Volume 15 Number 5 1987 Nucleic Acids Research

A new moderately repetitive DNA sequence family of novel organization

N.D.Epstein*, S.Karlsson, S.O'Brien[1], W.Modi[1], A.Moulton and A.W.Nienhuis

Clinical Hematology Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892 and [1]Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick Cancer Research, Frederick, MD, USA

Received September 9, 1986; Revised and Accepted December 24, 1986

ABSTRACT

In cloning adenovirus homologous sequences, from a human cosmid library, we identified a moderately repetitive DNA sequence family consisting of tandem arrays of 2.5 kb members. A member was sequenced and several non-adjacent, 15-20 bp G-C rich segments with homology to the left side of adenovirus were discovered. The copy number of 400 members is highly conserved among humans. Southern blots of partial digests of human DNA have verified the tandem array of the sequence family. The chromosomal location was defined by somatic cell genetics and in situ hybridization. Tandem arrays are found only on chromosomes 4 (4q31) and 19 (q13.1-q13.3). Homologous repetitive sequences are found in DNA of other primates but not in cat or mouse. Thus we have identified a new family of moderately repetitive DNA sequences, unique because of its organization in clustered tandem arrays, its length, its chromosomal location, and its lack of homology to other moderately repetitive sequence families.

INTRODUCTION

The homology between retroviral oncogenes and essential cellular proto-oncogenes has attracted considerable interest. There is little evidence that important elements of DNA viruses have corresponding cellular homologues. Prior studies had suggested no homology or a diffuse pattern of hybridization when Southern blots of human DNA were probed with adenovirus. (1,2) Recently Karlsson et al. reported discrete human DNA fragments having homology to adenovirus (3). We have cloned one of these fragments and found that it is part of a new moderately repetitive sequence family. Its hybridization to adenovirus DNA proved to be based on 15-20 nucleotide highly G-C rich segments, present in high copy number at a single position on the Southern blot.

Moderately repetitive DNA sequences comprise approximately 30% of the human genome. (4) Most such sequences are organized into families the individual members of which are widely dispersed within single copy DNA (see ref. 5 and 6 for reviews) whereas certain repetitive sequences are clustered. The Alu and Kpn I families represent the prototypes of short interspersed sequences (SINES) and long interspersed sequences (LINES), respectively. Individual Alu members are approximately 300 bps whereas Kpn I members vary up to as much as 6100 bps (7,8). Dispersed repetitive DNA sequences are not conserved during evolution but in a single species, indi-

vidual family members show high homology suggesting some mechanisms for homogenization. RNA homologous to dispersed moderately repetitive DNA sequences is present in most or all cells, but the function of these families is unknown. [9-15] In contrast, clustered moderately repetitive sequences are highly conserved during evolution and encode essential cellular products, eg. ribosomal RNAs, [16-18] small nuclear RNAs, [19-20] and histones. [21-23] The family we have discovered occurs in clustered arrays but is not conserved during evolution and therefore may be distinguished from previously described moderately repetitive DNA sequences.

MATERIALS AND METHODS

Materials

Restriction endonucleases and sequencing kits were obtained from Bethesda Research Laboratories, [^{32}P] dCTP, {^{32}P}dGTP and {^{35}S}dATP from Amersham. Deoxynucleotides were from Pharmacia Inc. and Bal 31 was obtained from International Biotechnologies Inc.

Southern Blot

Analysis was by standard techniques. (24) Hybridization was done in 3X SSC at 65° C. Four washes, each thirty minutes, were in 0.1XSSC at 55° C.

Cloning of the Adenovirus Homologous Sequences

Approximately 400,000 colonies of a human cosmid library [25] were screened using a modification of techniques of Grosveld et al. [26] and Woods. [27] Twenty ml of the library diluted to a concentration equivalent to 2000 colonies/ml were filtered onto each nitrocellulose filter through a modified Buchner Funnel. Replica filters were made and processed as described except that after lysis with NaOH they were allowed to dry for 5 minutes before neutralization with Tris/NaCl. Hybridization conditions were the same as in Southern blot analysis.

Sequencing

An isolated SstI repeat member initially cloned into pUC 12 was sequenced, using a modification of the method described by Poncz et al. [28] to construct a set of nested fragments. These were recloned into M13 and sequenced by the method of Sanger. [29] A BRL sequencing kit was used substituting {^{35}S}dATP for {^{35}p}dATP. The reaction was run on an ionic gradient acrylamide gel as suggest by the manufacturer.

In Situ Hybridization

The procedure followed was essentially that of Harper and Saunders. (30)

Statistical Analysis

Triplicate dot blots prepared with DNA samples from eleven subjects with diverse ethnic background were probed with isolated Alu, KpnI, or SstI repeat members. The spots were cut out and counted in a scintillation counter. The respective Dpms were converted to logarithms and the deviations from the expected value were calculated. An analysis of variance was performed; the source of variation estimated to be interaction using Tukey's procedure [31] was 0.001 (1df.) compared to a residual mean square of 0.01.

RESULTS

Human DNA Contains Adenovirus Homologous Sequences in Several Fragments

Human DNA derived from tissues or cell lines (K562, HL-60, HEL, and HeLa), when digested with Sst I and probed with adenovirus, yielded discrete bands of 12, 8.5, and 2.5 kilobases (kb) (Figure 1a). One individual appeared to be homozygous for a restriction fragment length polymorphism in that the 12kb band was replaced by one of 9.5 kb. The possibility that the DNA samples were contaminated with adenovirus DNA was considered but seemed highly unlikely based on this apparent polymorphism in one sample. The adenovirus genome was resolved into three fragments by digestion with Sal I followed by agarose gel electrophroesis. The 5′ 9 kb segment annealed to the 2.5 kb band (Figure 1B), the middle 7 kb segment did not reveal discrete bands and the 3′ 20 kb annealed to the 12 and 8.5 kb bands (data not shown). The 5′ fragment contains the coding sequences for proteins involved in DNA replication, transcriptional control,and cell transformation and therefore we chose to clone the sequences homologous to the fragment.

