The structure of a subterminal repeated sequence present on many human chromosomes

S.Cross+, J.Lindsey, J.Fantes, S.McKay, N.McGill and H.Cooke*
MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Received June 29, 1990; Revised and accepted September 14, 1990 EMBL accession nos X56275-X56278 (incl.)

ABSTRACT

All telomeres which have been studied consist of an array of simple G/C rich repeats. Human telomeres were shown to share sequence similarity with those of lower eukaryotes by cross-hybridization and human telomeric sequences have been cloned by complementation of telomere function in yeast. Analysis of human telomeric sequences cloned in this way is described here. The terminal part of the cloned human telomeric DNA consists of an array of simple repeats, principally of the sequence TTAGGG and derivatives. The very terminal part consists of yeast-type telomeric repeats which suggests that the human telomeric sequences have acted as a primer for the addition of additional telomeric repeats in the yeast. Subterminal sequences are shared between a number of clones and *in situ* data shows that these subterminal sequences are present at several different chromosomal ends. Related sequences are present at internal as well as telomeric positions. Differences in the hybridization patterns of subterminal sequences in somatic compared to germ-line tissues are described which indicate differential modification of these sequences during development.

INTRODUCTION

The telomere is the DNA-protein structure found at the end of a linear chromosome and it is important for the stability and the complete replication of the chromosome end. Interactions between different telomeres and between telomeres and the nuclear envelope may be important in the spatial organisation of the chromosomes in the nucleus (1).

Telomeres consist of tandem repeats of a short DNA sequence which has a marked strand asymmetry in base composition with the G rich strand running 5'-3' towards the end of the chromosome. In yeast this feature seems to be sufficient for telomere function and this has proven to be a useful method for selectively cloning telomeres from a variety of species including human (2, 3). The human telomeric repeat is TTAGGG (4) and somatic chromosomes have 5-15 kb of telomeric repeats whereas sperm chromosome telomeres appear to have an additional 10 kb of the repeats (2, 5).

In a number of species these short terminal repeats, which are common to all chromosomes, are followed by more complex repeated sequences which are present on some, but not all, chromosomes. No function has yet been clearly demonstrated for these sequences. In yeast these sequences are subject to frequent rearrangement through recombination and have ARS activity. The locus DXYS14 is located subterminally in the pseudoautosomal region of the human sex chromosomes and the number of loci found varies from sex chromosome to sex chromosome within the population although they are always located close to the chromosomal end (6, 7). Other, autosomal, subterminal regions have been cloned either in the same way as is described here (8) or directly in plasmid vectors (9). These sequences are repeated and are present at a number of genomic sites.

Here we describe the sequence and molecular structure of two human telomeres cloned by complementation in yeast. Sequence analysis shows that these repeats are principally of the sequence TTAGGG. The distal part of the clones consists of the yeast type of telomeric repeat, TG_{1-3}.

Immediately subtelomeric sequences from the two clones are related and represent members of a family of sequences which are present largely, but not exclusively on the terminal fragments of human DNA. One subtelomeric sequence displays different patterns of hybridising fragments when germ-line and somatic DNAs are digested with certain restriction enzymes.

MATERIALS AND METHODS

DNA isolation

DNA was prepared from human tissues by standard methods (10). For DNA fractions enriched for ova, ovaries from a 16 week foetus were rinsed in PBS, one was homogenised whole and DNA extracted, the other was teased open with tungsten needles in 0.2 ml of PBS and the resulting cell suspension and the remaining ovarian capsule used for separate DNA extractions. Preparation of yeast DNA and intact yeast chromosomes was as described (2).

YAC library construction

Construction of a collection of clones in which human sequences provide telomere function in yeast was carried out essentially as described in ref 2. After ligation of DNA to pYAC4Neo Not

* To whom correspondence should be addressed

+Present address: Institute for Molecular Pathology, Vienna, Austria

(11) the ligation mixes were used to transform yeast spheroplasts of strain YPH274. Ura^+ colonies were picked into selective media in microtitre wells and after overnight growth replicated onto nylon membranes on selective agar plates. After cell lysis and DNA denaturation these filters were hybridised with $(TTAGGG)_4$ and hybridising colonies were picked and grown as 50 ml cultures for DNA and chromosome analysis; those colonies containing short linear artificial chromosomes were used for further analysis.

Subcloning

Terminal fragments from chromosomes were cloned by a brief Bal 31 digestion followed by ligation (12) of an adaptor fragment resulting in molecules terminated by a Hind III site. After Bam HI digestion and size fractionation this DNA was cloned into pGEM7. Colonies containing human sequences were identified by hybridisation with $(TTAGGG)_4$. The human telomeric sequences in Hutel 1 were subcloned in a similar way using an EcoRI adaptor and cloned as an XhoI-EcoRI fragment in pBS.

Sequencing

DNA sequencing was carried out on fragments subcloned either in pBluescript or M13mp18 or 19 on either single or double stranded templates. Unidirectional deletions were also used starting from the terminal repeat end of the clones. These were sequenced on one strand only using 7-azaguanidine. On average sequence was determined from three independent clones using Sequenase V2.0. We were unable to completely sequence regions of GC rich repeats. Single stranded phage or phagemids from these regions were deleted and sequence was derived by double stranded sequencing. Sequences of a few copies of the repeats were determined using 7-azaguanidine and Taq polymerase.

Hybridisation and Southern analysis

Southern blotting was carried out by standard protocols onto nylon membranes (Amersham). Filters were hybridised with random primed probes (13) in 0.5 M Na_2HPO_4 pH 7.2, 7% SDS and 0.5% dried milk at 65°. Final washes were in 0.1×SSC 0.1% SDS at 68°. ^{32}P labelled oligonucleotide probes were hybridised as described (2). Final washes were at 65° in 4×SSC, 0.1% SDS.

In situ analysis

Metaphase spreads were prepared from short term peripheral blood cultures from individuals with normal karyotypes. Colcemid (0.1 mg/ml) was added for the final hour of culture; the cells were treated with hypotonic 0.075M KCl for 8 min at room temperature and then fixed three times in 3:1 methanol:acetic acid. Slides were stored under vacuum in a desiccator at room temperature for at least 5 days. Inserts were labelled with biotin by random oligonucleotide primed synthesis (13) using biotin-16-dUTP or biotin-11-dCTP. The slides were treated with RNase A (0.1 mg/ml for 1 hr at 37°C) and proteinase K (60 ng/ml for 4 min at 37°C) as described previously (14). 40 ml of hybridisation mixture (biotinylated probe at 100 ng/ml, 50% deionised formamide, 2×SSC, *E. coli* tRNA at 500 mg/ml and 20% dextran sulphate) was applied to each slide. Chromosome and probe DNA were denatured together at 70°C for 10 min before overnight incubation at 37°C. Slides were washed after hybridisation in 50% formamide; 2×SSC (4 washes of 3 min at 45°C) and 2×SSC (4 washes of 3 min at 45°C). Hybridised probe was detected with avidin FITC DCS (Vector) using the method of Pinkel (15) but substituting 4×SSC, 0.05% Triton×for bicarbonate buffer in all antibody dilutions and washes. Signal was amplified by successive incubations with biotinylated goat anti-avidin and avidin FITC and the best results were achieved with 3 layers of avidin FITC. After the slides were counterstained with 2 mg/ml propidium iodide in antifadent solution they were examined using a BioRad laser scanning confocal microscope. Laser light at 514 nm was used to excite both FITC and PI and separate images were obtained with narrow band pass filters. These images were merged electronically and enhanced by digital filtering. Only fluorescence signals at the same position on both chromatids were scored as sites of hybridisation to reduce the effect of random background. The slides were restained with DAPI (0.5 mg/ml) to give banded chromosomes and photographs of each cell were karyotyped and compared with stored images to identify the chromosomes with telomeric or internal sites of hybridisation.

