences can be made. However, in the case of unlooped hairpin structures there is ambiguity in interpretation, since these may have the same appearance as repetitive hybrids with only one tail. We study the distribution of these foldbacks by use of long strands of DNA renatured without any driver for short times.

Experimental Procedures

Preparation of DNA. Long DNA was isolated from rat ascites cells as described by Pearson et al. (1978). Shearing of the 0.9, 1.7, and 2.5 kbp drivers was performed in a Virtis homogenizer at 15 000, 5000, and 2000 rpm, respectively, at 5 °C in 0.05 M neutral phosphate buffer. Alkaline denaturation of the DNA followed the method of Manning et al. (1975). All

† From the Division of Biology, California Institute of Technology, Pasadena, California 91125. *Received July 29, 1977.* This work was supported in part by U.S. Public Health Service Research Grants GM 13762 and GM 20927.

‡ Present address: Department of Reproductive Medicine, University of California, San Diego, La Jolla, California 92037.

[1] Abbreviations used: kbp, kilobase pairs; rpm, revolutions per minute; EDTA, ethylenediaminetetraacetic acid; C_0t, concentration in moles of nucleotide from liter times time in seconds; EM, electron microscope; HAP, hydroxylapatite.

0006-2960/78/0417-0060$01.00/0 © 1978 American Chemical Society

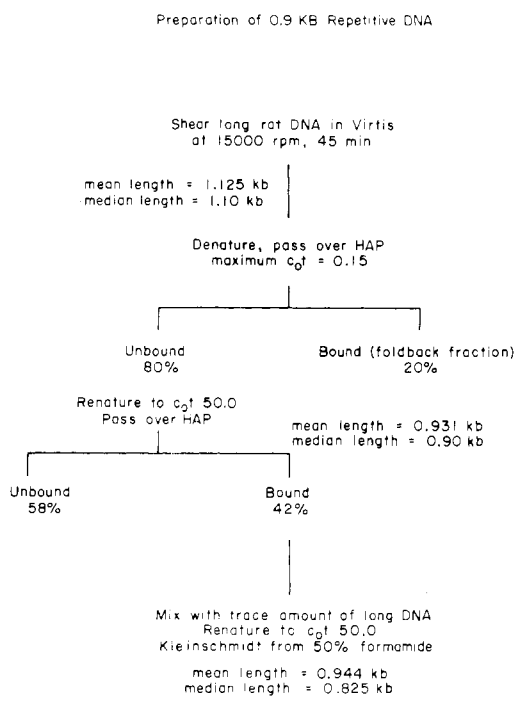

FIGURE 1: Preparation of 0.9-kbp repetitive DNA. Conditions of shearing, incubation, and HAP chromatography appear in Experimental Procedures. All length measurements were made by electron microscopy at the point in the isolation designated in the flowchart.

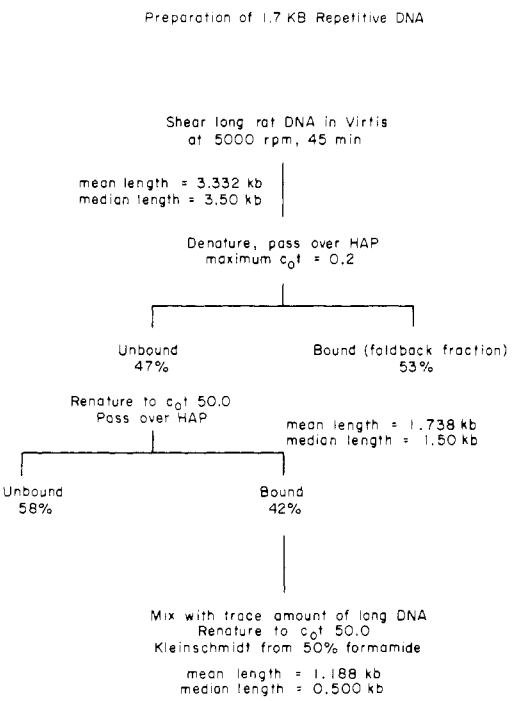

FIGURE 2: Preparation of 1.7-kbp repetitive DNA. See legend to Figure 1.

incubations and HAP chromatographic procedures were conducted at 60 °C in 0.12 M phosphate buffer. Elution of bound DNA from HAP was with 0.12 M Na_3PO_4 at 60 °C. Details about the pedigree of the driver DNA preparations are given in Figures 1–3.

Electron Microscopy. DNA was dialyzed against 0.01 M Tris, 0.001 M EDTA, pH 8.5, made 50% in formamide and spread for microscopy by the modified Kleinschmidt technique of Davis et al. (1971). The DNA was visualized in a Philips EM 201 at an accelerating voltage of 60 kV. The grids contained single- and double-stranded ϕX174 circular DNA molecules as internal length standards of known length 5.25 kbp. The ratio of the measured single to double strand length of a given DNA strand region was 0.97, so we treated single strand and duplex measurements without distinction. Micrographs were recorded on 35-mm film which was projected directly onto the electrosensitive platen of a Hewlett-Packard electronic digitizer.

Results

Characteristics of Rat DNA. Twenty percent of the rat genome is comprised of repetitive sequences; 75% of unique sequences; 5% of foldback sequences (Holmes and Bonner, 1974; Pearson et al., 1978). Virtually all the repetitive sequences reassociate in the C_0t interval 0.005–50. It is possible to remove foldback sequences from the DNA by renaturation to low C_0t followed by removal of duplex-bearing DNA by hydroxylapatite chromatography. Unique DNA may be removed by renaturation to C_0t 50 followed by isolation of hydroxylapatite bound DNA. This DNA now contains only strands containing repetitive sequences. Most of these repetitive sequences have two unique DNA flanking sequences.

We prepared three repetitive DNA fractions of 0.9, 1.7, and 2.5 kbp number-average length by the protocols shown in Figures 1–3. These lengths represent the average size of the input strands in the renaturation reactions described below and in Figures 1–3 at the bottom. However, it should be noted that

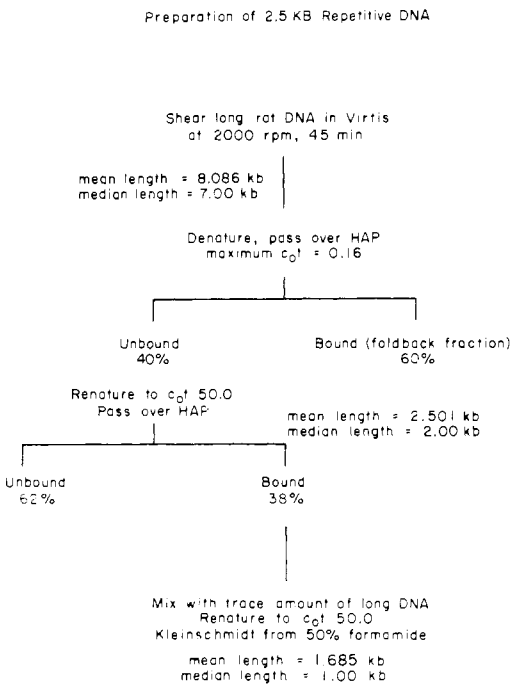

FIGURE 3: Preparation of 2.5-kbp repetitive DNA. See legend to Figure 1.

the average size of the overall strand populations after HAP chromatography and mounting for EM were reduced by breakage to 0.9, 1.2, and 1.7 kbp, respectively. The details of the isolation are as follows: 1.1, 3.3, and 8.1 kbp DNA strands were briefly renatured and applied to hydroxylapatite (maximum C_0t = 0.16). Respectively, 20%, 53%, and 60% of the DNA bound. The unbound strands measured 0.9, 1.7, and 2.5 kbp on the average after this procedure. These stripped preparations were then renatured to C_0t 50 and fractionated on HAP. The respective percents bound for the fractions were 42, 42, and 38%. These three bound fractions we designate the 0.9, 1.7, and 2.5 kbp repetitive DNA fractions.