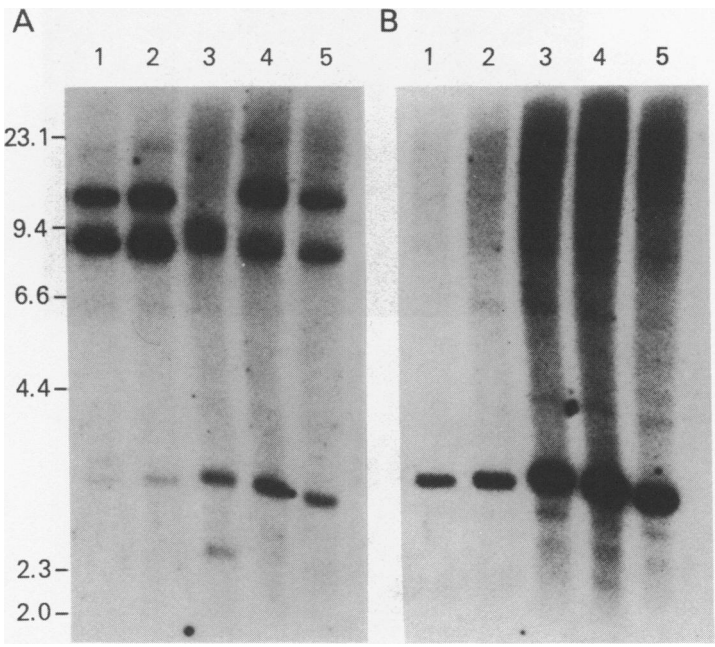
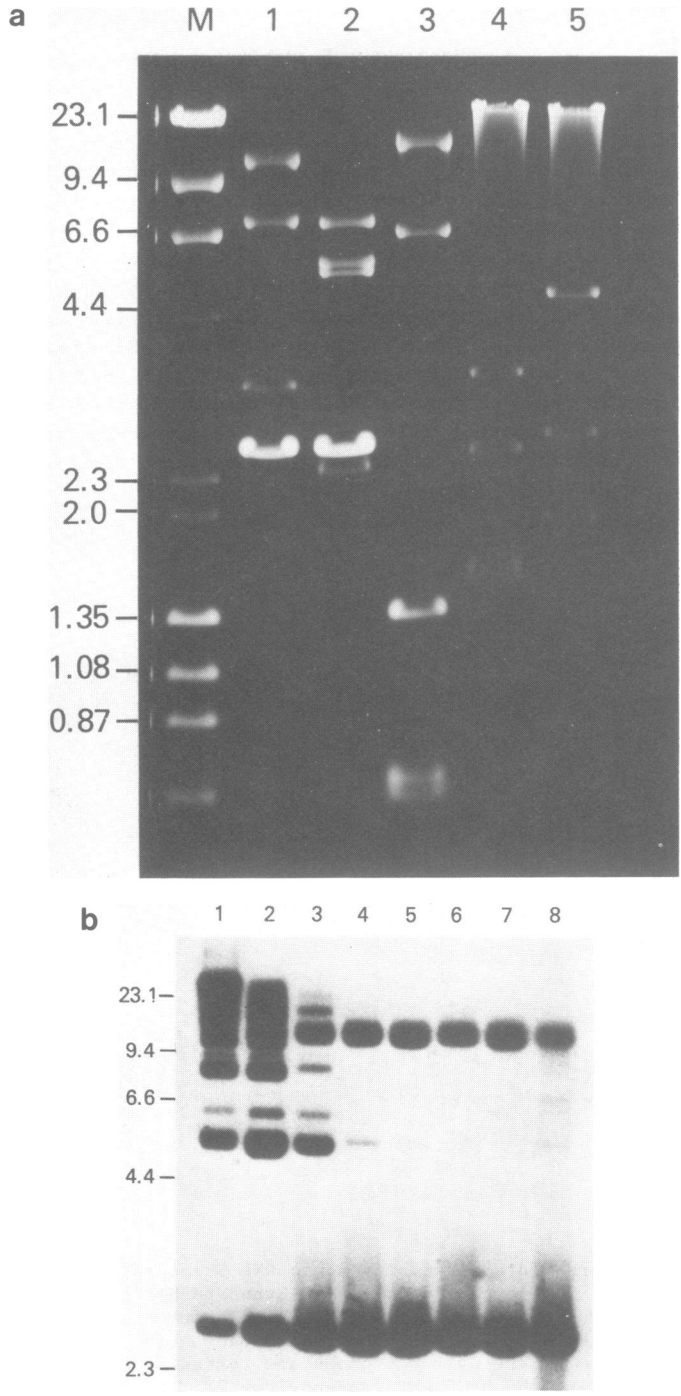


Fig. 1. (a) Southern blots of human genomic DNA probed with the entire adenovirus 5 genome. Locations of Standard Molecular weight markers are indicated at the left. Lanes are as follows: (1) K562, (2) HeLa, (3), (4),(5) normal humans. (b) The same filter washed free of adenovirus 5 probe and tested by 24 hr. exposure was reprobed with the 5′ 9 kb of adenovirus 2.


a
M
1
2
3
4
5
23.1
9.4
6.6
4.4
2.3
2.0
1.35
1.08
0.87
b
1
2
3
4
5
6
7
8
23.1
9.4
6.6
4.4
2.3

The 2.5 kb Band Reflects Hybridization of Adenovirus to Members of a New Moderately Repetitive Sequence Family

Fragments containing sequences homologous to the 5' 9 kb adenovirus segment were isolated from a human cosmid library. When purified DNA from a positive clone was restricted with either Sst I or Pvu II, a 2.5 kb fragment was released in vast molar excess (Figure 2a) suggesting the possibility of a tandem array. Consistent with this interpretation is the fact that Hind III and EcoR I released fragments of greater than 25 kb from which we infer that the 2.5 kb fragments released by Sst I or Pvu II lack EcoR I or Hind III sites. Partial digestion with Sst I verified the tandem array (Figure 2b). BamH I sites are present within the tandemly repeated 2.5 kb fragment as indicated by the presence of smaller fragments in molar excess (figure 2a - lane 3).