RESULTS

Analysis of human sequences in Hutel 1

We first analysed in detail a clone containing the human sequences which conferred telomere function on a telomere-deficient YAC vector. Briefly a modified YAC vector, which had telomeric sequences at one end and a cloning site at the other end, was ligated to human DNA enriched for telomeric sequences and transformed into yeast. Recombinant YACs with human telomeric sequence were maintained as linear chromosomes. The construction of such a yeast library and the isolation of Hutel 1 is described elsewhere (2). The human telomeric sequences in Hutel 1 were subcloned and the clone named pHutel-2-end. A restriction map is shown in Fig. 1. Probing of human DNA's with the XhoI-PstI (X-P) fragment of this clone showed that it contained a low copy number repeat (Fig. 2a).

Tissue and developmental differences

When the X-P fragment of pHutel-2-end was used to probe a Southern blot of various digests of blood and sperm DNA from the same individual a number of differences were apparent in the discrete fragments detected (Fig 2a). When digested with Sau3A some hybridizing fragments in the sperm DNA appear to be increased in intensity relative to blood DNA. With RsaI although basically the same hybridization pattern is detected in blood and sperm DNA digests, a group of fragments are larger in the blood digests. A similar phenomenon is seen with TaqI, although the effect is less marked. The same differences between blood and sperm DNA have been found with all individuals examined so far. The three restriction enzymes Sau3A, RsaI and TaqI all share a common feature; they either contain a CpG dinucleotide in their recognition site, as in the case of TaqI, or they can overlap a CpG sequence, as with Sau3A and RsaI. Most cytosines present in the CpG dinucleotide are methylated in mammalian genomes (16, 17). Cleavage by the restriction enzyme Sau3A is blocked by methylation whereas cleavage by its isoschizomer MboI is not. Hybridization of Sau3A and MboI digests of blood and sperm DNAs, shown in Fig. 2b, demonstrates that the differences in hybridization pattern seen with Sau3A are due to methylation as no differences are seen in the hybridization patterns of the blood sperm DNA digests with MboI. The sequences in sperm are less methylated than in blood. The restriction enzymes TaqI and RsaI are reported to be insensitive to methylation (18, 19). The same differences are

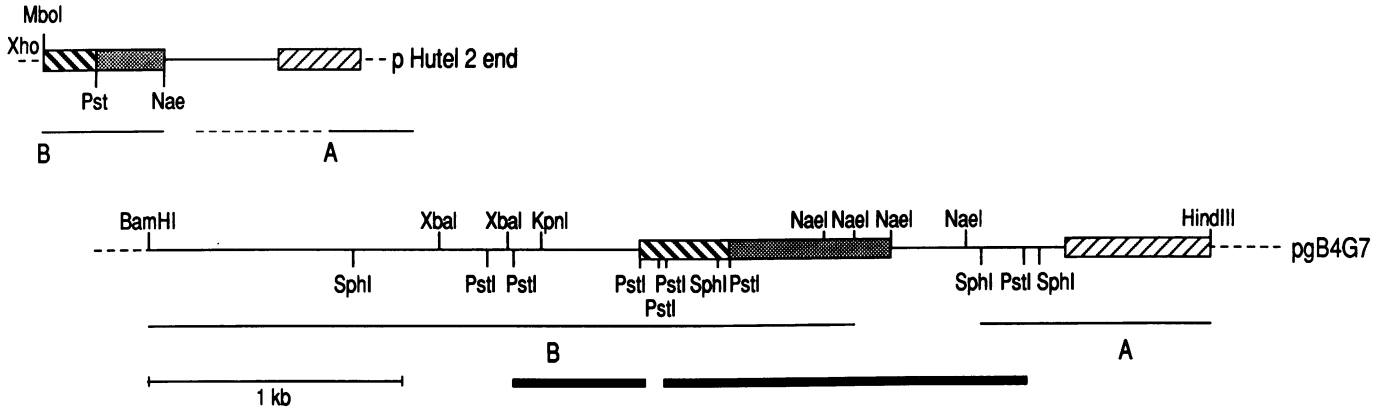


Figure 1. Restriction maps of pHutel-2-end and pGB4G7. Restriction maps were derived by restriction enzyme digestion, end labelling and partial digestion and were confirmed by sequence data. Hyphenated regions of the map are vector DNA. Shaded boxes represent regions of the two clones which cross hybridise, the telomeric ends of the clones are to the right of the figure. Sequenced regions are underlined and identified to correspond to the sequences in Fig. 4. The dashed line represents regions of GC rich repeats not sequenced but restriction site mapped. Bold lines represent the region of pGB4G7 used for *in situ* hybridisation probes.

also seen when the isoschizomer of TaqI, TthHB81, is used. TtHB81 is also reported to be insensitive to methylation.

To investigate the tissue specific differences seen in the hybridisation patterns of the X-P probe and to address the question of whether this difference was germ line specific, we have analysed a number of tissues from a single 16 week fetus, including a sample enriched for ova. The ova and placental DNA exhibit a pattern similar to that of sperm in Fig. 2a but other tissues exhibit a pattern like that of blood. $Ag^{2+}/CsSO_4$ density gradient fractionation of sperm and blood DNA (see below) demonstrates that these differences are in the copies of the subterminal repeat which are TTAGGG linked and thus immediately subterminal (data not shown).

Analysis of clones with subterminal homology to Hutel 1

In order to obtain a preliminary idea of the distribution of the subterminal sequences represented in pHutel-2-end on human telomeres we hybridized the X-P fragment from this clone to a collection of yeast clones selected as containing small linear artificial chromosomes and having terminal $(TTAGGG)_n$ homology. Of 58 colonies tested by colony hybridisation 29 were positive. Analysis of five clones showed that although the fragment patterns were similar they were not identical and therefore were likely to have been derived from different chromosomes. This analysis also showed that the TTAGGG array was abbreviated to the same extent irrespective of whether the human telomeres were derived from blood or sperm. For comparative purposes we have cloned a number of homologous regions from other clones and the complete segment of human DNA from yeast clone B4G7 as pGB4G7. A restriction map of this clone is shown in Fig. 1 and the regions which share homology with pHutel-2-end are indicated.

Sequence analysis of human telomere clones

Terminal repeats. We have sequenced the ends of pHutel-2-end and pGB4G7 clones from the added linker on the distal end. The Bal 31 digestion step in the cloning has removed varying amounts of the yeast telomeric sequences. Nevertheless both sequences start with the yeast terminal repeat and the first base which is unambiguously human (A) is preceded by G residues in both cases. There are a considerable number of variants of the basic TTAGGG repeat present and these variants occur in clusters. This is in contrast to the terminal repeats reported by De Lange *et al.* (9) which contained only TTAGGG repeats. Replication slippage on a small scale could account for such local spreading of variants introduced initially by divergence as single variant repeats.

