Volume 12 Number 10 1984 Nucleic Acids Research

Simple sequences are ubiquitous repetitive components of eukaryotic genomes

Diethard Tautz[1]* and Manfred Renz

European Molecular Biology Laboratory, Postfach 10.2209, 6900 Heidelberg, FRG

Received 14 March 1984; Accepted 2 May 1984

ABSTRACT

Simple sequences are stretches of DNA which consist of only one, or a few tandemly repeated nucleotides, for example poly (dA).poly (dT) or poly (dG-dT).poly (dC-dA). These two types of simple sequence have been shown to be repetitive and interspersed in many eukaryotic genomes. Several other types have been found by sequencing eukaryotic DNA. In this report we have undertaken a systematical survey for simple sequences. We hybridized synthetical simple sequence DNA to genome blots of phylogenetically different organisms. We found that many, probably even all possible types of simple sequence are repetitive components of eukaryotic genomes. We propose therefore that they arise by common mechanisms namely slippage replication and unequal crossover and that they might have no general function with regards to gene expression. This latter inference is supported by the fact that we have detected simple sequences only in the metabolically inactive micronucleus of the protozoan Stylonychia, but not in the metabolically active macronucleus which is derived from the micronucleus by chromosome diminution.

INTRODUCTION

Short stretches of simple sequences (mostly less than 100 bp long) have frequently been found by chance in regions of sequenced DNA. They occur near genes (1,2) in some introns of genes (3-5), in the spacers of the histone gene cluster in sea urchin (6,7) and Drosophila (8) as well as in the DNA regions between immunoglobulin genes (9,10). They were further found also within variants of the repetitive Alu-elements (11), within satellite sequences (12), as well as in other regions of the genome which cannot be related to any function (13-15). Two types of simple sequences, namely AA/TT* (16-19) and GT/CA (20) have been shown by means of hybridization to be interspersed repetitive elements in eukaryotic DNA.

Simple sequences are distinctly different from simple satellite sequences in that they are interspersed in the genome (19-21) and are usually transcribed into RNA (15,21). Additionally, different types

© IRL Press Limited, Oxford, England.

of simple sequence can be clustered within a small region of DNA (5,13-15).

Especially the simple sequence GT/CA has attracted some interest, because it is an alternating purine/pyrimidine sequence which has the potential to form Z-DNA (20,22,23). But also polypyrimidine sequences, which may be simple sequences (24) have been shown to be interspersed repetitive components of eukaryotic genomes (25). We therefore undertook a systematical search for simple sequences in eukaryotes.

MATERIALS & METHODS

Simple sequence probes

The simple polynucleotide probes poly (dG-dT).poly(dC-dA), poly(dG).poly(dC), and poly(dA).poly(dT) were purchased from Boehringer (Mannheim), poly(dG-dA).poly(dC-dT) was polymerized under the conditions given in (26) using the respective decamers (from P.L. Biochemicals, Wisconsin) as primers. All probes were labelled by nick-translation to a specific activity between 8.10^7 to 2.10^8 cpm/µg. At least two ^{32}p-labelled nucleotides were used in order to get both strands of the simple sequence DNA labelled. The restriction fragment containing the simple sequence CAG/GTC was isolated from an agarose gel (27) and also nick-translated.

Hybridization

Filters were generally preincubated in 5 x SSC/0.5% SDS/2 x Denhardt's solution (28) for 3-4 hours at 65°C; the hybridization solution contained SSC (1 x SSC = 0.15M NaCl, 0.015M Na-citrate) and Formamide as specified below, 0.5% SDS, 2 x Denhardt's solution and in some experiments 50 µg/ml poly-A in order to reduce background; the radioactive hybridization probes were taken up in 90% Formamide, heat denatured and added before any renaturation could occur; the filters were hybridized overnight, washed in the hybridization solution without Denhardt's solution at the hybridization temperature and then four times in 2 x SSC at room temperature.

The hybridization conditions for the different simple sequences were as follows: for GA/CT and GT/CA:50% Formamide/5 x SSC/37°C; for GG/CC: 50% Formamide/1 x SSC/45°C; for AA/TT: 5% Formamide/5 x SSC/37°C; for the restriction fragment bearing the CAG/GTC-element:50% Formamide/5 x SSC/42°C.