The tandem array of the 2.5 kb Sst I fragment was shown to be present in genomic DNA. One of the Sst I fragments from the cosmid clone was sub-cloned into pUC12 and used as a probe in Southern blots of human DNA, partially restricted with Sst I (Figure 3). Predicted bands of 2.5, 5.0, 7.5, 10.0, 12.5, 15, and 17.5 kb could easily be recognized, providing proof for the tandem array of this sequence in normal cellular DNA.

The sequence of the 2.5 kb Sst I fragment that had been sub-cloned was determined and is present in Figure 4a. A computer assisted search [32] against adenovirus DNA failed to reveal any long segments of homology. The truncated fragments that had been generated for sequencing were probed with the 5' 9 kb adenovirus fragment in dot blots. The results indicated that the annealing segment was located between nucleotide 1000 and 1200 (data not shown). This region is extremely G-C rich and contains several (14-20) nucleotide segments nearly identical to similar non-adjacent segments in the 5'end of the adenovirus genome. (fig. 4b) The T_m for several of the homologous segments was calculated as greater than 55° C and one has a T_m of greater than 65° C. We interpret the original "positive" signal on the Southern blot as reflecting (1) annealing of the short G-C segments and (2) the high copy number of the 2.5 kb fragment.

Appropriate computer analysis failed to reveal internal or inverted repeats within the 2.5 kb Sst I fragment. Furthermore, we found no homology to any known unique or repetitive DNA sequences including members of the Alu or Kpn I families and the genes for histones, or ribosomal or small nuclear RNA's. Thus the isolated 2.5 kb repeat is a member of a new moderately repetitive sequence family. We have designated these sequences the Sst I family.

Fig. 2 (a) Isolated cosmid clone was digested with several enzymes. Digests were run on a 1% agarose gel and stained with ethidium bromide. Lane (1) molecular weight marker (2) Sst I (3) Pvu II (4) BamH I (5) EcoR I (6) Hind III. (b) Southern blot of Sst I partial digest of isolated cosmid clone probed with 5' 9 kb of adenovirus 2 The persistent 10 kb bands represents pHC79 vector which is not cut by Sst I joined at Mbo I site to an internal fragment of Sst I repeat.

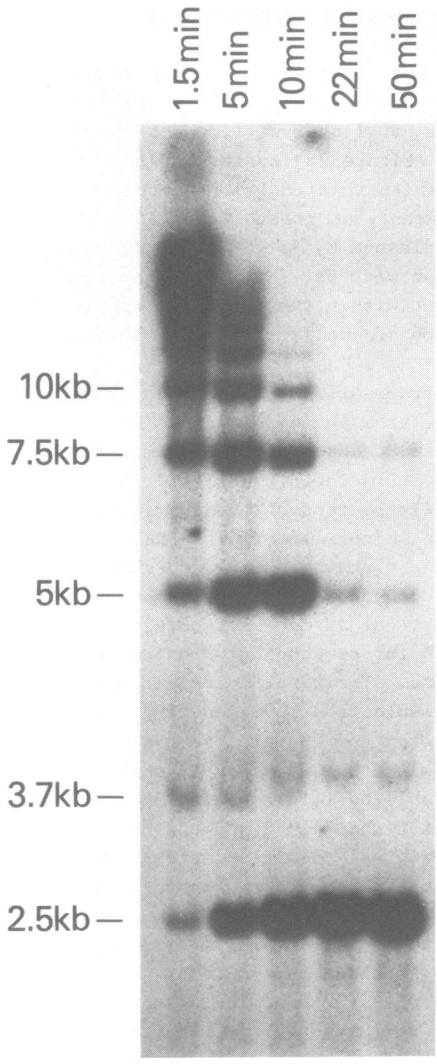


Fig. 3 Southern blot of Sst I partial digest of normal human DNA probed with an isolated member of Sst I repeat. Fragment sizes were estimated by semilog plots against standard molecular weight markers.

Members of the SstI Family Are Clustered At Only Two Chromosomal Locations

In order to determine the chromosomal location of the Sst I tandemly repeated members, a panel of 45 previously characterized rodent-human lymphocyte somatic cell hybrids was examined. The DNA from the hybrid cells were digested with Sst I and subjected to Southern analysis with the sub-cloned and sequenced 2.5 kb Sst I fragment. 30 cell lines were found to contain the 2.5 kb Sst I fragment whereas the remaining 15 cell lines were negative. Analysis of the discordancies between the presence or absence of the Sst I 2.5 kb fragment and the presence or absence of

specific human chromosomes in the hybrid cell lines failed to yield convincing data.

Localization was therefore approached through in situ hybridization (figure 5). A total of 125 grains were distributed throughout the chromosomes of the 56 cells examined. Of this total, a significant number of grains were found to be localized along two different chromosomes; 29 grains fell over chromosome 4 (7-8 expected; $p<0.001$ and 29 grains along chromosome 19 (2-4 expected; $p< 0.001$). Twenty-three of the 29 grains along chromosome 4 were found in region q28-q31 while 29 of the grains along chromosome 19 occurred in the region q13.1-q13.3. Subsequently when the somatic cell hybrid data was reevaluated, considering chromosomes 14 and 19 together, a 95% concordance was obtained confirming the in situ data. A graphic display of this data appears in figure 6. Thus the vast majority if not of the Sst I family members are found at these two chromosomal locations.

DNA from Several Individuals Contains Approximately 400 Copies of the Sst I Family

Variable numbers of short tandemly repeated sequences may form the basis for restriction fragment length polymorphisms. (18,33,34). We wondered whether an analogous dose polymorphism might exist within the Sst I fragment among humans as its members are present in tandem array and therefore might be susceptible to expansion or contraction by homologous, but unequal recombination. This hypothesis was investigated by probing dot blots of DNA from several individuals of various ethnic groups with the Sst I family member, a Kpn I family member and an Alu family member. The latter two families are dispersed and therefore likely to be present in equivalent copy number in individual humans. We found that the signal intensities generated with the three probes, when compared to one another, were perfectly proportional. Thus there is no evidence for dose polymorphism for the Sst I family.