Subterminal sequences. The subtelomeric sequences of two telomeres are shown in Fig. 3. The striking feature of pHutel-2-end is the presence of a region of 85% GC content composed of 27 repeats of a 28 bp sequence. This region of the clone contains restriction sites for several rare cutter enzymes and is similar in structure to many minisatellite arrays (20). We have been unable to sequence this region entirely but the presence of the repeat has been confirmed by restriction mapping. It is similar to that reported to be adjacent to TTAGGG repeats in other telomeric clones (9). The remaining sequence derived from this clone is not found in the data bases. There is considerable internal repetition within this region; for example the region of 54 bp repeats underlined in the sequence of pHutel-2-end. The corresponding region of pGB4G7 contains a related repeated motif (underlined) with a comparable periodicity. Figure 4 shows a dot plot analysis of pGB4G7, starting in this region. Database searches with the sequences derived from pGB4G7 only reveal the presence of an Alu sequence at positions 2369 to 2685 in segment B. No other homologies were detected. The telomeric repeated clone described by Cheng *et al.* (8) shows similarities in organisation to pGB4G7 and may be a member of the same family. We have sequenced shorter regions of two other telomeric clones which hybridise to the X-P fragment of pHutel-2-end and used the 'Bestfit' program of the UWGCG sequence analysis package (21) to compare these sequences to the longer sequences from pHutel-2-end and pGB4G7. Related sequences are found in the other telomeric clones. The degree of relatedness varies from clone to clone but the four sequences fall into two classes; within a class there is 98% or greater similarity between members, between classes there is 78% similarity. Several sequence duplication events with the possibility of intervening alterations in subrepeat content would account for this observation.

Localisation

We wished to know which telomeres carried this sequence family and which did not. If all human telomeres have this sequence then it might be of functional significance and might be expected to be conserved in other species. We have then asked if at a cytogenetic level these repeats are localised to telomeres and to which. Secondly we have asked if the distance of these repeats

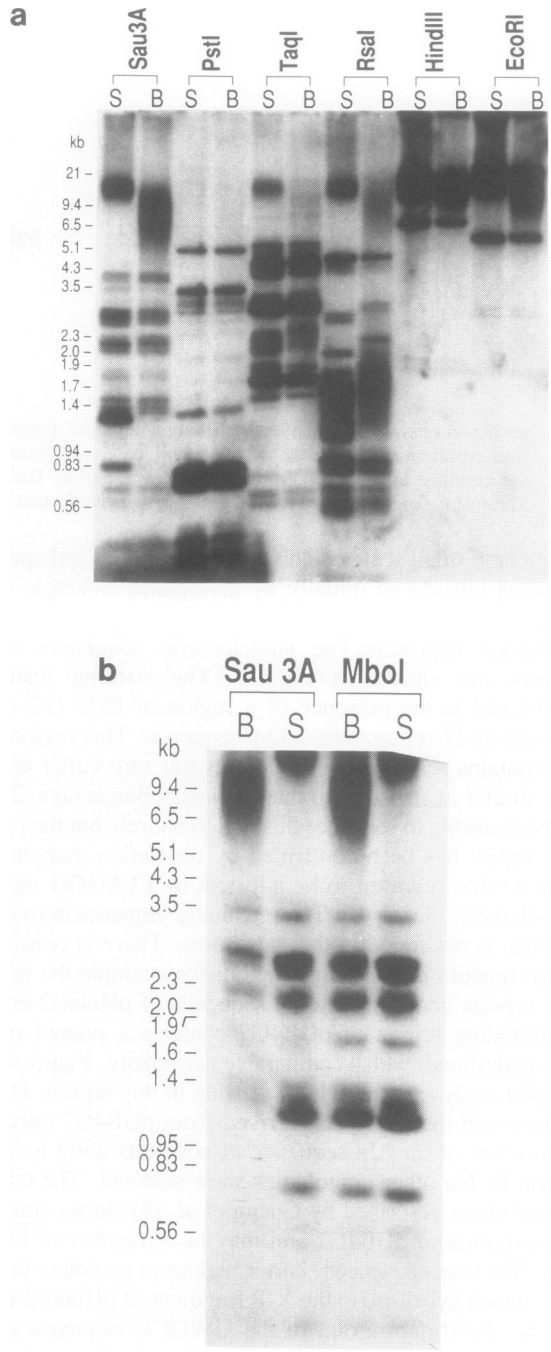


Figure 2. Hybridisation of the X-P fragment from pHutel-2-end. a) Hybridisation to restriction enzyme digests of blood and sperm DNA from a single individual. b) Hybridisation of the X-P fragments from pHutel-2-end to Sau3A and MboI digests of sperm and blood DNAs.

from the ends of the chromosomes can be measured at a molecular scale.

In situ hybridisation. For *in situ* hybridisation we have used three fragments from pGB4G7, indicated in Figure 1, which were known by sequencing and/or blotting to be free of both dispersed and terminal repeats. After labelling and *in situ* hybridisation, sites were scored which showed a signal at the same location on both chromatids. The cumulated sites on 108 spreads of three individuals are shown in Fig. 5a–c. There are differences in the localisation of hybridisation signal between individuals. For instance individual a has a strong signal on 8p and no detectable signal on 21p or q, individual b has a much weaker signal on 8p and strong signals on 21p and q. Given that exchanges between different telomeres occur in yeast this observation raises the strong possibility that this polymorphism of location of human subtelomeric sequences arises through a similar mechanism. With this probe one non-telomeric site at 2q13 is detected. This is the site of putative fusion of the two acrocentric ape chromosomes which gave rise to chromosome 2 (22). This site was not detected by *in situ* hybridisation with a terminal repeat probe (not shown). We conclude that this sequence is present on at least 50% of human telomeres and that it is found only at a limited number of cytologically internal sites. The variation in signal strength at individual telomeres could represent either variation in copy number of the overall repeat or of sub-repeat sequences within a repeat. Alternatively, there may be more divergent copies of this repeat not represented in our small sequenced sample. Telomeres of the sex chromosomes were not detected by *in situ* hybridisation. This finding was confirmed by Southern blot analysis of hybrid DNAs (not shown). Southern blot analysis of a telomere clone from chromosome 4p (23) confirmed that this is a site for a single copy of this sequence family (data not shown).