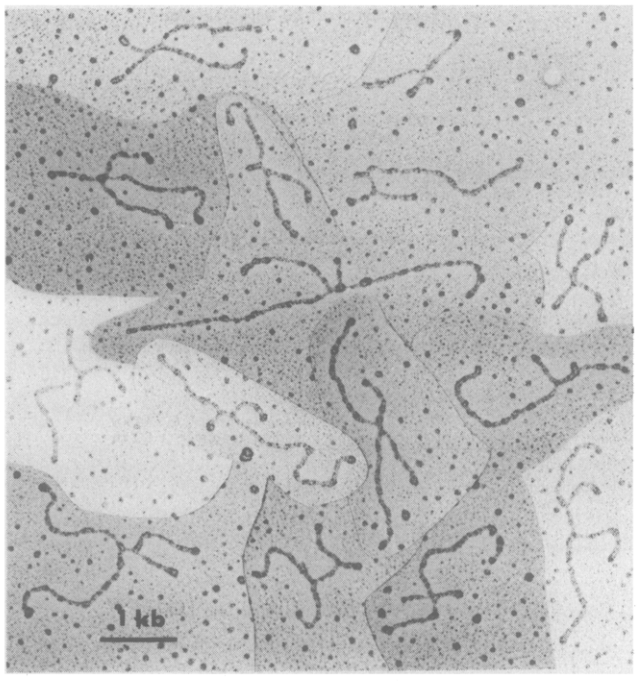

FIGURE 4: "H" structures formed among the 2.5-kbp repetitive DNA strands. A marker in the micrograph gives the equivalent length of 1 kbp.

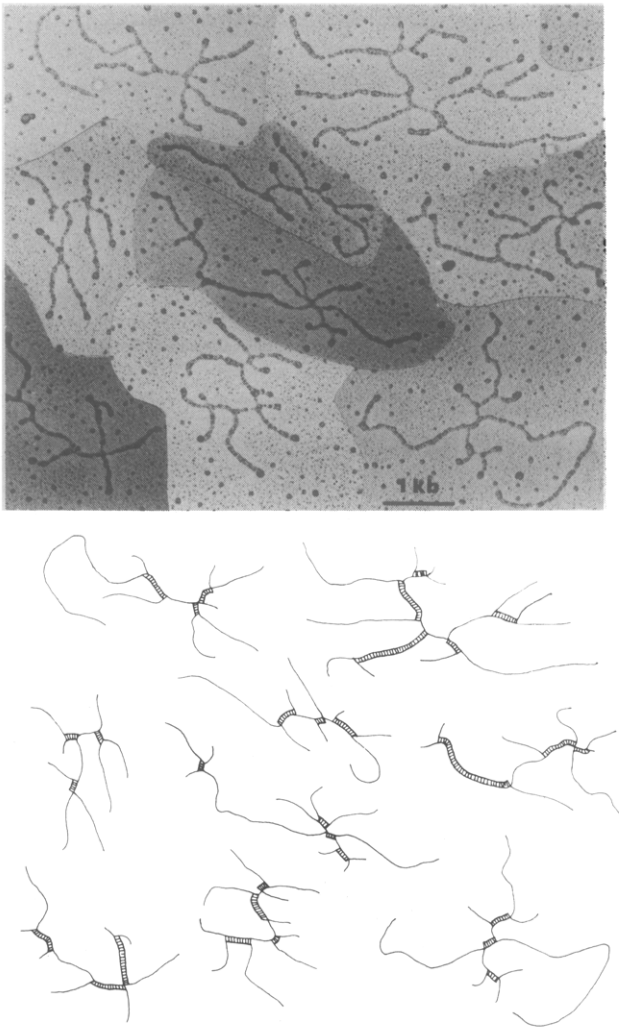

FIGURE 5: Multiple duplex structures formed among the 2.5-kbp repetitive DNA strands. Scale line drawings of the interpretations placed on the structures in the composite micrograph appear below; single lines are duplex strands; double, cross-hatched lines are duplex regions. When an ambiguity arose, e.g., an odd number of tails in the hybrid structure, one reasonable interpretation of the complex was recorded. In some cases, this required us to designate certain tails to be single-tailed structures analogous to the three-ended molecule class; however, no duplex measurement of one-tailed structures was attempted. We draw a short duplex and dotted line to indicate the presence of a one-tailed duplex structure.

The repetitive fragments were next denatured and then reassociated again to C_0t 50 in the presence of a trace quantity of long unsheared, unfractionated DNA single strands to form hybrids, which were spread for electron microscopy from 50% formamide. All incubations and HAP fractionations were carried out in 0.12 M phosphate buffer at 60 °C. Since the strand sizes prepared are greater than the repetitive sequence size (Bonner et al., 1973), many of the DNA strands will contain a repetitive sequence flanked on one or both sides by unique sequences. Reassociation of two strands each containing a repeat flanked by two unique sequences results in the formation of a repetitive sequence duplex terminated at each end by two noncomplementary unique DNA single strands. When visualized in the electron microscope, the structures have the appearance of an H. Length measurement of the duplex region terminated at each end by the two single-stranded "tails" yields the length of the repetitive sequence. Repetitive sequence lengths have been measured in this manner by Chamberlin et al. (1975) in *Xenopus* and by Manning et al. (1975) in *Drosophila*.

In the EM the reassociation products of the short repetitive DNA strands fell into four categories: (1) four-ended or H structures, (2) three-ended structures, (3) multiple duplex structures, and (4) linear strands. The DNA on the grids was scanned at random. All the DNA observed was measured and interpreted, excluding only strands found on technically imperfect areas of the grids, e.g., those with contrast failure. Only 1% of the molecules were found in uninterpretable structures.

A composite electron micrograph of some of the H structures studied is shown in Figure 4. When several strands participate in multiple duplex formation, the resultant structures have the appearance shown in Figure 5. Statistical data on the recovery of duplex classes in the two experiments are given in Tables I and II. These results show that the strands engaged in H and multiple duplex structures are about the same size as the input DNA. However, as mentioned above, the overall average lengths of all classes of structures for the 1.7 and 2.5 kbp repetitive fractions were 1.2 and 1.7 kbp, respectively.