DNA-Isolation and electrophoresis

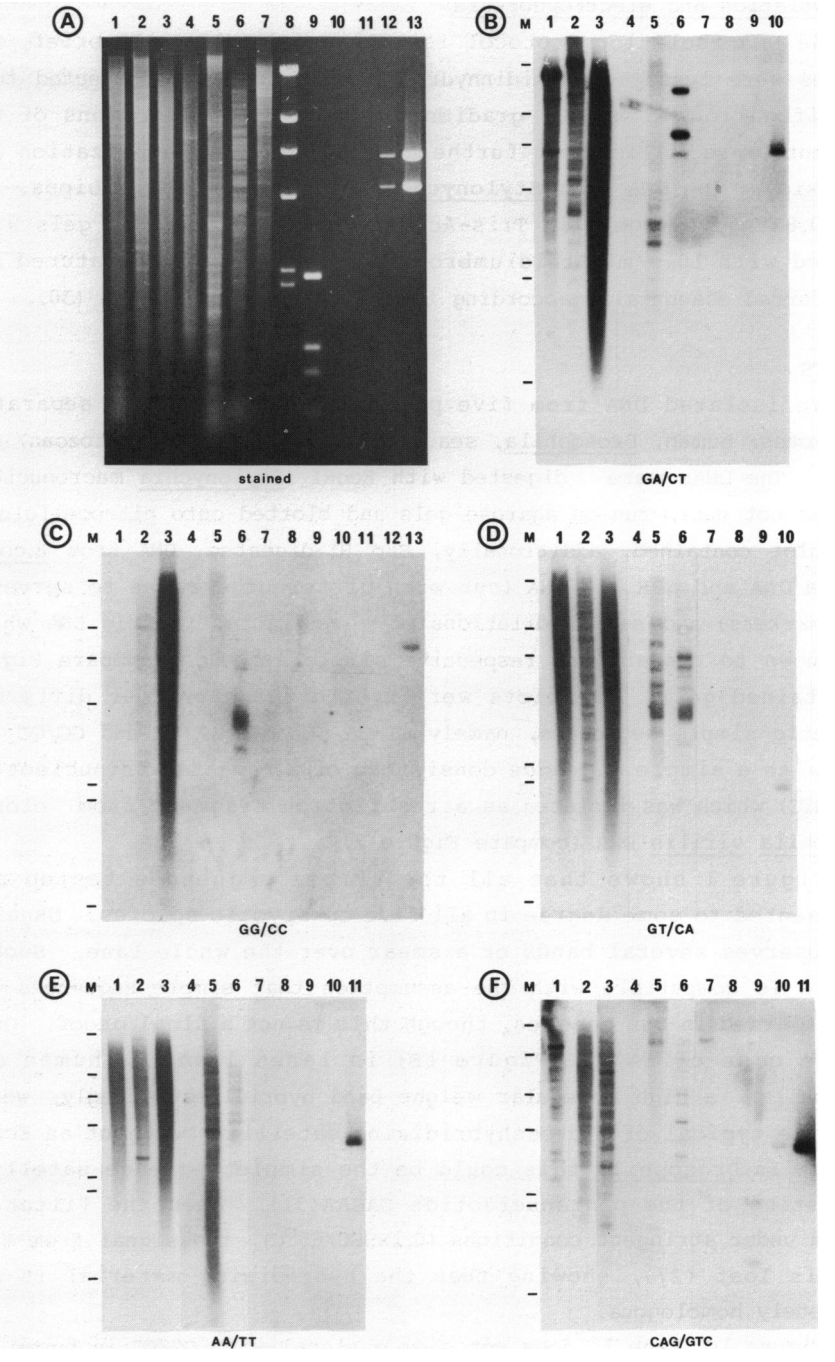
The DNA isolation protocol is outlined in (15). In brief, the tissues were lysed in Guanidinhydrochlorid-buffer and subjected to a centrifugation in a CsCl gradient. The viscous fractions of the gradients were utilized for further purification (phenolization and dialysis). The DNA from Stylonychia was a gift from H.J.Lipps. We used 0.8% agarose gels in Tris-Acetate buffer (29). The gels were stained with 10µg/ml Ethidiumbromide, photographed, denatured and transferred essentially according to the method of Southern (30).