Molecular localisation. Although some of these repeats are contained in the YAC clones and are therefore undoubtedly immediately subterminal, we do not know the distance to the telomere of all sequences which give an apparently terminal signal in *in situ* hybridisation. It is difficult to use Bal-31 digestion to examine sequences at greater than 15 kb from the terminus, as its action becomes asynchronous with increasing time of digestion. As an alternative, we have exploited the use of $Ag^{2+}/CsSO_4$ gradients as a crude mapping tool. These gradients separate TTAGGG containing fragments from the bulk of the DNA (3). For the purpose of analysis these gradients were divided into three fractions corresponding to the peak of TTAGGG hybridisation (I), the denser half (II) and the lighter half (III) of the main band peak. An equal amount of DNA from each

a

Terminal repeats of Hutel-2-end

$(C_{1-3}A)_{74}(CCCTTAA)_3(CCCTAA)_6CCCCTAACCCTA(CCCTA)_2C(CCCTGA)_3$
$(CCCTAA)_7CCCTA(CCCCTA)_3A(CCTAA)2CCCTGA(CCCTAA)_7CCCTGACCCTACCCCTACCCCT$
$GACCCTACCCCTGACCCTA(CCCTGA)_2(CCCTAA)_{18}$

Segment A

```
1   GGCAACTCCA CCCTCGCAAA GGCGGCGCGG CCGCGCAGGC GCAGAGAGG
                                      *
51  GCGGCGCGGC CGCAGGCGCA GAGACGGGGA GGCGCAGAGA GGCGCGGCGC
          *
101 GGCCCGAGGC GCAGAGATGG GGAGCGCGGC CGCCGAGGCG CAGAGAGGCG
      *                                 *
151 CGGCGCGGCC CGAGGCGCAG AGATGGGGAG GCGCGGCCGC G
            *
```

Segment B

```
      1  AGAGATGGAG AGGCGCCTCG CAGGGCATGA CGCACGCCGG CGCATCCCGG
     51  AGGGAGTGGT GGCTGGGCGG GCGGGATTGA GGCGCGGTGC AGGCGCAGAG
    101  ACGTACGTCG CTGGGCTGAG GGTGGCGGGC CGTGTTGCAG GCGCACAGTC
    151  GGATGGCGAC AGGCGGGGCG CGCGGGGGTG GCACGGTGAA GTGTAGAGAC
    201  ACACGTCCCC GGCGGCGCGG CGCAGAGACG GGTGGAACCT GAGTAATCTG
    251  AAAAGCCTTT TTCGAGCGCC CCCTGCTTGC AGCCGGGCAC TACAGGACCC
    301  AGTTGCACAC GGTGCTGTGC CATTACGCCC CCTGCTGGGC GACTAGGGCA
    351  ACTGCAGGGC CCTGTAGGTG ACAGTGGTGT CCAGCGCCCA CTGCTGGCGC
                                        ----------------------
    401  CGGGGCACGG CAGGGCTCTC TTGCTTGCAG TATAGTGGCG GCACCGCCTG
         ------------------------       ----------------------
    451  CTGGCAGCTA GGGACATTGC AGGGCCCTCT TGCTCACATC GTAGTGGCAG
         ------------------------------------  ---------------
    501  CACGCCCGCC TGCTGGCAGC TGCGGACACT GCCGGGCCCT CTTGCAGACA
         ----------------------------------------------------
    551  AGGAGGTAAA ATAACATTCT ACAAAATGCC TGACCAATCC TCCTCAGTAC

    601  TATCAAGATC
```

b

Terminal repeats

$(C_{1-3}A)_{74}AA(CCCTAA)_{10}(CCCGAA)_{10}CCCTAA(CCCGAA)_5$

$(CCCTAA)_6(CCCGAA)_{12}$

SEGMENT A

```
   1  CCCTCATTAT TCTCAGCTGC AAAGAGGAAG GATCTTTACC GTGGATTTGG
  51  CCCCAGTTGT CCAAAATGAA GCAGTGCCCC AACGTCTGTG GAGAGGCATG
 101  CGCTGCTCCA CCTTCGCGAT GTCCCCCGCG TCTGTGCTGA GCAGAATGCA
 151  GCTCCGTCAT CGCGTTCCCC CCGAAGTCTC TCGAGAGGAA AACGGAGCTC
 201  CTCCTTCGCG ATGCTCTCCA GGTCTGCGCT GAGGGGAACG CAGCTCCGCC
 251  CTCGCAAAGG CATAGCGCCA TCGCAGGCGC AGAAAAAAAC GTCGGTGCAG
 301  CGCAGGCGCA GAGAAAAACG ACGGCGCGTC CCTGGGGGGC GCGGCGCAGG
 351  CCCAGAGAGG CATGCCACCG TGGCGCCGGG GCGTGGGGCG CGACGCAGAC
 401  GAGAGACGCA CGCCGGCGCG GCGCCGGTGA TGGGAGCACG GCCCAGGCGC
 451  AGACACGGAC GGCAGCGTGG CGCCTGGCCG GAGGCGCACG CAGGCCCAGA
 501  GATACACGGC GGCGCGGTGC GCCATGATGG G
```