Table III shows the data pertaining to recovery of repetitive duplexes terminated by four single strand regions. The distributions of repetitive sequence lengths measured among the 1.7 and 2.5 kbp DNA strands are recorded. The number average repetitive sequence length in the rat is 0.4 ± 0.15 kbp according to these data. The great majority of measurements fall within the limits of 0.1–0.5 kbp. There appears to be a rather disperse, sparse population of longer duplexes. However, since the measured size of the larger duplexes is about the same magnitude as the DNA strand length, there is the possibility that any measurement on these structures would underestimate the true average length of the larger repetitive duplex class. Further, in many cases the distinction between single- and double-stranded DNA was insufficient to rule out the possibility that some structures interpreted as H structures were actually three-stranded structures. For the value of 0.4 kbp to accurately reflect the average size of the short repetitive sequences it must be established that the measured duplex lengths are independent of the lengths of the single strands on

TABLE I: Disposition of 1.7-kbp Repetitive DNA Strands Renatured to C_0t 50.

Type of structure	No. of strands scored	% of all strands scored	Total strand length in class (kbp)	% of total length in class	Strand lengths (kbp) No. av	Strand lengths (kbp) Wt av
Four-ended structures	190	8.5	317	12.0	1.667	2.094
Three-ended structures	364	16.3	373	14.1	1.024	1.738
Multiple duplex structures	259	11.7	543	20.5	2.096	3.170
Linear molecules	1413	63.4	1413	53.3	1.000	1.424
Lariats	2	0.1	4	0.1	2.000	2.022
Total	2228		2650		1.188	1.907

TABLE II: Disposition of 2.5-kbp Repetitive DNA Strands Renatured to C_0t 50.

Type of structure	No. of strands scored	% of all strands scored	Total strand length in class (kbp)	% of total length in class	Strand lengths (kbp) No. av	Strand lengths (kbp) Wt av
Four-ended structures	248	11.4	533	14.5	2.150	2.881
Three-ended structures	316	14.5	428	11.7	1.354	2.431
Multiple duplex structures	515	23.6	1394	37.9	2.707	4.008
Linear molecules	1101	50.5	1318	35.9	1.197	1.866
Circles	1	0.04	2	0.05	1.673	1.673
Lariats	1	0.04	3	0.08	3.202	3.202
Total	2182		3678		1.685	2.892

TABLE III: Recovery of Duplex Structures from Repetitive DNA Strands Renatured to C_0t 50.

Input DNA size (kbp)	Type of duplex-bearing structure	No. of duplexes scored	Total duplex DNA (kbp)	% of total DNA in duplex	Duplex lengths (kbp) No. av	Duplex lengths (kbp) Wt av
1.7	4-ended	190	154	5.8	0.406	0.781
1.7	Multiple duplex	307	220	8.4	0.357	0.601
1.7	Total	497	374	14.2	0.376	0.675
2.5	4-ended	248	238	6.4	0.478	1.006
2.5	Multiple duplex	715	600	16.4	0.420	0.757
2.5	Total	963	838	22.8	0.435	0.988

which they reside. If, for example, the average duplex size measured on a particular strand length increased continuously with the resident strand length, the possibility would exist that the true average sequence size has been underestimated. Thus, the existence of a substantial population of longer duplexes could go undetected, because the reassociating strands were insufficiently long to permit formation of a long duplex with four tails. That this is not the case is shown in Figure 6. The strands containing duplexes terminated by four single strand regions were divided into intervals measured at each strand length. Figure 6 shows that for both the 1.7 and 2.5 kbp DNA the average duplex length does not vary consistently over a wide range of strand lengths.

We can calculate the fraction of the genome composed of short repetitive sequences by a quantitative treatment owing to Chamberlin et al. (1975). The percent of the genome composed of short repetitive sequences is equal to the product of the percent of the total DNA found in duplex and the fraction of the genome constituted by the isolated repetitive DNA fractions. For the 1.7- and 2.5-kbp DNA fractions these products were 6.0 and 8.6%, respectively. The probability of formation of a duplex structure with four single strand regions at least 0.1 kbp long each is $[1 - (R + 0.2)/L]^2$ where R is the repetitive sequence length and L the strand length (Chamberlin et al., 1975). The mean size of the strands engaged in duplexes terminated by four single strand regions was 1.914 and 2.526 kbp in the 1.7- and 2.5-kbp DNA experiments (calculated from data in Tables I and II). Using these numbers for L and 0.4 for R, we calculate the probabilities of formation and multiply them by 15%, which is the approximated proportion of the genome comprised by short repeats (Pearson et al., 1978). The products represent the percent of the total DNA expected to be in duplex if recovery of repetitive sequences is complete (100%). The respective values obtained for the 1.7- and 2.5-kbp fractions were 7.0% and 8.7% in good agreement with observation (6.0% and 8.6%).

Distribution of Repetitive Sequences. The 0.9-kbp repetitive DNA fraction was renatured to 29.4 kbp long total DNA single strand in the following manner. Denatured repetitive and long DNA were mixed in a mass ratio of 100:1 and renatured to C_0t

TABLE IV: Recovery of Repetitive DNA Strands Hybridized to Long DNA.

	No. of strands scored	Experiment duplexes terminated by 2 single-stranded tails[a] No. av strand length (kbp)	No. av tail length (kbp)	No. of tails proximal to contiguous repeats[b]	No. av length of tails proximal to contiguous repeats (kbp)
Long vs. 0.9-kbp repeat DNA	296	1.314	0.490	54	0.496
Long vs. 2.5-kbp repeat DNA	618	2.392	0.937	64	0.865

[a] These data apply to cases we interpreted to be repetitive hybrid duplexes consisting of a duplex region terminated at both ends by single-stranded unique DNA tails radiating from the long backbone strand of the heteroduplex structure. In cases wherein the duplex region separating two tails could not be assigned a strandedness visually, it was assumed that this region is double-stranded provided the tails are in reasonably close adjacence to each other. In cases in which there was ambiguity of interpretation, e.g., an odd number of tails close together, one likely interpretation was recorded, assigning the odd tail to the one-tailed duplex category. In most cases no such ambiguities occur. [b] This category includes the two tails bounding the injunction between two contiguous repeats.

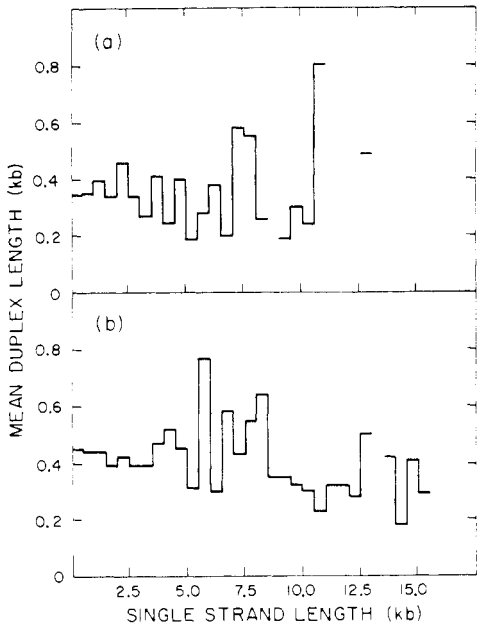

FIGURE 6: Mean duplex length as a function of single strand length. Mean length of duplexes terminated by four-ended as a function of single strand length. (a) For 1.7-kbp DNA fraction; (b) 2.5-kbp DNA fraction.