RESULTS

We isolated DNA from five phylogenetically widely separated eukaryotes: human, Drosophila, sea urchin, Stylonychia (protozoan) and yeast. The DNAs were digested with EcoRI (Stylonychia macronuclear DNA was not cut), run on agarose gels and blotted onto nitrocellulose. Each blot contained, additionally, Bam HI-digested DNA from E.coli, lambda DNA and pBR 322 DNA (cut with different enzymes to serve as size markers) and serial dilutions of a restricted plasmid DNA which was known to contain the respective simple sequence (compare Figure 1A, stained gel). The blots were hybridized with four different synthetic simple sequences, namely GT/CA, GA/CT, AA/TT and GG/CC and also with a simple sequence consisting of a repeated trinucleotide (CAG/GTC) which was isolated as a restriction fragment from cloned Drosophila virilis DNA (compare Figure 2).

Figure 1 shows that all the simple sequences tested are represented to some degree in all five eukaryotic species. Usually one observes several bands or a smear over the whole lane. Such a picture is compatible with the assumption that simple sequences are interspersed in the genomes, though this is not a final proof. Only in the case of GA/CT (Figure 1B) in lanes 1 and 2 (human and Drosophila), a high molecular weight band hybridizes strongly, which would be typical of a crosshybridizing satellite (without an EcoRI site). In Drosophila this could be the simple sequence satellite consisting of the pentanucleotide GAGAA(31). When the filter is washed under stringent conditions (0.1xSSC/65°C), the signal from this band is lost (27), showing that the hybridizing material is not completely homologous.

Figure 1C, lane 1 does not show a signal with GG/CC in human DNA.


A
1 2 3 4 5 6 7 8 9 10 11 12 13
stained
B
M 1 2 3 4 5 6 7 8 9 10
GA/CT
C
M 1 2 3 4 5 6 7 8 9 10 11 12 13
GG/CC
D
M 1 2 3 4 5 6 7 8 9 10
GT/CA
E
M 1 2 3 4 5 6 7 8 9 10 11
AA/TT
F
M 1 2 3 4 5 6 7 8 9 10 11
CAG/GTC

Further exposure of the filter is not possible because GG/CC exhibits a high tendency to stick to nitrocellulose by itself, yielding a high background. But using a dot blot assay, however, we did detect a weak signal (15), GG/CC tracts have been also detected in mammalian genomes before (17).

We found no hybridization with the DNA from E.coli, lambda or pBR 322, except for the experiment with the restriction fragment (containing the simple sequence CAG/GTC). The signal from pBR 322 in the CAG/GTC hybridization is doubtless due to a contamination of the isolated fragment with vector sequences. The signal from E.coli might have the same origin although it looks quite specific.

The hybridization conditions employed (see Methods) were determined such that runs of simple sequences shorter than 15 bp

Figure 1 Hybridization of simple sequences to genomic blots from different organisms (note the different exposure times of the lanes). 1 µg of each of the genomic DNAs was applied. Lanes 1: human, lanes 2: Drosophila, lanes 3: sea urchin, lanes 4: Stylonychia (macronucleus), lanes 5: Stylonychia (micronucleus), lanes 6: yeast, lanes 7: E.coli, cut with Bam HI, lanes 8: lambda, cut with Hind III, (fragment sizes: 23.15 kb, 9.42 kb, 6.56 kb, 4.38 kb, 2.32 kb, 2.02 kb and 0.56 kb), lanes 9: pBR 322, cut with EcoRI, Bam HI, Ava I and Pst I, (fragment sizes: 2.2 kb, 1.05 kb, 0.75 kb and 0.375 kb), lanes 10-13: serial 1 to 25 dilutions of a restricted plasmid (specified below) starting with 0.5 µg in lanes 13, 20 ng in lanes 12, 0.8 ng in lanes 11 and 30 pg in lanes 10. The bars at the left side in B - F indicate the positions of the lambda Hind III marker fragments.

A: Agarosegel, stained with ethidiumbromide.

B: Blot from the gel in A, hybridized with the simple sequence GA/CT. The reference plasmid (lane 10) is pDv-1, cut with Bam HI (compare Fig.2). Exposure times: lanes 1-3 for 15 hours at room temperature, lanes 4-10 for 40 hours at -70°C, with intensifying screen.

C: Blot hybridized with the simple sequence GG/CC. The reference plasmid is 191-6, cut with Bam HI (compare Fig.2). Exposure times: all lanes 16 hours at -70°C with intensifying screen.

D: Blot hybridized with the simple sequence GT/CA. The reference plasmid is pDv-161, cut with Bgl I (compare Fig.2). Exposure times: lanes 1-3 and lane 6 for 15 hours at room temperature, lanes 4, 5 and lanes 7-10 for 64 hours at -70°C with intensifying screen.