SEGMENT B

```
   1  CCAGCGCGGC GCCGGCGCAA AGACGGGCGC AGGCGCAGAG TCGGGCGCTG
  51  GCGCGTCGCC GAGGTGGGGG CGCGATGCAC GCGCAGAGAC GCACGGCTGC
 101  GTGGCGCAGA CGCAGAGAAG AACGCGAGCG CGGCGCCGAG GACAAGGCGC
 151  GGGTGCGGAG ACGCACGCCA GCGCGGGGGC GAGGCGCAGG CGCGGAGATG
 201  CACTCCGCCA GGCGCGGGGA GGGGGGCGCG GCGCAGGCGC AGTGACGCAC
 251  GCCGCCTGGG GCCAGCGCAG AGATAGGCGG AACCTCAGTA ATCTGAAAAG
 301  CCAGGTTGCC CCCTCCTTGC GGCCGGGCAC TAAAGGGCCC ACTTGCTGAA
 351  GGCGCTGTGC CAGCGTGCCC CCTCGTGGTG ACTGGGGCAA CTGCAGGGTT
 401  CTCTTGCTTC CATTAGTGGC CAGCGGCCCT GCTGGCGGCG GGGCACCGCA
 451  GGGTCCTCTT GCACACAGTA TAGTGGCGGC ATGCCGCCTG CTGGCAGCTG
 501  GCGACATTGC AGGGCCCTCT TGCTCATAGT ATAGTGACAG GACGCCCGCC
 551  TGCTGGCAGC TGGGGACACT GCCGGCCACT CTTGCTCCAA GTGTAGTGGC
 601  TGTTGGCTCC CCTGCTGGCA GCTGGGGACA CTGCCGGGCC CTCTTGCTTG
 651  CAGTTTACTG GGGGCACGCC CCCTTCTGGC CGCTTGGGGC ACTACAGGAT
 701  GCTCTTGCTC ACAGTGTAGT GGCAGCTCGC CGCCTGCTGG CAACCAGGGT
 751  ACTGCAGGGT TCTCTTGCTC ATGGTGTGGT GCCCGTCCAC CACCTGCTGG
 801  CAGCTAAGGA CACTGCAGGG CCCTCTTGCT CAGAGTGTAG TCGTCGTACA
 851  CCCCCTGCTG GCAGCTGGGG ACGCTGCCGG GACTTTTGCT GGCACTGTCG
 901  TGGCAGCACA CTACCTGCAG GCAGATGGGG ACTATGCAGG GACCTCTTGT
 951  TCAGGGTGTG AGGGCTGGCA CGCCCTACTG GCCGCCTCCT GCACCACTTA
1001  AAGTCGGAGC GCCAGTTGTT AAGCACCATC AGTTCTGGAA ATTGAAACTG
1051  AAATGGAGCT ATTACTGAGG AGAGTTGATG TCCCAGTTCT TGTCTAACTT
1101  GGAAGAAAGA TTTTTCACCA AGAGGCAGCA AAAACATGGC AGATAACTTC
1151  ATTGAAAACA AATACAGTGT AAAGAGCTTA TTGTAGAATA ATAGGGAGGA
1201  GTGGGCTGAT TGTGCAGGAA AACAGCCTGA GAGTCCTGTG CAGGGAATTT
1251  TATTTTGGAC TTCTTCACAT TTCTGCCTCT GTCTCAAGTC TCCACCTGTT
1301  TTCTTTGTCT GGTTTTCCTG CTACTGCCTT AGGTCCCTGA GTTGCCCCAC
1351  TTAGGCTTAT GGGACCTCCT CACTGTTGGT TGAGGCACAT GTGTGGTGAT
1401  CAATCCGAAT CCACTCTGGT ACCAGGCTCC TTCCCCCCAT CCCAGGCAGG
1451  CTGACAGCGG TCATGTTTCT GCCTACAGCG CCTGCCTATC TCTTTTGAAT
1501  GTCCTTCTCT ACCCTACTCT GTACTTATGG TGCCAGGTTT CTCTTAAGAA
1551  TGTCCCCTTT GTCCTTCTTA TCAGCATGTA GCCAGCAATA TTGTGACATT
1601  TTTACTGCAG AGTGAATGAT GACTGGGGCA TCTTAAATGG AGTTCTGGGG
1651  TGTTTCTTTC TGCATAGGTC CCTCTGCAGT AGTAGTTTCC AAAATACTTT
1701  TGGTAATTTT TAACCTTAAA GTTAACCTTA AAGTTAAGCT AAGTAAAAGA
1751  TTTGCATTAA ATATCTAGAC CATTTATAAA TAAGATACAA TACTAAAACA
1801  TTACTGAAGA TAAATAATTC AAGTTTACAT ACTTTTGGCT ACTTATTTTT
1851  ACAGAGAAAC TAAAGATATT TTAGCCCATT AATAAACATG TTTTTGTCTA
1901  CCACACTGAG AAATTGTACT ATGAGGAAAC ACATCCCTCT AGATGTTGGG
1951  AGATGGTATA CTCATACATT TTCTAACCTA CTATAGAATG CTAACATATG
2001  ACAGTTTATA ACTGTCTACT TCCTAGTTTT CTCTGGAAAA TAAAAGATTA
2051  CTAAGTATTA AAATTATAAT CAATATGTGT AAATAAAACT ACTGGAAATA
2101  ACAGAATAAC TAGAAACAAC TCTATGCAAA GCATGCAAGA AAAGTAGTGC
2151  ATGTTTTGCA AGTAAAGTAG GACGTATTTT TTATAAGGAA AACCATACAA
2201  AAGATACAAA TAAAAAGAGA TACCTAGCCT TCCCTGTGTT ATATTTGTAT
2251  GGGTAAAATG TCATGTTTTC AGAAATTATA TAAAATTCCT GGAAATTTGT
2301  CAATGTTCTC CTTATCCATG CTATGTGCCA GTATAGAGTT ATGAGTCATA
2351  ATTCCAATTA TTATTTTAAA TGTTGTGCTG GGTGCAGTGG CTCACGTCTG
2401  TAATCCCAGC ACTTTGGGAG GCCTAGGCAG GTGGATCACA AGGTCAGGAG
2451  ATCGGAGAAC ATCCTGGCTG ACATGGTGAA ACCCCATCTC TACTAAAAAT
2501  ACAAAAAATT AGCCAGGCGT GGTGGTGGAC ACCTGTACTC CCAGCTACTC
2551  AGGAGGCTGA GGCAGGAGAA TGGCATGAAC CAGGGAGGCA GAGCTTGCAG
2601  TGAGCCAAGA TAGCGCCACT GCACTCCAGC ATGGGCAACA GAGCGAGACT
2651  CTGTCTCTAA ATAAATAAAT AAATAAATAA ATGTTGTATC CCACAGAAAA
2701  AATCGAATAT CCTTGTCAGT TGTGGTATAA TGAACTCTCA TCAGGTCTTT
2751  CATCACAGCC ATTTCATATT CTTTATCATT TAGATATTAT TTCCCCCTGA
2801  TGCTTTCCTG AAAGCTCCTG CAATCAGCTA CAGGTCAGAA TGTTCATCTC
2851  CATCACGGGA TTCCCTCTGA GACACACAGA AAAGAGTATG CAAGATAGTC
2901  TGGTTATAGG CTTCTGATGA TATTGTTTAA ATAACTTTAA GACCATACAC
2951  TTCGCTCAGT GAAGATCTCC AGAAGTCTGC TTCAGAAATT GATGGGTTCA
3001  TGACACTGCT AACCCAAGAT GCAACAAGAC TGGAATTGAT TACATGGTAC
3051  TGAATGAACT GATGAAAATT GATTATAATT TTATAGCTTT TTGGAGCATT
3101  GCTGGTTCTT TAATGTTCTA GTTTCTGGAC TTAAGAAATC TCTTTCTCTT
3151  AACCTAACTG TAACATACAA TTTAGTAGAT TATACTTTTG AAAACAGAAG
3201  TGAAGCATTT ATCTTTTTTC CCCTGCCTGA TTTTTCCAGA ATTTTGAAAT
3251  CCTTACTGAA CACTCTTATT TTCACGATGA TATAGTTGTT AGAAAAGTCC
3301  AATAAGAATC TGTTCACCTT GAACAGAGAC CTCAGAAATA ATGCCGCATA
3351  TCTACAACCA TCTGATCTTT GACAAACCTG ACAAAAACAA GCAATGGGAA
3401  AGATTCCCTA TTTAATAAAT GGTGCTGGGA AAACTGGCTA GCCATATGTA
3451  GAAAGCTAAA CTGGATCC
```

Figure 3. Sequence of telomeric clone. a)Sequence derived from pHutel-2-end. The sequence is presented 5′−3′ starting with the most centromere distal region of the clone. Segment A contains the minisatellite array one element of which is underlined and the start of subsequent elements asterisked. Repeated elements in segment B are underlined with dashes. b)Sequence derived from pGB4G7, repeats in segment A are underlined with dashes and those in segment B with continuous underlining. The Alu repeat in segment B is underlined with dashes. EMBL/GenBank accession numbers for these sequences are X56275 to X56278.

fraction was loaded onto a gel. Hybridisation with the X-P fragment of pHutel-end-2 reveals strong hybridisation to fraction I with no discrete bands visible (Fig. 6). Secondary digests of this fraction with RsaI or Sau3A gave the characteristic set of fragments. This is the expected result if most copies of this sequence are distal to the last BamHI site before the chromosome end. Pulsed field gel analysis of a BamHI digest of human DNA probed with TTAGGG gives a size distribution with a mean of approximately 25 kb. About 5% of the hybridisation signal is fractionated in the main band consistent with the *in situ* data.