50 in 0.12 M phosphate buffer at 60 °C. Repetitive sequences comprise 20% of the long tracer (Holmes and Bonner, 1974). The short driver contains on average 45% repetitive sequences. Therefore the driver to tracer repetitive sequence ratio is ca. 200:1. At the end of the reassociation, part of the DNA was prepared for electron microscopy, while the remainder was mixed with freshly alkali-denatured and neutralized driver DNA. The driver DNA from the first phase of the reaction was considered to be inert with respect to further reaction. In this second phase of reaction the mass and sequence excesses of fresh driver to tracer were 1700:1 and 3400:1, respectively. The reaction was carried to C_0t 50 as before. These DNA samples were mounted for electron microscopy at 25 °C from 50% formamide by the modified Kleinschmidt method of Davis et al. (1971). This formamide concentration maintains the criterion of duplex formation equivalent to 0.12 M phosphate buffer at 60 °C.

The 2.5-kbp repetitive strands and long DNA were renatured to C_0t of 50 in mass and sequence ratios of 120:1 and 200:1, respectively. Conditions of renaturation were as above. No second reaction of the repetitive and long DNA hybrids was performed for the 2.5-kbp DNA.

Recovery of Hybrid Structures. The grids containing the various DNA preparations were scanned at random to the exclusion only of areas of obvious technical imperfection. All the DNA in the scanned regions was photographed and interpreted. The self-reassociation products of the driver DNA had been described above. Uninterpretable structures accounted for 1% of the total DNA and were disregarded. To collect additional repetitive/long DNA structures we scanned the grids further, photographing only the strands substantially longer than the largest of the driver molecules.

The structures found under the electron microscope fell into three main categories: (1) driver/long DNA hybrids with two tails or more, (2) duplex structures between driver strands and the tails of driver strands already renatured to a long strand, and (3) looped foldback structures. In addition, there was a substantial number of duplex structures with only one tail. These could represent either hybrids of repetitive strands flanked only at one end by a unique sequence or foldback structures with no loops. There were in the data from 0.9- and 2.5-kbp fractions, respectively, 296 and 618 cases of two-tailed hybrid structures (Table IV) compared with 62 and 120 of the one-tailed events.

Figure 7 shows hybrid structures between 0.9-kbp repetitive and long DNA and our interpretations thereof. Micrographs of 2.5-kbp repetitive/long DNA hybrids appear in Figure 8.

The number of long strands scored and their lengths are listed in Table V.

Recovery of Repetitive Duplexes. The length of the duplex regions demarcated by two tails gives a measure of the repetitive sequence length of rat DNA. We have shown above evidence of a class of repetitive sequences 0.40 ± 0.15 kbp long. A similar conclusion has been reached by means of physical chemical techniques (Pearson et al., 1978). Data on duplex recovery and in these experiments are displayed in Table VI and are consistent with this result. Among the 2.5 kbp/long DNA hybrids there are apparently a number of large repetitive duplexes (>2 kbp). However, any measurements on this larger class of sequences may strongly underestimate the true average sequence length, since the average input strand is approximately the same size as the observed sequence length. Pearson et al. (1978) present physical chemical evidence of a class of repetitive sequences in rat at least 1.5-kbp long and comprising roughly 5–10% of the genome. If we eliminate such larger repeats from the 2.5-kbp data, the resulting size for the short repetitive sequence class is 0.43 kbp. The 0.9-kbp data indicate 0.34 kbp as the number average size of this class of repetitive

TABLE V: Sequence Organization of Rat DNA.[a]

Experiment		Type of organization: Interspersed repeat and unique sequence	Interspersed repeat, foldback, and unique sequence	Interspersed foldback and unique sequence	Unique sequence	Total[b] DNA
Long vs. 0.9-kbp repeat DNA	No. scored	19	28	8	16	71
	Mean length (kbp)	15.623	12.071	20.360	17.206	15.117
	DNA in type (kbp)	297	338	163	275	1073
	% of total	27.7	31.4	15.2	25.7	
Long vs. 2.5-kbp repeat DNA	No. scored	64	60	7	8	139
	Mean length (kbp)	16.560	21.967	21.047	19.834	19.308
	DNA in type (kbp)	1060	13.8	147	159	2684
	% of total	39.5	49.1	5.5	5.9	

[a] Each of the long strands analyzed in the various experiments was sorted into four categories according to whether they possess (1) only repetitive hybrids, (2) repetitive structures and foldbacks, (3) only foldbacks and unique sequences, or (4) no duplex structures at all. The statistics generated in the table were calculated from the separate data pools. [b] The data in this column pertain to all the long tracer strands scored.

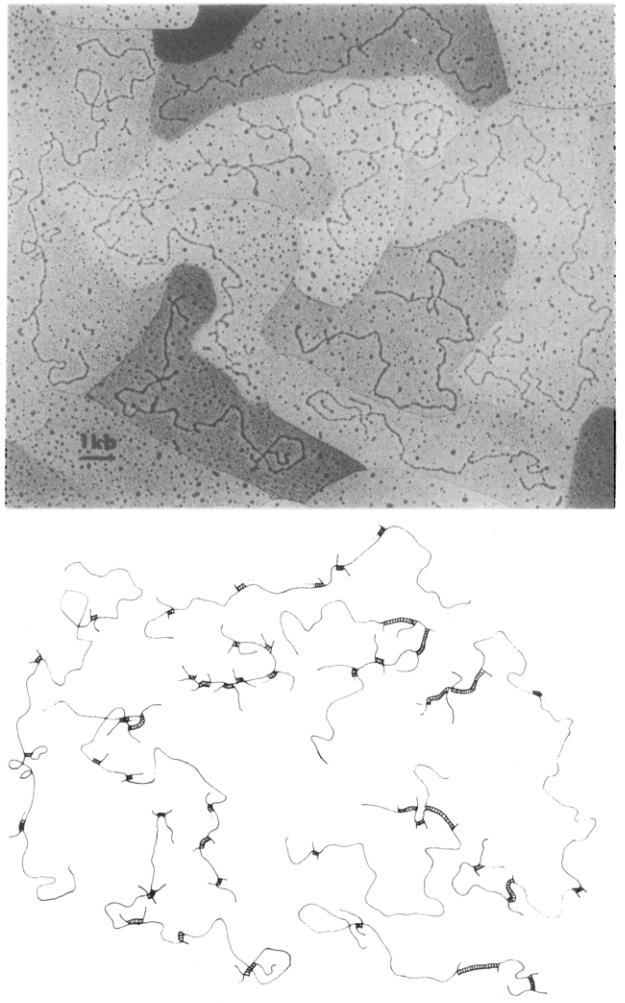

FIGURE 7: Heteroduplex molecules between 20 kbp total DNA and 0.9-kbp repetitive sequence-bearing strands. Long total DNA strands were reacted to C_0t 50 with a 100-fold mass excess of 0.9-kbp repetitive sequence-bearing DNA. Renaturation was at 60 °C in 0.12 M phosphate buffer. The sample was spread for electron microscopy from 50% formamide at room temperature. Long strands were located by scanning the grids at random and scoring any strands substantially greater than the driver molecules. A bar in the composite micrograph shows the length corresponding to 1 kbp. Below are scale line drawings of our interpretations of the structures. Single lines represent DNA single strands; double, cross-hatched lines show duplex DNA. Duplexes without tails are foldbacks. In the case of apparent one-tailed repetitive hybrids a short duplex is drawn ended by a dashed line. This was done for illustrative purposes, and no duplex measurements from one-tailed structures were attempted.