The copy number of the Sst I family was estimated by comparison of the signal intensities generated with the Sst I and Kpn I fragment probes. The Kpn I fragment probe was derived from the 5' end of the 6.1 kb repeat downstream from the beta globin gene; there are approximately 4000 copies of this portion of the repeat in the human genome. By comparison of the signal intensities, we concluded that the Sst I family has approximately 400 members.

DISCUSSION

The Sst I repeat family is clearly distinct from other repetitive DNA by virtue of its clustered tandem array, location, and lack of sequence homology to known repetitive sequences (Table 1). It resembles dispersed repetitive DNA in that it is not conserved in evolution. Of the homogeneity of the Sst I family we know only that it's length is constant and that Sst I, Pvu II and BamH I sites are preserved. The Sst I family is similar to the set of clustered multigene families in its structure but differs in that it has no known function.

a

```
   1 GAGCTCTGGG CTTCCATACC TGTGTGGGAC AGGGAAGCTC TCTCGGTCTC
  51 CATGGCCCAA GTGATGGCTG CACGCTCGGT CCAGGAAGAG GCGGAGGAAG
 101 CCCACCGCTC CTGACATTGG CCTTCTAGGA AAGGCGGTGT TGCATCCCAC
 151 CTGCACTTCC TCTCTGATTC TTGAGGGCCA ACCGCTTCCT CCGCTCCTGG
 201 GGAAAGTGCC TTCTAGCACC GAATCTTTTG GCTGCCACGG ATGTCAGGGA
 251 GCCAACGGGA CTGGGTTTTG GCTGGGTGCA GGGGAGGTTG CGTCAGGGGT
 301 ACTAGCCGGC GGCGGGCTGG GGGTGGGGTG TACTTTGTCC AAACTCCCGG
 351 CTCCTCTGGC GGGCCTCCCT GAACGTGGCG TGGACTCGCG CACAGGCCCT
 401 GTCTCGCAGG TTTTCAGGTG CGCTTGGCTT TTCCTCCGCT TTGTGGGGCA
 451 GGTCTCCAGT GCCCCCGGCG CACGCCTGGA CATCACTGTC CGTCTCGTCG

 501 TCGCCCCTAC GGCCTCAAAG ACACACGCTG CCTGCATGTG CTCTTGGGGG
 551 ACGACAGTGC ACATGTGGAC ACACTGGCTC CAGCTCGGAC TCGCCTCTGT
 601 CTCTCTTTGC CCGTGTCGCC GGAAGCCGCC TCGGGTTGCC GGAGCCCTCG
 651 GGCCTTGGAG ATGAAGGCAG GCCCCTGCTC CTGCCAGGAA GGAGGGAGGC
 701 AGTGGGCTCA TGGGTCGGTG CCTTTGCAGC CGACAGCACG TCCGCCCCTG
 751 GGGATCTTCC TGTGCCCCGG CGAGACCCTT TCCGCCTCAC TGCATTGGAA
 801 CCCCATTCCC GATCACCCGC TGGGATCCAT CATCGGACTC CAAGAGGAGT
 851 CCGCGCAGCC AGCCGGCACC CCGAAGCTCC TCCTTCAGCG GGAACCGAAG
 901 CAGAAGAGCG ATCAAGGAGG TCCTCACCAC AGGACTCCTA TGGGTCCGAC
 951 CCTGGGTCTC CCGCAGGCCC CTCTGGCAGT CCTCTTCCCA CCCGCCGCCT

1001 CGGCTTCGCC GCCGCCGCCG CAACCTCCAG CACCGCCCCC CAGGCCCCGC
1051 AGCCGCCGTC GCCGCCATTT TTTAAAGGGT CGCAGCCTGA CTCTGCGGAG
1101 TAAGGGGGGG TGGAGCGGGG GAGTCGCTCG CCAGCATGCG CGAGCCCGAG
1151 CCGCCGCTTG GGTCACAGTG AAAGCCACCG TTGCCCCGGG ATGGGTCCCT
1201 GACACTTGGG GAAGTAGGAG CCCTGTGTGA TCGTGCGTCT GAGTCTGGGC
1251 TGAGACCAGT CCTGGCCAGG GCAGTTACCA GGACGGTCTC GGAGGCCGGG
1301 ATTCGCCGGAG GGTCCAGCAG CAGGAAGAAA CCCCAGGAGG AAGAAACCTC
1351 AGACAGATCG CCGGCGAGGC AGCGCGGGAT CCCAGCCTCA GGCGTGCGCG
1401 GACGGTGTGC GGGTGAGTCT CCCCAAAAGT GGAGCCCTTG TGATGACGAG
1451 CACAGGTCCG CCTGCGTGCC CGTGGGCGGC TCTCTCACCG GTGGCTCTCA

1501 GTCGCGGAGA GCAGAACCCG CAGCTTCAGG GGCTGCTGCG GGAGGGTGTT
1551 CCCTGCTGTA CGTGTGTGTT CGTCATGGGG GTGTGTGTGT GTGTGTTGGG
1601 GAGGGTACGT CTGTGTGTGT GTCTGTGTGT GTGCGCGCGC AGTGCCTGTC
1651 TGTGTCGCGA CTTCTGTCTC TCTCTCACGT CTCTCTCTCT CTCTCTCTCT
1701 CCCTCTCCCT CTCCCTCTCT CCCTTCTCGC TCTTTCCGTG GCCCTCTCTT
1751 TCTGTCTCTG TCCGTCTGTG TGTCGTGCGC CTCGGGACAC ATGTGCCCTG
1801 TGCGCCGGAG GGTGGGTTTC TTGCACGTCG GCCTTTCTTC TGGTCAGCCT
1851 CTCCCCGCGT CTCTGCCTGG GTCGTGTGGC CGGTTGGCAG TCGTCGTCCC
1901 GGCGGTTCCA GTTTGGGGGT CTGTGAAGGC CTGGGCAACG TGGGCATCGG
1951 CGTCGGACCG CAGGGGTTTT CATCCCCTCC CCATCCGGAG CAGCCTCTTT

2001 GCTAGGCTGG ATCCAGACGA CCGCTCCCCA ACCAAGGACA ACGGCCTCCC
2051 AGGCGCTCAT CGTCCACCCG CAGGAGGGTG CCCGCAGAGC TTCAAGAAGG
2101 TGGTTGTCAC GCGTGTCGCC CTCTGCCCTC ATCGAGAAAT GTAGCCACAG
2151 TCGACGCAGG GACGGAGAAG GAAGCCGGCA AGGGGATGGG GCAAGCATGT
2201 CTGTCTCTCA AAGGCTGCTT CCCGGCCGAG TCACCCGTTT GACACTCCTC
2251 CCGGATGCCG GTGGTGGTGC ATGGCCCCCC CGTATCCTGC CTGGGCTCTG
2301 GCCTCTGCTC TGACCTCCCT CTTGCTGTGT CTGCCCCGTC TCTGAGCTGG
2351 CGGCTTCTTA GTGTGGCTCA GTGTCTTCCA CAAAGAAGAC TTCCCCGTCC
2401 ATCAGGGAGA AACGTCGTGC GGTCCGCGTC ATGATTGTTT CCCTCTCCAC
2451 ACCTCTTTCT GGATGATTGG GCAGCTGTGG TGATCCTGGA GCTC
```