Internal sites

Although the majority of sequences homologous to the X-P fragment of pHutel-end-2 are located subterminally there are

some internal sites. An attempt to isolate sequences from one of these sites was made by screening a partial Sau3A human DNA library cloned in lambda EMBL3 with the X-P fragment. Twelve positive clones were identified. One of the most strongly hybridising was analysed in detail. A restriction map of this clone is shown in Fig. 7. It contains approximately 1.2 kb of DNA homologous to pHutel-2-end and to pGB4B7 and this DNA is organised in a similar manner. The restriction map was confirmed

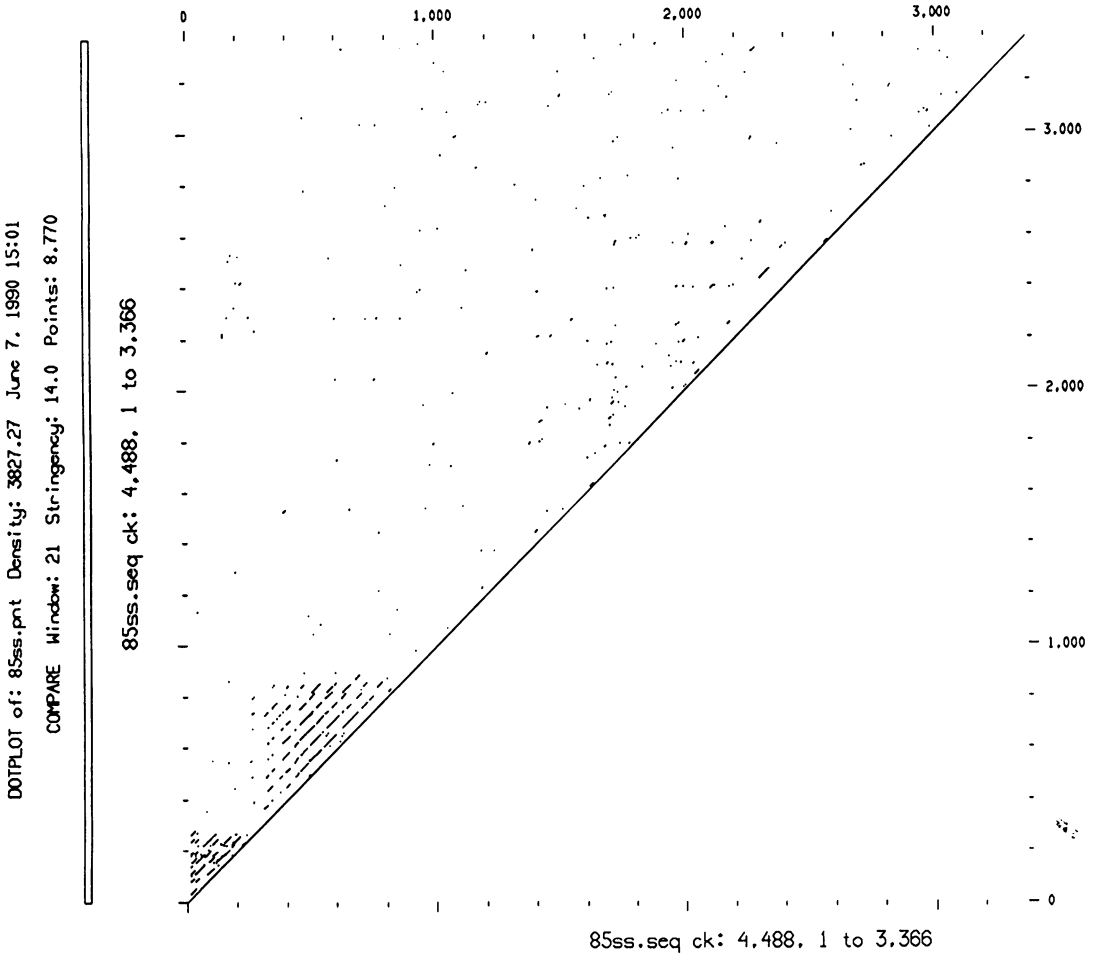
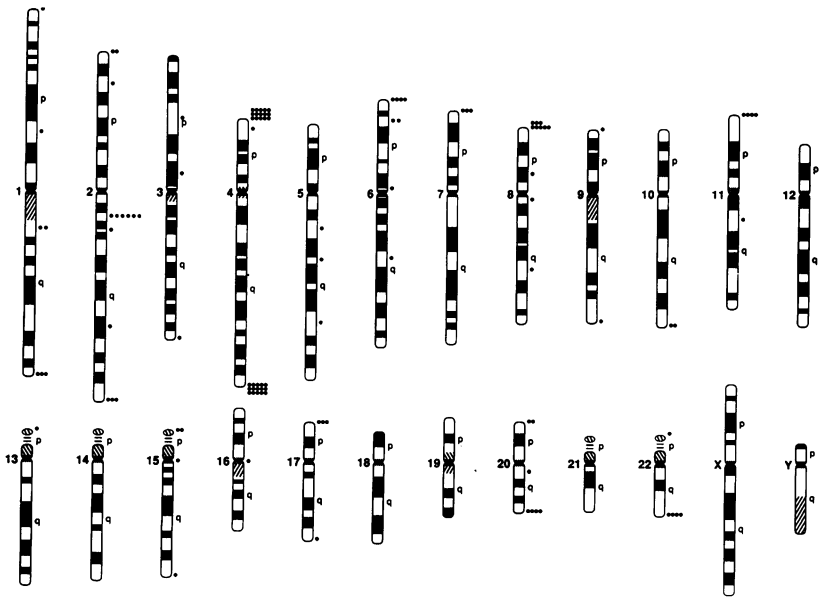


Figure 4. Dot plot self comparison of segment B of pGB4G7 with a stringency of 14 and window of 21.