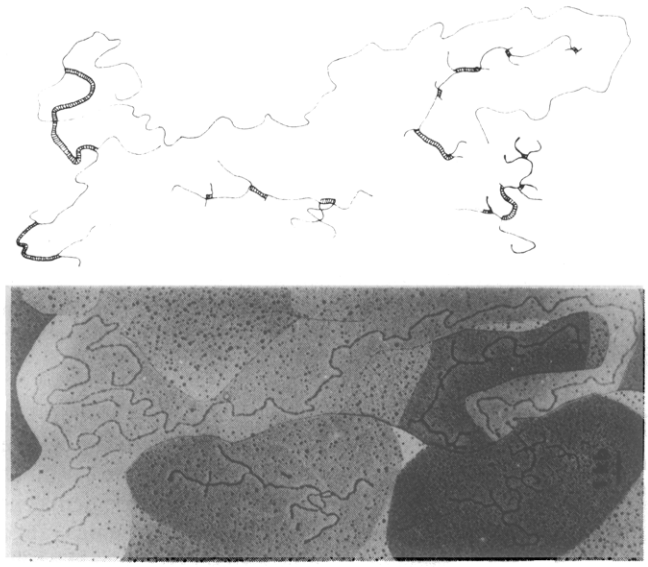

FIGURE 8: Heteroduplex molecules between 20-kb total DNA strands and 2.5-kbp repetitive sequence-bearing DNA, including long and short repetitive duplexes. The 20-kbp total DNA strands were reacted to C_0t 50 with a 120-fold mass excess of 2.5-kbp repetitive DNA. Renaturation and microscopy procedures, as well as conventions used in drawing our interpretations, are as in the legend to Figure 7.

sequences. Both numbers are in good agreement with the findings above derived from analysis of short strand reassociation.

Before we can proceed to the question of the average unique sequence length between repeats, we must establish whether the renaturation of repetitive sequences on the long strands was complete. The presence of an appreciable fraction of unrenatured repeats could result in a drastic overestimate of the actual length of unique sequences between repeats. We apply the calculation of Chamberlin et al. (1975) to measure the recovery of DNA in four-ended H structures. Setting R equal to 0.4 kbp and using the values of L from Table IV, we calculate the expected percent of DNA in duplex for the 0.9-kbp and 2.5-kbp hybrid experiments to be 8.2 and 11.2 on the assumption that 15% of the genome consists of short repetitive sequences. The observed values 10.2 and 12.9 (Table VI) are in agreement with expectation. We conclude that recovery of repetitive duplex structures was essentially complete.

Spacings between Repetitive Sequences. By measuring the single strand contour length from the proximal tail of one

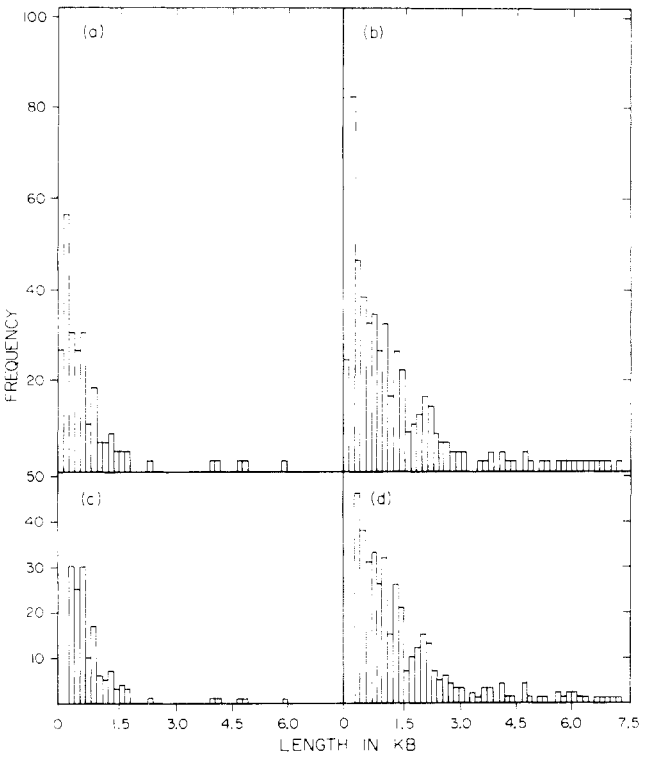

FIGURE 9: Distribution of spacings between repetitive sequences. (a) These data were derived from the molecules of the 0.9-kbp repeat/long DNA experiment, by measuring the single-stranded region bounded by the proximal tails of two two-tails hybrids. (b) As in a for the 2.5-kbp repeat/long DNA hybrids. (c-d) After removing all the spacings less than 0.3 kbp which is only 1.5% of the total DNA for the 0.9- and 2.5-kbp repeats, respectively.

TABLE VI: Recovery of Duplex Structures between Repetitive and Long DNA Strands.

Experiment	No. scored	Total[a] duplex DNA (kbp)	% of[b] long DNA in duplex	No. av	Wt av
Long vs. 0.9-kbp repeat DNA	327	110	10.2	0.336	0.473
Long vs. 2.5-kbp repeat DNA	652	345	12.9	0.529	1.542

[a] This is the sum of the total contour length of repetitive duplex DNA found between two single strand tails; thus, it excludes any contribution from one-tailed structures. [b] The divisor in this computation is the total contour length of all the long tracer strands measured; this number is given in Table V.

TABLE VII: Spacings between Duplexes Terminated by Two Single-Stranded Tails.

Experiment	Class of[a] spacing	No. scored	No. av	Wt av
Long vs. 0.9-kbp repeat DNA	Smaller than 0.3 kbp	82	0.192	0.214
	Larger than 0.3 kbp	148	1.157	5.506
	Total	230	0.813	5.060
Long vs. 2.5-kbp repeat DNA	Smaller than 0.3 kbp	105	0.204	0.223
	Larger than 0.3 kbp	396	1.825	4.921
	Total	501	1.485	4.786

[a] The total spacing measurements were simply sorted into those greater and less than the value of 0.3 kbp to form two separate data pools, which were analyzed independently to yield the numbers cited for the less and greater than 0.3 kbp categories.

two-tailed repeat, we can determine the length of the unique sequence between two repetitive sequences. The spacings measurements from 0.9- and 2.5-kbp DNA experiments exhibit the distributions shown in Figure 9. The spacings range from the smallest observable value (~0.05 kbp) to 30 kbp. In addition 14.8% of the total repetitive sequences appeared to be located in physical contiguity (spacing = 0) to each other in groups of 2 and 3 (see below). The present data cannot establish whether spacings for longer than those observed are present in rat DNA since the average tracer strand lengths in the two experiments were 15 and 20 kbp for the 0.9 and 2.5 kbp data, respectively.

