Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q

J. Guy1, C. Spalluto1, A. McMurray3, T. Hearn1, M. Crosier1, L. Viggiano2, V. Miolla2, N. Archidiacono2, M. Rocchi2, C. Scott3, P.A. Lee4, J. Sulston3, J. Rogers3, D. Bentley3 and M.S. Jackson1+

1Human Genetics Unit, School of Biochemistry and Genetics, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK, 2DAPEG, Sezione di Genetica, Università di Bari, Via Amendola 165/A, 70126 Bari, Italy, 3Sanger Sequencing Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and 4Department of Computer Science, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK

Received 24 April 2000; Revised and Accepted 22 June 2000 DDBJ/EMBL/GenBank accession nos AL022344, AL022345, AL031601 and AJ290103–AJ290167

The organization of centromeric heterochromatin has been established in a number of eucaryotes but remains poorly defined in human. Here we present 1025 kb of contiguous human genomic sequence which links pericentromeric satellites to the RET proto-oncogene in 10q11.2 and is presumed to span the transition from centric heterochromatin to euchromatin on this chromosome arm. Two distinct domains can be defined within the sequence. The proximal ∼240 kb consists of arrays of satellites and other tandem repeats separated by tracts of complex sequence which have evolved by pericentromeric-directed duplication. Analysis of 32 human paralogues of these sequences indicates that most terminate at or within repeat arrays, implicating these repeats in the interchromosomal duplication process. Corroborative PCR-based analyses establish a genome-wide correlation between the distribution of these paralogues and the distribution of satellite families present in 10q11. In contrast, the distal ∼780 kb contains few tandem repeats and is largely chromosome specific. However, a minimum of three independent intrachromosomal duplication events have resulted in >370 kb of this sequence sharing >90% identity with sequences on 10p. Using computer-based analyses and RT–PCR we confirm the presence of three genes within the sequence, ZNF11/33B, KIAA0187 and RET, in addition to five transcripts of unknown structure. All of these transcribed sequences map distal to the satellite arrays. The boundary between satellite-rich interchromosomally duplicated DNA and chromosome-specific DNA therefore appears to define a transition from pericentromeric heterochromatin to euchromatin on the long arm of this chromosome.

INTRODUCTION

Heterochromatin is defined cytologically within the eucaryotic nucleus as densely packed and dark staining chromatin, distinct from the more diffuse and lightly staining euchromatin (1,2). This distinction reflects significant differences in biological activity and sequence organization. Heterochromatin can both repress the expression of genes via position effects (3–5) and suppress recombination in a wide variety of eucaryotes (6). Gene density is also known to be low in the heterochromatin of organisms where detailed mapping (7) or sequence data (8–10) are available. For instance, a combination of pulsed field gel electrophoresis (PFGE) and sequencing indicates that Drosophila heterochromatin is composed of satellite arrays interspersed with retrotransposable elements (9,10). A similar organization has been established in Arabidopsis where centromeric heterochromatin largely consists of a 180 bp satellite repeat and retrotransposons, mostly from the Athilia family (8,11). A heterochromatic knob from the short arm of chromosome 4 in this species has also been completely sequenced, revealing 22.5 tandemly arranged copies of a 1950 bp satellite repeat and retrotransposons, mostly from the Athilia family (8,11). A heterochromatic knob from the short arm of chromosome 4 in this species has also been completely sequenced, revealing 22.5 tandemly arranged copies of a 1950 bp repeat surrounded by DNA rich in retrotransposons (12). Heterochromatic knobs have also been partially characterized in maize and found to have comparable organizational features (13). It is presumed that sequence repetition is responsible for conferring heterochromatic properties, a hypothesis supported by the fact that repetitive sequences are capable of inducing gene silencing in a number of eucaryotes including mammals (14,15).

In human chromosomes the most prominent sites of heterochromatin are the centromeric regions, the large C bands present on chromosomes 1, 2, 9, and 16, and the long arm of the Y chromosome. Consistent with data from other eucaryotes, all of these regions of the genome are rich in tandem repeats (16).

*To whom correspondence should be addressed. Tel: +44 191 222 8005; Fax: +44 191 222 6662; Email: mjackson@hgmp.mrc.ac.uk
To date, a detailed understanding of pericentromeric sequence organization has been hindered by the lack of contiguous sequence data, largely due to repetitive sequences causing clone instability and contig branching during mapping (17). Only the alpha satellite which is present at the primary constriction of all human chromosomes has been analysed in detail, due to its possible role in centromere function (18). Despite this, the basic organization of centromeric repeats has been established on several chromosomes by PFGE and fluorescence in situ hybridization (FISH); in each case a central array of alpha satellite is flanked by arrays of other satellites such as the classical satellites (I–IV) and the beta satellite (16).