E: Blot hybridized with the simple sequence AA/TT. The reference plasmid is pDv-161 cut with Bgl I (compare Fig.2). Exposure times: lanes 1-3 and lane 6 for 15 hours at room temperature, lanes 4, 5 and lanes 7-11 for 7 days at -70°C with intensifying screen.

F: Blot hybridized with the restriction fragment from pDV-161 (indicated in Fig.2, part 5) containing the simple sequence CAG/GTC. The reference plasmid is pDv-161 cut with Bgl 1 (compare Fig.2). Exposure times: lanes 2 and 6 for 15 hours at room temperature, lane 1, lanes 3-5 and lanes 7-11 for four days at -70°C with intensifying screen.

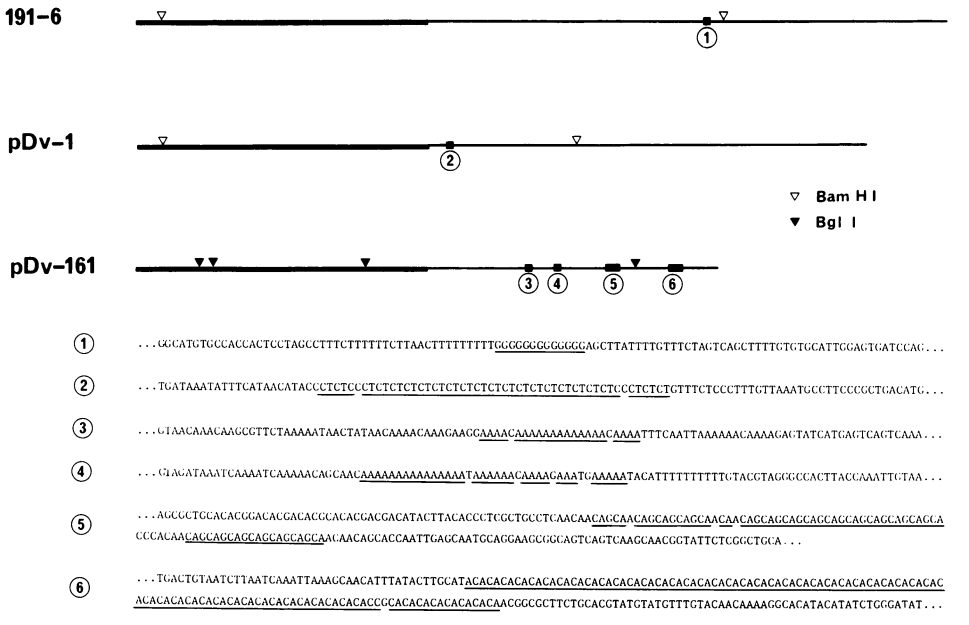


Figure 2 Reference clones used for the hybridization experiments shown in Fig.1. The relative locations for those simple sequences which are relevant for the hybridization experiments are indicated by the numbered black boxes in each clone. These sequences are shown underneath. All three clones additionally contain other simple sequences which are not shown. The vector (pBR 322) is drawn by a bold line. Clone 191-6 is a Hind III fragment from clone 2.14 which contains the H-2k^d gene of mouse (5) cloned in the Hind III site of pBR 322. The clone yields two Bam HI fragments with a size of 8.5 kb and 3.7 kb respectively. Clones pDv-1 and pDv-161 originated from a D.virilis microlibrary and are described in more detail in (15) pDv-1 yields two Bam HI fragments with sizes of 6.2 kb and 4.8 kb. pDv-161 yields four Bgl I fragments with sizes of 4.7 kb, 2.3 kb, 2.0 kb and 0.3 kb. Part 5 from this clone shows the sequence of the entire restriction fragment (flanked by a Hae II and a Pst I site) which was used for the CAG/GTC hybridization experiments.

should give no signal, because their melting point would be below the hybridization temperature. To calculate this, we used empirically determined formulas which are summarized in (32). However, the tract of only 13 G-residues in clone 191-6 gave an extremely weak signal with GG/CC as probe (Figure 1C, lane 13; note that 0.5µg of plasmid DNA was loaded in this lane). On the other hand, this clone contains some G and C rich, but not homopolymeric regions also on the second Bam HI-fragment (5), which do not yield a signal. This shows that the hybridization is still specific.