Statistical data pertaining to the spacing measurements are found in Table VII. Comparing the spacing distributions in the 0.9- and 2.5-kbp data (Figure 9), one may observe a close correspondence in the spacing frequencies less than roughly 2 kbp in magnitude. However, in the range of larger spacings (>2 kbp) there are far fewer measurements in the 0.9 kbp than in the 2.5-kbp results. This is reflected in the substantially lower average spacing values in the 0.9-kbp data on Table VII. It may be due to the considerably smaller tracer strand size in this experiment (15 vs. 20 kbp), since the longer spacings have a decreased probability of observation in shorter molecules. The possibility exists that experiments with even longer tracer strands might yield an even larger value for the mean spacing. Nonetheless the distributions shown in Figure 9 indicate that most spacings fall in the range 0.05 to 3 kbp.

In general, we assume any spacing between two-tailed duplexes to be equivalent to the length of that unique sequence. However, since many of the inter-repeat spacings are short (0.05–0.3 kbp), we must consider a second possibility. It could be that the renaturation of a repetitive sequence located between two others, which have already renatured, might be sterically hindered, if the distance between the two previously formed duplexes is short. Thus, while the contour length of DNA contained in all the spacings less than 0.3 kbp, for example, is only 1.5% of the total DNA, replacement of all these apparent spacings with repetitive hybrids would have a very significant effect on the total spacing distributions (Figure 9c-d). Table VII shows that the actual average spacing range would be 1.2–1.8 kbp instead of 0.8–1.5 kbp, if this were the case. Another implication of this alternative interpretation is that the proportion of clustered repeats might be even higher than the 15% observed (see below). The 2.5-kbp data more accurately reflect the total unique sequence distribution than do the 0.9-kbp results for reasons previously mentioned, though they may themselves underestimate the true average spacing. Therefore, regardless of the nature of the short spacings, the average unique sequence in rat DNA is at least in the range of 1.5–1.8 kbp. Using the physical-chemical techniques pioneered by Britten and co-workers (Davidson et al., 1973; Graham et al., 1974), Pearson et al. (1978) have found a value of 1.8–2.5 kbp for the average unique sequence in the rat. The two results are in reasonable accord.

It is noteworthy that an appreciable fraction of the two-tailed repetitive duplexes resided continuously to each other in pairs and infrequently also in triplets. In this regard the 0.9- and 2.5-kbp DNA experiments yielded consistent results. Using the combined data, 12.6% of the total number of duplexes scored resided in doublets and 2.2% in triplets. However,

TABLE VIII: Electron Microscopic Analysis of 29.4-kbp DNA Renatured to Low C_0t.

Class	No. scored	Total DNA in class (kbp)	% of total DNA	No. av length (kbp)	Wt av length (kbp)
Long DNA strands	396	11 600	100	29.421	47.630
Looped foldback duplexes	176	156	2.7	0.884	7.689
Loopless foldback duplexes	150	92	1.6	0.616	10.428
Total foldback duplexes	326	248	4.3	0.740	8.910
Interstrand duplexes	18	65	0.6	3.618	11.383
Loops	176	901	7.8	5.117	15.136
Foldback duplexes in loops	29	26	0.2	0.449	0.737
Loops of foldbacks within foldbacks	12	73	0.6	6.066	17.609
Spacings between foldbacks	152	1 471	12.7	9.679	21.246
Long DNA strands driven by short repeat DNA	41	35.2	0.95	0.42	

these measurements could be the spurious result of partial displacement of a two-tailed repeat by a second incoming driver strand. This hypothesis would predict that the average size of sequences in doublets and triplets would be smaller than that of the total. This is not the case. The mean duplex length of 0.393 kbp among the repeats in multiplets does not differ from the total average.

Foldback Sequences. The rat genome contains a considerable number of foldbacks. We have studied their number, size, and arrangement. We briefly renatured long DNA single strands, mounted them for electron microscopy from 50% formamide, and visualized the hairpin and looped hairpin foldback structures. We can measure the number and size of the foldback duplexes, the spacings between noncontiguous complementary inverted repeat sequences, i.e., the loop lengths, and the spacings between renatured foldback duplexes. Data on the loop-containing foldbacks can also be derived from the experiments in which short repetitive fragments drive long tracer fragments. Unsheared, unfractionated DNA single strands at a concentration of 25 μg/mL were renatured for 1 min, quenched by addition of 4 volumes of ice-cold deionized H_2O, and spread for electron microscopy.

The long single strands displayed numerous hairpin structures, looped and unlooped, as shown in Figure 10. Long strands, 396, of 29.4-kbp number average length were analyzed. Data on foldback structures studied are tabulated in Table VIII. The preponderance of duplex structures consisted of looped and unlooped foldbacks, though a small class of foldback structures formed within the loops of other foldbacks. Also, a small class of apparent intermolecular hybrids was in evidence. The latter structures consisted of four-ended H-type structures. These may have arisen in part from the breakage of foldback loops, as the maximum C_0t of 0.005 reached in this experiment is not favorable for interstrand duplex formation; or they could be due to a small amount of genuinely highly repetitive sequences.

The percent of the total DNA recovered as foldback duplex was 4.5%. The duplexes ranged in size from the smallest measurable to several kbp. The distribution of foldback lengths consists of a numerous class of sequences 0.1–0.6-kbp long and a sparse, disperse class encompassing the remaining foldbacks. If, for the purposes of comparison to hydroxylapatite data, we consider the foldback duplexes respectively less and greater than 2 kbp independently, then the short class averaging 0.3 kbp in length contains 39% of the foldback duplex DNA. The other 61% falls to the long category of 6.1 kb mean length. There are 1.9×10^5 foldback duplexes in the genome. The overall average inverted repeat sequence size is 0.71 kbp.