b

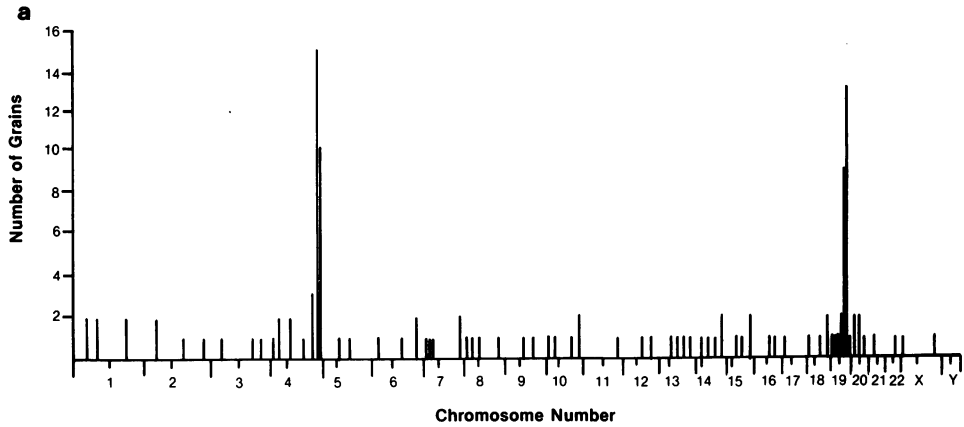
```
1          1010                                    +
     GCGGCGGCGGCGGC                          SstI member
     GCGGCGGCGGCGGC                          Adenovirus 2
       2180

2.       1060                                    +
     GGCGGCGACGGCGGC                         SstI member
     GGCGGCGGCGGCGGC                         Adenovirus 2
        2180*

3.       1000       1010
     CCGCCTCGGCTTC-GCCGCCGCCGCC              SstI member
     CCCCCGCCGCCTCAGCCGCTGCAGCC              Adenovirus 2
        3790*      *      * *

4.            1020
     GCCGCCGCCGCCGC                          SstI member
     GCCTCCGCCGCCGC                          Adenovirus 2
        *   3790

5.        1040
     CCGCCCCCCAGGCCCCGCAG                    SstI
     CCGCGCCGCAGGCCCCACAG                    Adenovirus 2
         *  * 5600

6.   1046               1064
     CCCGCAGCCGCCGTCGCCG                     SstI
     CCCGCCGCCGCCGTCGCCG                     Adenovirus 2
     17527*
```

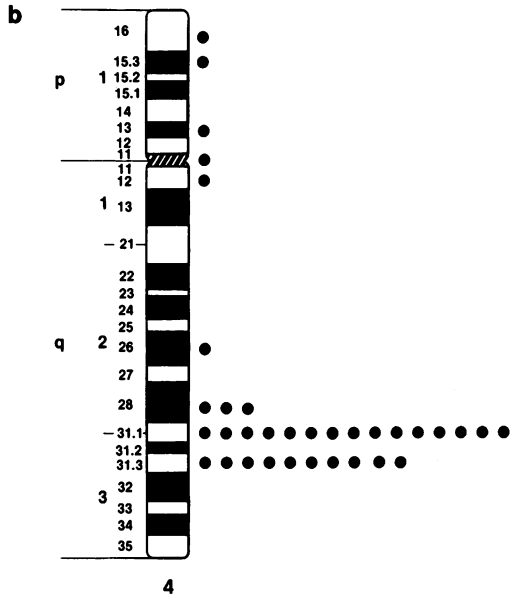
The absence of any known function for an amplified sequence makes it difficult to determine if the sequence has been preserved through natural selection at the organismal level or is merely an example of selfish DNA. (45-49) The theory of selfish or parasitic DNA has grown out of opposition to the tendency of ascribing phenotypic benefit, either immediate or evolutionary, to classes of repetitive DNA. It rests on the reasoning that if the only selection pressure DNA encounters is to survive within cells, that sequences will arise whose only function is self preservation. Whether this class of amplified DNA is truly unnecessary to the organism, a by-product of essential operations of recombination and amplification, or a pool from which new genes can arise through mutation and recombination is a matter for speculation. Roberts and Axel (50) have shown that an essential gene transfected into eukaryotic cells is amplified as a cluster under selective conditions. They have furthermore demonstrated the existence of a nonreciprocal process of gene conversion which can maintain the homogeneity of this amplified family. Gene conversion in higher eukaryotes has been implicated in homogenization of small repeated families such as encode hemoglobins, immunoglobulins, and the proteins of the major histocompatibility loci. (51-54) It is not known if it influences repetitive DNA in general although reciprocal recombination between partially homologous repeats has been used in computer modeling to predict an intrinsic tendency towards simple periodicity in DNA under nonselective conditions. (55) The possibility that gene conversion also occurs in moderately repetitive DNA (56) makes evaluation of a repetitive family without known function all the more difficult.