Simple sequences up to a length of about 20-25 bp yield an disproportionally low signal under the employed conditions. This is seen in Figure 1E, lane 11. The hybridizing restriction fragment here contains only comparatively short tracts of uninterrupted A-residues (compare Figure 2) which show a hybridization signal 10-20 times lower than the longer GA/CT or GT/CA-stretches (Figure 1B and 1D, lanes 10; note that 25 times more DNA is loaded in lanes 11 than lanes 10). The same effect was observed when the hybridization efficiency of short or interrupted tracts of GT/CA and GA/CT was compared with the signal from long tracts (21 and unpublished results).

Finally, under the conditions employed, we never observed a cross hybridization between different simple sequences.

Considering all this, we feel sure that any signal observed in the genomic blots must originate from a sufficiently long (> 25 bp) tract of DNA which is highly homologous to the probes, or, if less homologous, is from a highly repeated DNA, due to the summation of weak signals.

In additional experiments we also observed some signals from the simple sequences GC/CG and AT/TA on dot blots (15) but also partly on genomic blots. However, one has to keep in mind that these probes are self-complementary and can not be compared with usual hybridization data.

It seems that a certain correlation exists between the number of positive bands (or extent of smear) and the kinetic complexity of the respective genome. However this can not be taken as a general rule, in that the exceptions are too obvious; for example in the GG/CC hybridization (Figure 1C) or the general low simple sequence content in Stylonychia (Figure 1, B-F, lanes 5) or the high content of CAG/GTC in Drosophila and yeast (Figure 1F, lanes 2 and 6) (note the different exposure times of several lanes).

A particularly interesting detail in Figure 1 is the comparison of the micronuclear and macronuclear DNA of the Stylonychia genome. The macronuclear DNA in this ciliate protozoan is derived from the micronuclear DNA by chromosome diminution (33) and contains only 5%-10% of the complexity of the micronucleus (33,34) but nonetheless retains all the sequences needed for normal cell metabolism and vegetative growth (33). Comparison of lanes 4 and 5 in Figure 1 shows that simple sequences are obviously present only in the micronucleus

Table 1 Estimation of the relative portions of simple sequences in different eukaryotic genomes. The figures given are only rough estimates. Especially the values in the GG/CC column were estimated without a good control (see text).

	genome size [bp]	GT/CA [%]	GA/CT [%]	AA/TT [%]	GG/CC [%]	CAG/GTC [%]
man	3.3×10^9	0.5	0.2	0.3	0.0002	0.005
Drosophila	1.4×10^8	0.2	0.1	0.2	0.01	0.8
sea urchin	8.6×10^8	0.8	0.8	0.7	0.2	0.01
Stylonychia	6.2×10^8	0.01	0.002	0.02	0.005	0.001
yeast	4.0×10^7	0.1	0.006	0.1	0.02	0.1

(lanes 5), which gives a strong indication that they are not usually needed for gene expression in the normal metabolic cell functions. (The signal in Figure 1B, lane 4 is a background spot).

This observation is also interesting with respect to a second point. The simple sequence GT/CA has been shown to form left-handed Z-DNA in vitro (22,23) and has therefore been proposed to be the major Z-DNA forming component in eukaryotes (20). Z-DNA has been detected in Stylonychia, but only in the macronucleus, not in the micronucleus (35) just the reverse of the distribution of the GT/CA sequences. This result shows that GT/CA sequences do not necessarily acquire Z-DNA conformation in vivo or under the fixation conditions employed in the experiment (35), but that other, or very short sequences must account for Z-DNA formation in Stylonychia, and possibly also elsewhere.

We have tried to estimate the amounts of the individual simple sequence variants in the different genomes. This was done by comparing the hybridization signals from a cloned DNA with a known simple sequence content, with those from the genomic DNAs in a dot blot assay (36,21). But for the reasons outlined above (nonlinear relationship between the length of the hybridizing DNA and the corresponding signal) we think that the results summarized in Table 1 are only very rough estimates. They may vary by an order of magnitude. Nonetheless it shows that the several simple sequence variants may add up to a few percent and thus should be considered to be a substantial ccomponent of eukaryotic DNA.