The distribution of spacings between proximal termini of renatured foldback duplexes appears in Figure 11. Foldbacks

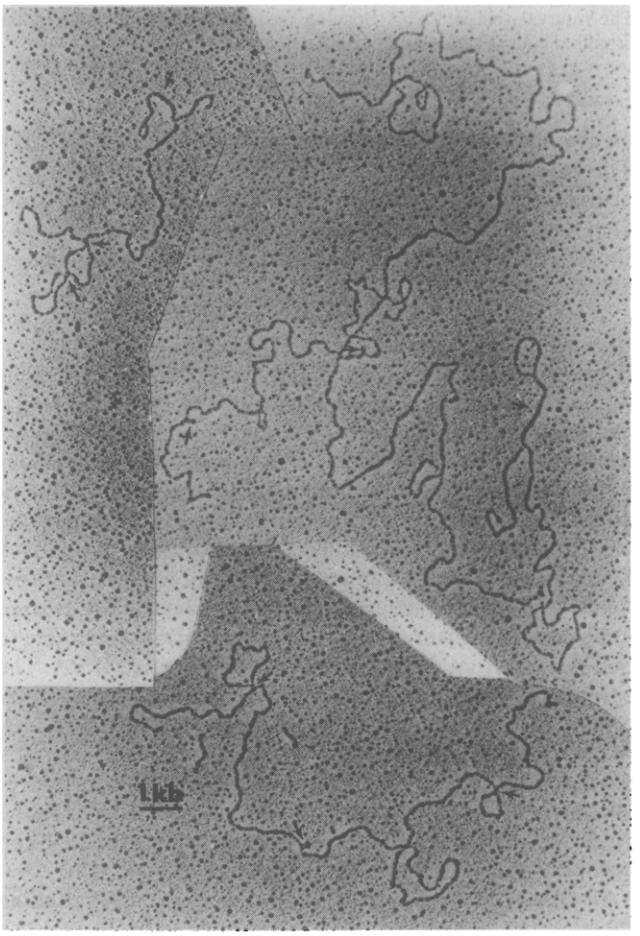

FIGURE 10: Foldback structures among 29.4-kbp DNA single strands. After renaturation as described in the text and quenching with ice-cold H_2O, the DNA was dialyzed against 0.01 M Tris, 0.001 M EDTA (pH 8.5), made 50% in formamide and spread for electron microscopy. ϕX174 circular DNA was present on the grids as a known internal length standard of 5.25 kbp. Arrows on the micrograph point to the hairpin duplexes. A bar indicates the equivalent length of 1 kbp.

are interposed by single strand regions ranging from the very short to greater than 60 kbp. The mean length of the strands among which the spacings were measured, i.e., those bearing at least two foldbacks, was 45.0 kbp. The average inter-foldback spacing is 9.7 kbp (Table VIII); therefore, it is unlikely that a strong bias toward shorter spacings has influenced the data. The relationship between spacing and strand length appears to be random (data not shown).

Fifty-four percent of the foldbacks scored possessed a measurable loop. The looped duplexes are somewhat longer

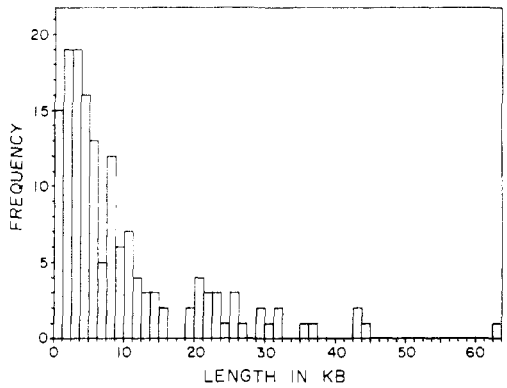

FIGURE 11: Distribution of spacings between foldback duplexes among the 29.4-kbp DNA strands. The frequency is the number of spacings actually scored. Statistics concerning the distribution are given in Table VIII. The interval size is 1 kbp. Spacing measurements were between the proximal ends of two adjacent foldback duplexes.

on average than the simple hairpin structures (Table VIII). The mean loop length is 5.1 kbp (Table VIII). There is no obvious relationship between the foldback duplex length and loop length (data not shown). In 58% of the cases a duplex shorter than 0.5 kbp was terminated by a loop less than 4 kbp.

Hairpin structures in HAP-Fractionated DNA. Further electron microscopic data were gathered from the HAP fractionation products of 29.4-kbp foldback DNA. The strands were renatured as before and fractionated on HAP. The bound and unbound fractions were separately mounted for microscopy and analyzed as before. The bound displayed various types of hairpin structures. The product of the percent foldback duplex in the bound fraction and the fraction of the genome constituted by this fraction, 4.9%, is the amount of foldback DNA recovered. This is in good agreement with the 4.5% estimate from the unfractionated DNA experiment. The average foldback size measured here, 0.61 kbp, is also comparable to that derived from the results with unfractionated DNA. The parameters of spacing and loop length from the bound DNA cannot be compared with the unfractionated DNA results, however, because the bound DNA molecules had suffered a reduction of mean strand length to 7.7 kbp. This length reduction means that these results must tend to deemphasize the longer spacings and loop lengths observed among the 29.4-kbp unfractionated DNA strands.

Comparison of Electron Microscopic to HAP Binding Results. Pearson et al. (1978) have found with S1 nuclease experiments that 6% of rat DNA is duplex by C_0t 0.05. This agrees with the 5% duplex estimate from the electron microscope data. By separating the foldback duplexes shorter and longer than 2–3 kbp with agarose A-50 chromatography, these authors find 42% and 58%, respectively, of the foldback duplex to belong to the shorter and longer categories. This can be compared with the 39–61% distribution determined by EM with the 29.4-kbp unfractionated strands.

Discussion

A Model of the Rat Genome. We propose this model of the rat genome. A numerous class of repetitive sequences 0.40 ± 0.15 kbp long is interspersed among unique sequences at least 1.5–1.8 kbp long throughout at least 90% of the genome. Foldback sequences are interspersed with repetitive and unique sequences throughout at least half the genome. Fifteen percent of the repetitive sequences are organized into doublets and a few triplets of contiguous sequences.

The 4.5% of rat DNA constituted by foldback sequence is divided 39% among a numerous class of sequences an average 0.3 kbp long and 61% among foldbacks at least 6.1 kbp on average. However, the parameters describing the larger foldback class are somewhat uncertain, in that these sequences are few in number. Further, extremely long foldback sequences would appear as linear molecules and go undetected by the present technique. Therefore, the true average size and proportion of the total foldbacks of this larger class may be greater than the figures cited. Foldback pairs are interspersed throughout the genome at an average spacing of 9.7 kbp; the minimum average spacing between noncontiguous complementary inverted repeats is 5.1 kbp. Foldbacks are interspersed between both repetitive and unique sequences in half of the genome (Table V).

From these findings we conclude that rat DNA may contain as little as none and and much as 0.6% highly repetitive DNA.

Comparison with Other Organisms. Davidson et al. (1975) summarize evidence that the *Xenopus* pattern of sequence interspersion appears in organisms found on many of the major branches of the phylogenetic tree. Recently, the genomes of the mollusc (Angerer et al., 1975) and human (Schmid and Deininger, 1975) have also been shown to conform to this pattern. Only the DNA of *Drosophila* (Manning et al., 1975) stands as the notable exception among genomes studies until now. However, Walbot and Dure (1976) have reported that the mean length of interspersed repetitive sequences in cotton is 1.25 kbp—a value intermediate to the short repeats of most animals and the 5-kbp repetitive sequences in *Drosophila*. The organization of the rat genome conforms to this general pattern of sequence arrangement. It is noteworthy, however, that rat DNA contains a much larger portion of contiguous repeats than does *Xenopus*.

Chamberlin et al. (1975) point out that the electron microscopic determination of repetitive sequence length affords a direct measure of the distribution of sequence lengths. In contrast, the physical chemical approach offers ease and an accurate estimate of the overall population of sequences. This comparison is applicable also to the EM and physical-chemical approaches to sequence organization. The former gives a direct measurement of individual repetitive and unique sequence lengths, while the latter provides a good estimate of the mean unique sequence length.