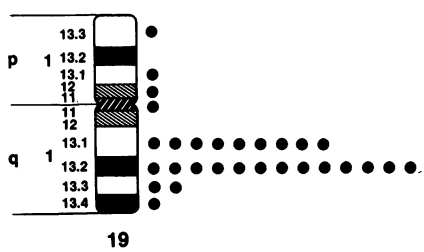
There is both evidence for and against the Sst I repeat family encoding a protein. Analysis of the sequenced member shows several open reading frames of approximately 600 bp and one greater than 700 bp (Table 2). Simple probability theory predicts a 1 in 20 chance of finding a stop codon at any one point in a random sequence of DNA. This should preclude such relatively long open reading frames in unselected DNA. Whether in actuality random sequences of DNA exist with open reading frames of this size is not known. A search for control signals has failed to show typical RNA polymerase II or III control sequences. ON the other hand all pol II dependent transcripts do not require a TATA box and deviation from a consensus sequence is not uncommon. Preliminary Northern blots have in fact shown an RNA fragment of 6 kb size which anneals to the Sst I family member (unpublished observations).

Other intriguing information relevant to the Sst I repeat includes the presence of a constitutive fragile site at 4q31. (57) This is interesting in view of the tandem array structure and the theoretical possibility for

Fig. 4 (a) Sequence of isolated Sst I repeat member. The region between 1001 and 1078 contains short G-C righ segments homologous to nonadjacent segments of the left end of adenovirus 2. (b) Some of the short G-C segments in the Sst I repeat member homologous to the left side of the adenovirus genome. The numbers indicate the position in the sequence. * indicates a mismatch, + indicates the complement of the sequence as it appears in figure 4a.


a
Number of Grains
16
14
12
10
8
6
4
2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome Number


b
p
q
1
2
3
16
15.3
15.2
15.1
14
13
12
11
11
12
13
21
22
23
24
25
26
27
28
31.1
31.2
31.3
32
33
34
35
4


p
q
1
13.3
13.2
13.1
12
11
11
12
13.1
13.2
13.3
13.4
19

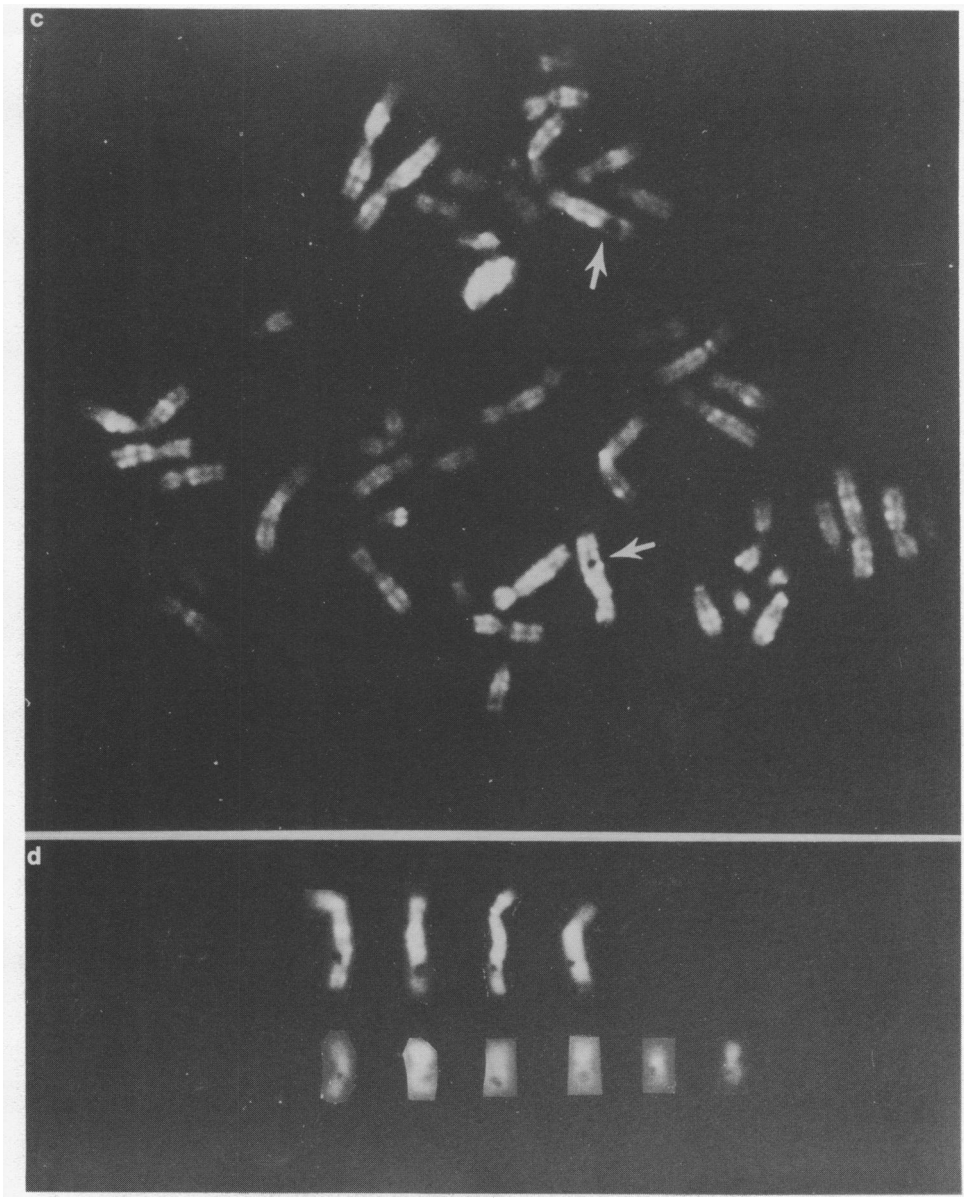


Fig. 5 Localization of Sst I repeat member using in situ hybridization. (a) Histogram of the distribution of 125 grains from 56 metaphase cells. (b) Ideograms depicting grain distributions along chromosomes 4 and 19. (c) Q banded metaphase cell with both members of pair 4 labeled (arrows). (d) A series of chromosome 4 (upper row) and chromosome 19 (lower row) from several different Q banded cells demonstrating typical labeling pattern.