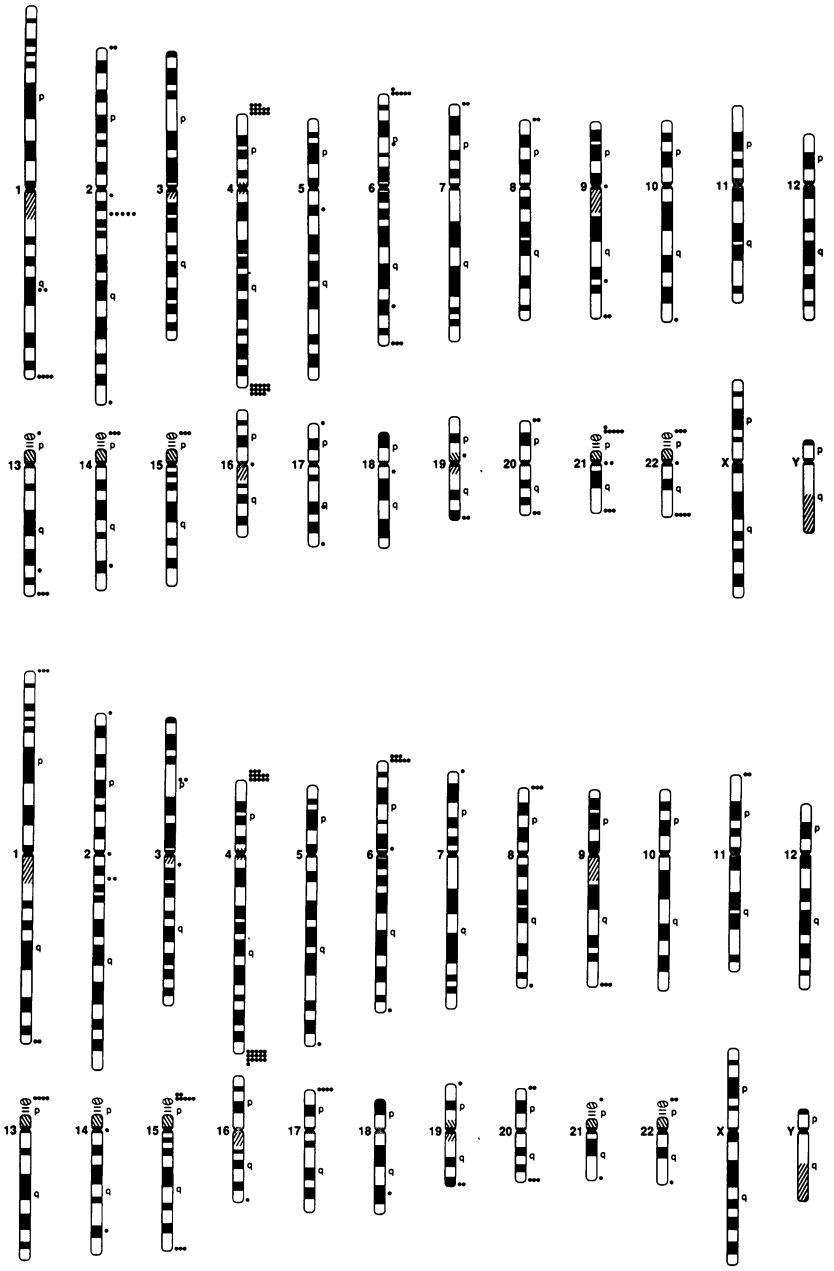


Figure 5. *In situ* hybridisation of pooled fragments of pGB4G7 to metaphase chromosomes of three caucasians.

in genomic DNA to rule out the possibility that the clone represented a cloning artefact. 800 bp of this region was sequenced and compared to pHutel-2-end and pGB4B7 using the Bestfit program of the UWGCG sequence analysis package. The sequence was approximately equally homologous to both pHutel-2-end and pGB4B7. $(TTAGGG)_4$ hybridisation detected only a weak signal adjacent to the subtelomeric homologous region. Sequencing of this region revealed no TTAGGG repeats although there were a large number (22 in 800 bp sequenced) of imperfect 4/6 bp matches to TTAGGG which may have accounted for the weak signal. The fact that the region homologous to the subterminal sequences is located 9 kb from the end of a clone which contains no TTAGGG repeats suggests that this DNA is not immediately subterminal. This was confirmed by probing BamHI digested human DNA with a fragment from this clone outside the X-P homology region. A discrete fragment was detected indicating that this DNA is proximal to the last BamHI site before the chromosome end. We have not yet assigned this clone to a unique site by *in situ* hybridisation because of repetitive sequences present in the clone.

Conservation

Hybridisation of the X-P probe from pHutel-2-end to DNA of a number of species showed that only primate genomes contained closely related sequences. This in contrast to the terminal repeat which is found in many species (24).

DISCUSSION

We report here on the structure, location and developmental changes observed in a class of human sequences isolated as being adjacent to $(TTAGGG)_n$ repeats by complementation of telomere function in yeast and on the structure of the (TTA-

$GGG)_n$ repeats themselves. We have previously shown that some of these sequences (subterminal in yeast clones) are indeed immediately subtelomeric in the human genome (2). It is clear from the analysis reported here that more than 95% of these sequences are localised to within 30 kb of the end of the chromosomes. The internal site, detected by *in situ* hybridisation, of these sequences on chromosome 2 has also been reported to be a site of hybridisation for the Tetrahymena terminal repeat TTGGGG (25). Our studies suggest that TTAGGG repeats are not present in large number here. The sequencing data suggest that variant repeats may be concentrated at the proximal end of the terminal repeat array. These observations would be consistent with a process of chromosome fusion in which the terminal repeats were deleted. This process would also account for the presence of sequences related to the subterminal repeats at the internal site which we have characterised.

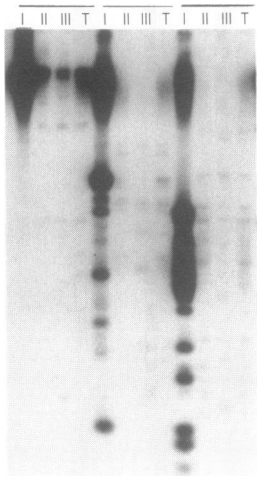


Figure 6. Hybridisation of the X-P fragment from pHutel-2-end to gradient fractions and total human DNA (T). First panel BamHI digested, second panel BamHI/Sau3A digested, third panel BamHI and RsaI digested.

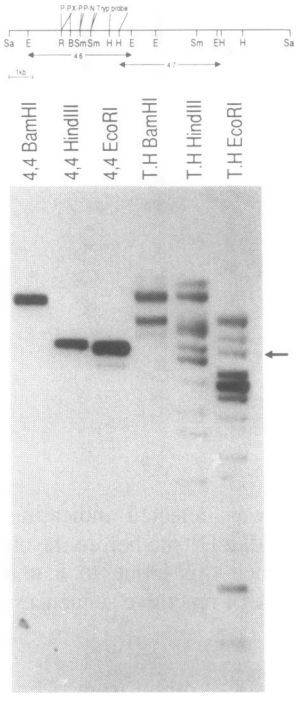


Figure 7. Restriction map of an internal copy of the subterminal repeat family. The restriction map of this region of lambda clone 4.4 is shown with regions of homology to pGB4G7 above and predicted EcoRI and Hind III fragments and their sizes shown below. Sa — SalI, E — EcoRI, R — RsaI, Sm — SmaI and B — BamHI. Southern blot analysis of digests of 4.4 DNA and human genomic DNA (T.H) with the HindIII to EcoRI fragment adjacent to the region of Trypanosome terminal repeat homology used as a probe. Fragments of the predicted size in the human DNA are arrowed.