DISCUSSION

We have shown that all possible types of simple sequences which are composed of only one nucleotide or two alternating nucleotides are present to at least some extent in eukaryotic genomes (for GC/CG and AT/TA we assume that they exist, but that they can not be readily traced by hybridization). We demonstrated that the same is true for a simple sequence which is composed of a repeated trinucleotide (CAG/GTC). Others have shown that even higher ordered simple sequences are shared by diverse organisms (10,37). Several suggestions have been made concerning a possible function of simple sequences, for example, in chromatin-folding (38) homogenisation of repetitive gene arrays (39), as hot spots for recombination (3,40), in the evolution of new genes (41), in telomere formation (40,42) and in gene regulation (43). All these proposals are concerned only with certain types of simple sequence. But the very fact that probably all types of simple sequences exist, suggests that they all might arise by the same mechanisms and not because they have a primary function. It is reasonable to assume that these mechanims are slippage reactions and unequal crossovers which take place at randomly occurring short runs of these sequences. Both mechanisms would lead to a constant formation and deletion of simple sequences and one would expect to find them in all regions of the genome which do not undergo strong selection. Hence, the occurrence of simple sequences in eukaryotes is not a matter of evolutionary conservation, but instead depends on a number of factors including (i) the frequency of accidental amplifications and deletions, (ii) the extent to which such mechanisms spread the sequences between homologous chromosomes (44), (iii) the degree to which the sequences are tolerated in the genome and (iv) on the amount of possible formation sites for simple sequences, namely redundant DNA. The absence of large amounts of simple sequences in prokaryotes could be due to any one of these factors, singly or in combination.

It is possible that with respect to some simple sequences there are additional mechanisms by which they arise and are distributed in the genome. For example AA/TT may equally well arise by reverse transcription of poly-A tails of mRNA and reintegration into the genome (45). An integration mechanism has been proposed also for GT/CA, because some of the stretches are flanked by short direct

repeats (40). However, we should like to emphasize that subsequent slippage and unequal crossover must be expected to occur in all simple sequence regions regardless of the actual mode of origin.

A general, indirect function of simple sequences is most probably that they serve as hot spots of recombination, as has been shown in certain cases (3,46,47). This is supported by the fact that simple sequences may easily form single stranded regions (48), which is probably due to slippage. These might serve as hot spots for strand invasion during initiation of the recombination event. They might also be able to combine different chromosome regions which otherwise share no homology, a mechanism which has been proposed for the switching region of immunglobulin genes (46). Simple sequences should therefore be regarded as a source of naturally occurring rearrangements and variation.

ACKNOWLEDGEMENTS

We thank H.J. Lipps for the gifts of micronuclear and macronuclear DNA from Stylonychia and G.A. Dover for discussion and helpful suggestions on the manuscript. We thank June Hunt for her patience with preparing the manuscript.

[1]Present address: University of Cambridge, Department of Genetics, Downing Street, Cambridge CB2 3EH, UK

*To whom reprint requests should be sent

*ABBREVIATIONS

We write homopolymeric sequences as two nucleotides from both strands, for example AA/TT instead of poly (dA). poly (dT). For higher ordered simple sequences we write the repeat unit from both strands, for example GT/CA instead of poly (dG-dT). poly (dC-dA).

REFERENCES

1. Nishioka, Y. and Leder, P. (1980). J. Biol. Chem. 255, 3691-3694.
2. Miesfeld, R., Krystal, M. and Arnheim, N., (1981). Nucl. Acids Res. 9, 5931-5947.
3. Slightom, J.L., Blechl, A.E. and Smithies, O. (1980). Cell 21 627-638.
4. Hamada, H., Petrino, M.G. and Kakunaga, T. (1982). Proc. Natl. Acad. Sci. U.S.A. 79, 5901-5905.
5. Kvist, S., Roberts, L. and Dobberstein, B. (1983). EMBO Journal 2, 245-254.
6. Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H.O. and Birnstiel, M.L. (1978). Cell 14, 655-671.