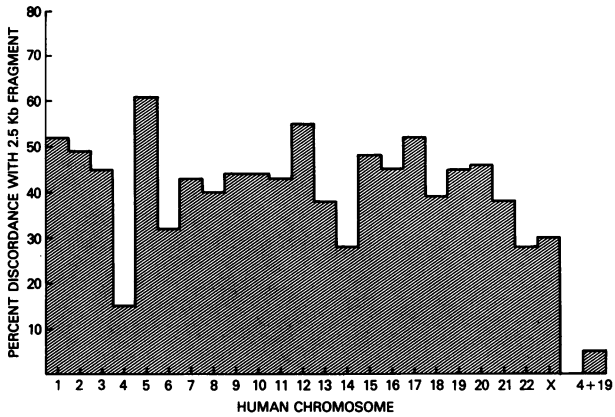


Fig. 6 **Analysis of a hybrid panel of 45 rodent-human somatic cell hybrids for the presence or absence of 23 human chromosomes and of 2.5 kb Sst I fragment. The last entry on the X axis shows the % discordance of the 2.5 kb fragment vs. human chromosome 4 and/or 19.**

Table 1. Families of Dispersed and Clustered Repetitive DNA*

	Repeat	Length	Copy Number	Comment	Reference
Dispersed	Alu	approx. 300 bps	5×10^5	Prototype of SINES. Recent evidence suggest SINE families are derived from mammalian tRNA genes.	35-36
	Kpn	Ave.2 kb longest-6.1 kb	4000 - 50,000	Prototype of LINES.	5-8
	CA	40 bps	4500	Present in all eukaryotic DNA.	6
	O	373-435	4500	Recent evidence suggests O repeats may be solitary LTRS. Some are linked to longer transposon-like elements.	6,37
Clustered	U1	164 bps	30-125 bonafide genes 10 fold greater pseudogenes	Role in splicing of mRNA precursors. Genes are loosely clustered with intergenic distances greater than 15 kb. Majority of bonafide U1 genes are on chromosome 1. Pseudogenes are dispersed throughout the human genome.	38-41
	U2	188 bps coding region imbedded in 6 kb repeat unit	10-20	Role in splicing of mRNA precursors. Bonafide U2 genes organized in a tandem array on chromosome 17 bands q21-22.	19-20 42-43
	Ribosomal 18S 5.8S 28S	40-45 kb repeat approx. 31 kb is non-transcribed spacer	150-300	Arranged in tandem clusters located in the satellite regions of acrocentric chromosomes 13, 14, 15, 21, 22.	16-18
	Histone		30-40 copies	A cluster of H_4 coding sequence localized to the long arm of chromosome 1. Other clusters containing H_3, H_1, H_{2A}, and H_{2B} with or without H_1 have been assigned to chromosomes 1, 6 and 12.	21-23 44

*Does not include the set of endogenous retroviral like elements whose identifying characteristics exclude them from this discussion.

crossover. It does not fit with either the lack of any known transposition between 4q31 and 19q 13.1-q13.3 or the conservation of dose that we have indicated. Recent findings that adenoviral 12 sensitive sites correspond to clusters of tandem arrays of U1 and U2 genes on chromosome 1 and 17 (10) suggest other possible consequences for tandem arrays. Of note are studies of somatic cell hybrids that suggest the human cell susceptibility to Coxsackie B3 virus is determined by a locus on chromosome 19(58) as is polio virus. (59) echo virus, (60) and baboon virus. (61)

ACKNOWLEDGEMENT

We thank D. Alling for his help with the statistical analysis and J. Maizel and C. Buchler for their help with the sequence analysis. We are grateful to Ms. R. Mayes, Ms. M. Gaizband and Ms. J. Danzik for their assistance in preparation of the manuscript.

*To whom correspondence should be addressed: Building 10, 7C103 NIH, NHLBI, 9000 Rockville Pike, Bethesda, MD 20892, USA

REFERENCES

1. Grodzicker, T. and Klessig D.F. (1980) Cell, 21 453-463.
2. Braithwaite, A.W., Lejeune, S. and Hiroto, N. DNA (1984) 3 223-230.
3. Karlsson, S., Humphries, R.K., Gluzman, Y. and Nienhuis, A.W (1985) Proc. Natl. Acad. Sci. (USA) 82 158-162.
4. Lewin B. (1980) Gene Expression 2nd edition, Eukuryotic Chromosomes, Wiley, New York.
5. Singer, M.F. (1982) Cell 28 433-434.
6. Sun, L., Paulson, K.E., Schmid, C.W., Kadyk, L. and Leinwand, L. (1984) Nucleic Acids Res. 12 2669-2690.
7. Adams, J.W., Kaufman, R.E., Kretschner, P.J., Harrison, M. and Nienhuis, A.W. (1980) Nucleic Acids Res. 9 6113-6128.
8. Hattori, M., Hidaka, S., Sakaki, Y. (1985) Nucleic Acids Res. 21 7813-27.
9. Kole, L.B., Haynes, S.R. and Jelinek, W.R. (1983) J. Mol. Biol. 165 255-270
10. Shafit-Zagardo, B., Brown, F.L., Zavodny, P.J. and Maio, J.J. (1983) Nature 304 277-280.
11. Tashima, M., Calabretta, B., Torelli, G., Scofield, M., Maizel, A. and Saunders, G.F. (1981) Proc. natl. Acad. Sci. (USA) 78 1508-1512.
12. Jagadeeswaran, P., Forget, B.G. and Weissman, S.M. (1981) Cell 26, 141-142.
13. Edler, J.T., Pan, J., Duncan, C.H. and Weissman, S.M. (1981) Nucl. Acids Res. 9 1171-1189.
14. Jelinck, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Biro, P.A., Choudary, P.V., Weissman, S.M., Rubin, C.M., Houck, C.M. Deininger, P.L. and Schmid, C.W. (1980) Proc. Natl. Acad. Sci. (USA) 79 1398-1402.
15. Ullu, E., Murphy, S. and Melli, M. (1982) Cell 29 195-202.
16. Henderson, A.S., Warburton, D. and Atwood, K.C. (1972) Proc. Natl. Acad. Sci. (USA) 69 3394-3398.
17. Wellauer, P.K. and Dawid, I.B. (1979) J. Mol. Biol. 128 289-303.