Sequencing of the termini of two yeast clones shows that the structure is largely as expected (26, 27). The most terminal sequence, which before Bal 31 digestion probably amounted to about 200 base pairs, consists of yeast terminal repeats of the $(TG_{1-3})_n$ type and this is followed immediately by the $(TTAGGG)_n$ sequence expected of human terminal repeats. We presume that the human sequence has been recognised as a substrate for the mechanism responsible for elaboration of telomeres in yeast, probably on the basis of strand assymetry in base composition and orientation of the G rich strand towards the chromosome end. The tracts of TTAGGG and closely related repeats in the primary yeast clones are about 500 bp in length. Since we and others (2, 3, 25) have previously shown that the terminal restriction site barren region of human chromosomes consists of 5−15 kb of DNA in blood and 15−20 kb in sperm all but the most proximal sequences must have been lost in the cloning process. The human telomeric repeats were abbreviated to the same extent irrespective of whether they were derived from blood or sperm. Given that a telomerase-like activity has been isolated from a human cell line which specifically adds TTAGGG (28), it is most likely that the whole of this missing DNA is composed of TTAGGG repeats. This loss of sequence could be due to either the yeast cell being unable to replicate a long stretch of repeats or could reflect the relative probabilities of degradation and protection from degradation by addition of a TG_{1-3} repeat and subsequent elaboration of a yeast telomere. The region of repeats sequenced would presumably not have been recently added by a telomerase activity in human cells and divergence could account for the TTAGGG related repeats which we have observed. Hybridisation experiments by Allshire *et al.* (29) suggested that extensive arrays of TTAGGG-related repeats are found adjacent to the terminal TTAGGG repeats. Such arrays are not present in the human telomeric sequences which we have examined. Although there are some occurrences of repeats such as TTGGGG they are not sufficient to be detected by hybridisation using a $(TTGGGG)_4$ oligonucleotide probe (data not shown).

Sequencing of the internal regions of yeast clones with homology to sequences from pHutel-2-end reveals that there are at least two subclasses of a repeated sequence representing either a duplication event or a correction mechanism which operates between a limited number of sets of telomeres. Telomere exchange followed by conversion is one possible example of such a correction mechanism but we have not observed such an exchange. Differences between these subtelomeric sequences arise through variations in the number and arrangements of a number of subrepeats.

Our *in situ* hybridisation data show a variable number of sites for these sequences in different individuals. This polymorphism could arise by a change in the copy number of the subterminal

repeats by an inter or an intra chromosomal mechanism. Most subterminal repeat sequences are within a few tens of kb from the ends of the chromosomes and a long telomere clone from chromosome 4p contains only one copy of this sequence. If this pattern of distribution is representative of all chromosomes an intra chromosomal mechanism seems unlikely and telomere exchange between chromosomes is probably the cause of the polymorphisms which we have seen. About 50% of telomeres hybridise *in situ* with the probe used, a figure comparable with the proportion of clones which hybridise to this probe in our telomere library.

There is no compelling evidence for function in the region of subterminal repeats which we have studied. Their lack of conservation between chromosomes and species argues against a functional role. There is a short region of the subterminal repeat which is unmethylated in both male and female germ lines. These sites are subsequently de novo methylated in somatic tissues by 16 weeks of development. We have no explanation for the differences seen with enzymes such as Taq I. Possibly a methylation induced conformational change is affecting the ability of the enzymes to cleave the DNA at these sites. The *de novo* methylation observed is reminiscent of that found in satellite DNA sequences. We find no evidence for other changes during development in these subterminal sequences. We have previously reported the loss of terminal repeats during formation of somatic tissues (5). Comparably positioned subterminal repeat sequences in some species are exchanged between chromosomes. Analysis of unrelated individuals at the level of *in situ* hybridisation suggests that in humans the location of these sequences is polymorphic but the Mendelian inheritance of polymorphisms revealed by these sequences (not shown) implies that this does not happen at a high frequency in humans.

The major conserved feature of mammalian chromosomes remains the terminal repeats themselves (24). *In vivo* studies of cloned mammalian telomeres are underway and the construction of mammalian linear episomes with telomeres will answer the many outstanding questions about their function.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Council. We would like to thank those who donated tissue samples, C. Gosden for help with foetal material and all who contributed by discussion and criticism of the manuscript.

REFERENCES

1. Blackburn,E.H. and Szostak,J.W. (1984) *Ann. Rev. Biochem.* **53**, 163–194.
2. Cross,S.H., Allshire,R.C., McKay,S.J., McGill,N.I. and Cooke,H.J. (1989) *Nature* **338**, 771–774.
3. Brown,W.R.A. (1989) *Nature* **338**, 774–776.
4. Moyzis,R.K., Buckingham,J.M., Cram,L.S., Dani,M., Deaven,L.L., Jones,M.D., Meyne,J., Ratliff,R.L. and Wu,J.R. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6622–6626.
5. Cooke,H.J. and Smith,B.A. (1986) *Cold Spring Harbor Symp. Quant. Biol.* **51**, 213–219.
6. Cooke,H.J., Brown,W.R.A. and Rappold,G.A. (1985) *Nature* **317**, 687–692.
7. Inglehearn,C.F. and Cooke,H.J. (1990) *Nucl. Acids Res.* **18**, 471–476.
8. Cheng,J.F., Smith,C.L. and Cantor,C.R. (1989) *Nucl. Acids Res.* **17**, 6109–6127.
9. De Lange,T., Shiue,L., Myers,R.M., Cox,D.R., Naylor,S.L., Killery,A.M. and Varmus,H.E. (1990) *Molecular and Cellular Biology* **10**, 518–527.
10. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) N.Y., Cold Spring Harbor, *Molecular Cloning: A Laboratory Manual.*
11. Cooke,H.J. and Cross,S. (1988) *Nucl. Acids Res.* **16**, 11817.
12. Lathe,R., Kieny,M.P., Skory,S. and Lecocq,J.P. (1984) *DNA* **3**, 173–182.
13. Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.* **132**, 6–13.
14. Fantes,J., Gosden,J. and Piper,J. (1989) *Cytogenet. Cell Genet.* **48**, 142–147.
15. Pinkel,D., Straume,T. and Gray,J.W. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2934–2938.
16. Bird,A.P., Taggart,M.H., Nicholls,R.D. and Higgs,D.R. (1987) *EMBO Journal* **6**, 999–1004.
17. Bird,A.P. (1987) *TIGS* **3**, 342–347.
18. McClelland,M. and Nelson,M. (1988) *Gene* **74**, 291–304.
19. Nelson,M. and McClelland,M. (1989) *Nucl. Acids Res.* **17S**, r389–r415.
20. Simmler,M.C., Johnsson,C., Petit,C., Rouyer,F. and Vergnaud,G. (1987) *EMBO Journal* **6**, 963–969.
21. Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucl. Acid. Res.* **12**, 387–395.
22. Yunis,J.J. and Prakash,O. (1982) *Science* **215**, 1525–1529.
23. Bates,G.P., MacDonald,M.E., Baxendale,S., Sedlacek,Z., Youngman,S., Romano,D., Whaley,W.L., Allitto,B.A., Poustka,A., Gusella,J.F. and Lehrach,H. (1990) *Am. J. Hum. Genet.* **46**, 762–775.
24. Meyne,J., Ratliff,R.L. and Moyzis,R.K. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7049–7053.
25. Allshire,R.C., Gosden,J.R., Cross,S.H., Cranston,G., Rout,D., Sugawara,N., Szostak,J.W., Fantes,J. and Hastie,N.D. (1988) *Nature* **332**, 656–659.
26. Shampay,J., Szostak,J.W. and Blackburn,E.H. (1984) *Nature* **310**, 154–157.
27. Szostak,J.W. and Blackburn,E.H. (1982) *Cell* **29**, 245–255.
28. Morin,G.B. (1989) *Cell* **59**, 521–529.
29. Allshire,R.C., Dempster,M. and Hastie,N.D. (1989) *Nucl. Acids Res.* **17**, 4611–4627.