Pearson et al. (1978) used nuclease and hydroxylapatite binding procedures to estimate the short repetitive sequence length as 0.3 kbp and the average inter-repeat spacing as 1.8–2.5 kbp. The agreement between the two measurements of repetitive sequence length in the rat is good.

Implications of Sequence Organization for Gene Regulation. The Britten–Davidson model of gene regulation (Britten and Davidson, 1969; Davidson and Britten, 1973) proposes, in part, that the existence of a few contiguous repeats might provide the basis for coordinate control of certain structural genes. The organization of rat DNA is consistent with this hypothesis.

Acknowledgment

The authors thank Professors Norman Davidson and Eric Davidson for their council and suggestions.

References

Angerer, R. C., Davidson, E. H., and Britten, R. J. (1975), *Cell 6*, 29–39.

Bonner, J., Garrard, W. T., Gottesfeld, J., Holmes, D. S., Sevall, J. S., and Wilkes, M. (1973), *Cold Spring Harbor*

Symp. Quant. Biol. 38, 303-310.
Britten, R. J., and Davidson, E. H. (1969), *Science 165,* 349-357.
Cech, T. R., and Hearst, J. E. (1975), *Cell 5,* 429-446.
Chamberlin, M. E., Britten, R. J., and Davidson, E. H. (1975), *J. Mol. Biol. 96,* 317-333.
Davidson, E. H., and Britten, R. J. (1973), *Q. Rev. Biol. 48,* 565-613.
Davidson, E. H., Galau, G. A., Angerer, R. C., and Britten, R. J. (1975), *Chromosoma 51,* 253-259.
Davidson, E. H., Hough, B. R., Amenson, C. S., and Britten, R. J. (1973), *J. Mol. Biol. 77,* 1-23.
Davis, R. W., Simon, M., and Davidson, N. (1971), *Methods Enzymol. 21D,* 413-428.
Graham, D. E., Neufeld, B. R., Davidson, E. H., and Britten, R. J. (1975), *Cell 1,* 127-137.
Holmes, D. S., and Bonner, J. (1974), *Proc. Natl. Acad. Sci. U.S.A. 71,* 1108-1112.
Manning, J. E., Schmid, C. W., and Davidson, N. (1975), *Cell 4,* 141-155.
Pearson, W. R., Wu, J.-R., and Bonner, J. (1978), *Biochemistry 17* (preceding paper in this issue).
Schmid, C. W., and Deininger, P. L. (1975), *Cell 6,* 345-358.
Schmid, C. W., Manning, J. E., and Davidson, N. (1975), *Cell 5,* 159-172.
Walbot, V., and Dure, L. S., III (1976), *J. Mol. Biol. 101,* 503-536.
Wilson, D. A., and Thomas, C. A. (1974), *J. Mol. Biol. 84,* 115-144.
Wu, J.-R., Hurn, J., and Bonner, J. (1972), *J. Mol. Biol. 64,* 211-219.

Complexity of Nuclear and Polysomal Polyadenylated RNA in a Pluripotent Embryonal Carcinoma Cell Line†

Michel Jacquet,* Nabeel A. Affara, Benoît Robert, Hedwidge Jakob, François Jacob, and François Gros

ABSTRACT: The base-sequence complexities and relative abundance of polysomal and nuclear polyadenylated [poly(A+)] RNA sequences have been analyzed in a pluripotent embryonal carcinoma cell line. Polysomal RNA and nuclear poly(A+) RNA have a complexity representing respectively 0.5% and 2.5% of the single copy component of haploid mouse DNA (1.8×10^6 K base pairs). By hybridization with specific cDNAs, three abundance classes were found in polysomal poly(A+) RNA, representing respectively 31%, 33%, and 36% of the RNA, with base sequence complexities of 0.1×10^3, 0.9×10^3, and 14.5×10^3 kilobases. This corresponds to 7000-8000 different mRNA species of an average length of 2000 nucleotides, present on an average of 5 to 600 copies per cell. In nuclear RNA, a major class of abundance was found with a complexity of 100×10^3 kilobases, each sequence being present in 1 copy per nucleus. The majority of the polysomal poly(A+) RNA sequences are represented in the nuclear poly(A+) RNA but are present in a more restricted range of relative abundance implying posttranscriptional mechanisms of quantitative modulation: polysomal RNA sequences appear to be preferentially transcribed into nuclear cDNA suggesting a preferential location of these sequences close to poly(A) sequences. The presence of a specialized gene product, globin specific RNA, could not be detected either in the nuclear or polysomal compartments of embryonal carcinoma cells, even at levels that would have detected one sequence per 50 cells.

RNA/DNA hybridization has become a useful tool for evaluating the extent of genetic information expressed in different eukaryotic cell types and tissues. By the application of either saturation hybridization of single-copy DNA (Brown and Church, 1971; Gelderman et al., 1971; Hahn and Laird, 1971; Grouse et al., 1972; Galau et al., 1974, 1976) or the analysis of hybridization kinetics between cDNA and its template polyadenylated RNA populations (Birnie et al., 1974; Bishop et al., 1974a), the resulting elucidation of base-sequence complexity and abundance classes in defined RNA populations has begun to furnish quantitative information on the extent of genomic DNA expressed in eukaryotic cells.

The emergent pattern from such studies on a variety of cultured cell types and tissues would imply that base-sequence complexities of mRNA populations (in the range of $8-15 \times 10^3$ different "average-sized" mRNA sequences primarily transcribed from single-copy DNA) represent a very small proportion of the genome coding potential and that most of these RNA sequences are present in a few copies per cell (Galau et al., 1974, 1976; Birnie et al., 1974; Ryffel and McCarthy, 1975; Williams and Penman, 1975; Levy and McCarthy, 1975; Axel et al., 1976; Getz et al., 1976; Young et al., 1976; Hastie and Bishop, 1976; Bantle and Hahn, 1976). In contrast, total or poly(A)-containing nuclear RNA populations have, on the average, a complexity which is five to ten times greater than that of the corresponding mRNA populations, thus implicating some post-transcriptional mechanism controlling the flow of genetic information from the nucleus to the cytoplasm (Getz et al., 1975; Hough et al., 1975; Bantle and Hahn, 1976; Herman et al., 1976; Ryffel and McCarthy, 1975; Levy et al., 1976). It is not clear, however, whether this represents a choice between different protein-coding sequences

† From the Institut Pasteur, Department of Biologie Moléculaire, 25, rue du Docteur Roux, 75015 Paris, France. *Received July 26, 1977.* This work was supported by grants from the Délégation Générale à la Recherche Scientifique et Technique, the Centre National de la Recherche Scientifique, the Institut National de la Santé et de la Recherche Médicale, le Commissariat à l'Energie Atomique, the Ligue Nationale Française contre le Cancer, the Muscular Dystrophy Associations of America, and the National Institutes of Health, Grant number CA 16355. N.A.A. is an EMBO fellow and B.R. is a recipient of a Fondation Roux scholarship.

0006-2960/78/0417-0069$01.00/0 © 1978 American Chemical Society