18. Erickson, J.M. and Schmickel, R.D. (1985) Am. J. Hum. Genet. 37 311-325.
19. Lindgren, V., Ares, M. Jr., Weiner, A.M. and Francke, U. (1985) Nature 314 115-116.
20. Van Arsdell, S.W. and Weiner, A.M. (1984) Mol. and Cell Biol. 4 492-499.
21. Green, L., Van Antwerpen, R., Stein, J., Stein, G., Tripputi, P., Emanuel, B., Selden, J., Croce, C. (1984) Science 226 838-840.
22. Marashi, F., Prokopp, K., Stein, J., Stein, G. (1984) Proc. Natl. Acad. Sci. (USA) 81 1936-1940.
23. Sierra, F., Lichtler, A., Marashi, F., rickles, R., Van Dyke, T., Clark, S., Wells, J., Stein, G. and Stein, J. (1982) Proc. Natl. Acad. Sci. (USA) 79 1795-1799.
24. Southern, E.M. (1975) J. Mol. Biol. 98 503-517.
25. Chen, M.J., Shimada, T., Davis-Moulton, A., Cline, A., Humphries, R.K., Maizel, J., and Nienhuis, A.W. (1984) J. Biol. Chem. 259 3933-3943.
26. Grosveld, F.G., Dahl, H.M., deBoer, E. and Flavell, R.H. (1981) Gene 13 227-237.
27. Woods, D. (1984) Focus (Published by Bethesda Research Labs) 6 1-2.
28. Ponc̀z, M., Solowiejczyk, Ballantine, M., Schwartz, E. and Surrey, S. (1982) Proc. Natl. Acad. Sci. USA 79 4298-4302.
29. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74 5463-5467.
30. Harper, M.E., and Saunders, G.F. (1981) Chromosoma (Berl.) 83 431-439.
31. Tuckey, J.W. Biometrics (1949) 5 232-242.
32. Queen, C. and Korn, L.J. (1984) Nucleic Acids Res. 12 581-599.
33. Bell, G.I., Selby, M.J. and Rutter, W.J. (1982) Nature 295 31-35.
34. Spritz, R.A. (1981) Nucl. Acids Res. 9 5037-5047.
35. Schmid, C.W. and jelinek, W.R. (1982) Science 216 1065-1070.
36. Daniels, G.R. and Deininger, P.L. (1985) Nature 317 819-822.
37. Paulson, K.E., Deka, N., Schmid, C.W., Misra, R., Schindler, C.W., Rush, M.G., Kadyk, L. and Leinwand, L. (1985) Nature 316 359-361.
38. Busch, H., Reddy, R., Rothblum, L. and Choi Y.C.A. (1982) Rev. Biochem. 51 617-654.
39. Naylor, S.L., Zabel, B.U., Manser, T., Gesteland, R. and Sakaguchi, A.Y. (1984) Somat. Cell Molec. Genet. 10 307-313.
40. Lund, E., Bostock, C., Robertson, M., Christie, S., Mitchen, J.L. and Dahlberg, J.E. (1983) Molec. Cell Biol. 3 2211-2220.
41. Denison, R.A., Van Arsdell, S.W., Bernstein, L.B. and Weiner, A.M. (1981) Proc. Natl. Acad. Sci. USA (78) 810-814.
42. Oshima, Y., Itoh, M., Okada, N. and Miyata, T. (1981) Proc. Natl. Acad. Sci. USA 78 4471-4474.
43. Branlant, C., Krol, A., Ebel, J.P. Lazar, E., Haendler, B. and Jacob, M. (1982) Embo J. 1 1259-1265.
44. Tripputi, P., Emanuel, B.S., Croce, C., Green, L.G., Stein, G.S. and Stein J.L. (1986) proc. Natl. Acad. Sci. USA 83 3185-3188.
45. Orgel, L.E., Crick, F.H.C. and Sapienza, C. (1980) Nature 288 645-646.

46. Dover, G. and Doolittle, W.F. (1980) Nature 288 246-247.
47. Jain, H.K. (1980) Nature 288 247-248.
48. Orgel, L.E. and Crick, F.H.C. (1980) Nature 284 604-607.
49. Doolittle, W.F. and Sapienza, C. (1980) Nature 284 601-603.
50. Roberts, J.M. and Axel, R. (1982) Cell 29 109-119.
51. Slighton, J.L., Blechl, A.L. and Smithies, O. (1980) Cell 21 627-638.
52. Baltimore, D. (1981) Cell 35 157-165.
53. Liskay, R.M. and Stachelek, J.L (1983) Cell 35 157-165.
54. Weiss, E.H., Meller, A., Golden, L., Fahrner, K., Simpson, E., Hurst, J. and Flavell, R.A. (1983) Nature 301 671-674.
55. Smith, G.D. (1976) Science 191 528-535.
56. Edelman, E.M. and Gally, J.A. (1970) In Schmitt, F.O. (ed.) Neurosciences Second Study Program Rockefeller University press, New York.
57. Yunis, J.J. and Soreng (1984) Science 226 1199-1203.
58. Gerald, P.S. and Bruns, G.A. (1978) Birth Defects Orig. Art. Ser. XIV(6A) 1-7.
59. Miller, D.A., Miller, O.J., Dev, V.G., Shahna, H. and Tantravahi, R. (1974) Cell 4 167-170.
60. McKusick, V.A. (1983) In Mendelian Inheritance in Man 6th ed. Johns Hopkins Univ. Press, Baltimore p. 148.
61. Schnitzer, T.J., Weiss, R.A., Juricek, D.K. and Ruddle, F.H. (1980) J. Virol. 35 575-580.