7. Sures, I., Lowry,^U J. and Kedes, L.H., (1978). Cell 15, 1033-1044.
8. Goldberg, M., (1979). Ph.D. dissertation, Stanford University, U.S.A.
9. Richards, J.E., Gilliam, A.C., Shen, A., Tucker, P.W. and Blattner, F.R. (1983). Nature (London) 306, 483-487.
10. Gebhard, W. and Zachau, H.G., (1983). J. Mol. Biol. 170, 567-573.
11. Saffer, J.D. and Lerman, M.I., (1983). Mol. Cell Biol. 3, 960-964.
12. Skinner, D.M., Bonewell, V. and Fowler, F.F. (1982). Cold Spring Harbor Symp. Quant. Biol. 47, 1151-1157.
13. Shen, S. and Smithies, O. (1982). Nucl. Acids Res. 10 7809-7818.
14. Höchtl, J. and Zachau, H.G. (1983). Nature (London) 302, 260-263.
15. Tautz, D. and Renz, M. (1984). J. Mol. biol. 172, 229-235.
16. Shenkin, A. and Burdon, R.H., (1972). FEBS Letters 22, 157-160.
17. Shenkin, A. and Burdon, R.H., (1974). J. Mol. Biol. 85, 19-39.
18. Bishop, J.O., Rosbash, M. and Evans, D., (1974). J. Mol. Biol. 85, 75-86.
19. Flavell, R.A., van den Berg, F.M. and Grosveld, G.C., (1977). J. Mol. Biol. 115, 715-741.
20. Hamada, H., Petrino, M.G. and Kakunaga, T. (1982). Proc. Natl. Acad. Sci. U.S.A. 79, 6465-6469.
21. Tautz, D. (1983). Ph.D. thesis, University of Tübingen, FRG.
22. Arnott, S., Chandrasekaran, R., Birdsall, D.L., Leslie, A.G.W. and Ratliff, R.L., (1980). Nature (London) 283, 743-745.
23. McIntosh, L.P., Grieger, I., Eckstein, F., Zarling, D.A., van de Sande, J.H. and Jovin, T.M. (1983). Nature (London), 83-86.
24. Deugau, K.V., Mitchel, R.E.J. and Birnboim, H.C., (1983). Analyt. Biochem. 129, 88-97.
25. Straus, N.A. and Birnboim, H.C. (1976). Biochim. Biophys. Acta 454, 419-428.
26. Morgan, A.R., Coulter, M.B., Flintoff, W.F. and Paetkau, V.H., (1974). Biochemistry 13, 1596-1603.
27. Tautz, D. and Renz, M. (1983). Analyt. Biochem. 132, 14-19.
28. Denhardt, D.T., (1966). Biochem. Biophys. Res. Commun. 23, 641-646.
29. Loening, U.E., (1967). Biochem. J. 102, 251-257.
30. Southern, E. (1975). J. Mol. Biol. 98, 503-517.
31. Sederoff, R., Lowenstein, L. & Birnboim, H.C (1975). Cell 5, 183-194.
32. Chow, L.T. and Broker, T.R., (1981). Electron microscopy in Biology, Vol.1 (John Wiley & Sons, New York), 139-188.
33. Ammermann, D., Steinbrück, G., von Berger, L. and Henning, W., (1974). Chromosoma 45, 401-429.
34. Lauth, M.R., Spear, B.B., Heumann, J. and Prescott, D.M. (1976). Cell 7, 67-74.
35. Lipps, H.J., Nordheim, A., Lafer, E.M., Ammermann, D., Stollar, B.D. and Rich, A. (1983). Cell 32, 435-441.
36. Kafatos, F.C., Jones, C.W. and Efstratiadis, A. (1979). Nucl. Acids Res. 7, 1541-1552.
37. Epplen, J.T., McCarrey, J.R., Sutou, S. and Ohno, S., (1982). Proc. Nat. Acad. Sci. U.S.A. 79, 3798-3802.
38. Johnson, D. and Morgan, R., (1978). Proc. Natl. Acad. Sci. U.S.A. 75, 1637-1641.
39. Kedes, L.H., (1979) An. Rev. Biochem. 48, 837-870.
40. Rogers, J. (1983). Nature (London), News & Views 305, 101-102.

41. Ohno, S. and Epplen, J.P., (1983). Proc. Nat. Acad. Sci. U.S.A. 80, 3391-3395.
42. Walmsley, R.M., Szostak, J.W. and Petes, T.D. (1983). Nature (London) 302, 84-86.
43. Russel, D.W., Smith, M., Cox, D., Williamson, V.M. and Young, E.T., (1983). Nature (London) 304, 652-654.
44. Dover, G.A. (1982). Nature (London) 299, 111-117.
45. Sharp, P.A., (1983). Nature (London) 301, 471-472.
46. Nikaido, T., Nakai, S. and Honjo, T. (1981). Nature (London) 292, 845-848.
47. Stringer, J.R. (1982). Nature (London) 296, 363-366.
48. Hentschel, E.C., (1982). Nature (London) 295, 714-716.