However, it is clear that pericentromeric regions of human chromosomes do not consist solely of repetitive sequences. Tracts of gene-related sequences, ranging in size from 60 to 200 kb, have been identified proximal to arrays of satellite II and the 48 bp satellite (D22Z3) on chromosome 22 (19) and at the terminus of the centromeric alpha satellite array on chromosome 16 (20). Some of these sequences have been formed by recent pericentromeric-directed duplication events, examples of which involve the duplication of NF1- and ALD-related sequences (17,21–23). Phylogenetic analysis of human and primate ALD sequences indicates that they have been amplified by a two-step process where the duplication of ~10 kb of DNA from Xq28 into a pericentromeric location has been followed by rapid distribution of this sequence between chromosomes (20). There is some evidence that repetitive sequences are involved in these processes: sequence identities between the ALD gene and its autosomal paralogues terminate within a GCITTTTGCG repeat (21), the genomic distribution of human NF1-related sequences correlates with the position of alpha satellite suprachromosomal family 2 (22), and several pericentromeric duplication events have been shown to terminate close to or within the GC-rich CAGGG and HSREP522 repeats (24). In addition to these interchromosomally duplicated sequences, large intrachromosomal duplications flank the centromeric repeats of several human chromosomes (25–29) which suggests that pericentromeric instability may involve a number of mechanistically distinct phenomena.

We have previously developed a high resolution physical map of the 10p11–q11 region which includes a PFGE map across the centromeric repeats (30) and two yeast artificial chromosome (YAC) contigs which link pericentromeric satellites to genes on both chromosome arms (31). In addition to interchromosomally duplicated sequences, these contigs contain two large (90–250 kb) duplications which flank the centromere (32). To further investigate the organization of these sequences, and to attempt to define human pericentromeric heterochromatin at the sequence level, we have developed 1.025 Mb of contiguous genomic sequence which extends from satellite arrays in 10q11.1 to the RET proto-oncogene in 10q11.2. We have used this sequence to analyse transcriptional activity and to establish the distribution of paralogous sequences within the human genome. The data provide the first integrated picture of mammalian sequence organization, evolution and gene expression at a boundary between pericentromeric repeats and a chromosome arm.

The sequence extend from pericentromeric satellites in the G-dark band 10q11.1 to the RET proto-oncogene in the G-light band 10q11.2 (32). The GC content of the entire sequence is 43.18%. However, it rises steadily from 39.48% in the proximal 200 kb (38.34% after removal of all tandem repeats) to 54.04% in the most distal 100 kb (Fig. 1B), consistent with the general observation that G-dark bands are GC poor relative to G-light bands (36). In total, 48.46% of the entire sequence is accounted for by interspersed repeats, with SINES, LINES and long terminal repeats (LTRs) accounting for 10.60, 24.28 and 10.87% of the sequence, respectively (Fig. 1B and C). These

RESULTS

Sequence acquisition and fidelity

A total of ~1.025 Mb of contiguous sequence was generated using three overlapping YACs as templates for shotgun cloning (33), a technique which has proved invaluable for sequencing repetitive regions within the Caenorhabditis elegans genome (34). The YACs form part of a 10 Mb map of 10p11–q11, have been FISH mapped, and possess internally consistent restriction maps which suggests that they have not rearranged during cloning (31). The position of these YACs relative to the final sequence, and to the centromeric alphaltid array (D10Z1), is shown in Figure 1A. The 1–2 Mb region between D10Z1 and the sequence presented here remains poorly characterized but is known to include an ~1 Mb satellite 2-rich domain (31). Because YACs are known to rearrange at high frequency (35), several steps were taken to ensure the fidelity of the data (see Materials and Methods). Once the sequence was completed, the large overlap between Y738F9 and Y214H10 was compared as a further check on sequence quality. The overlap between these two sequences share 99.92% similarity with a total of 150 nucleotide differences and 25 insertions/deletions (indels) in 231 540 bp of aligned sequence. This figure must be considered to be a conservative estimate of sequence quality as it will include errors in both sequences. Furthermore, a significant proportion of the sequence variation is likely to represent allelic polymorphism. For instance, the two largest indels consist of an 11 bp difference within a poly(A) tract at the terminus of an Alu element (position ~661 330 bp). This conclusion is supported when the sequence is aligned with cDNAs from two genes from the region: RET (GenBank accession nos X15262 and X12949) and KIAA0187 (GenBank accession no. D80009). There are a total of nine differences between the cDNAs and the genomic sequence. However, all are within untranslated regions or result in silent substitutions (data not shown) and one is a known polymorphism (C→A at position 3959 of GenBank accession no. D80009).

GC content and the distribution of interspersed repeats

The sequence extends from pericentromeric satellites in the G-dark band 10q11.1 to the RET proto-oncogene in the G-light band 10q11.2 (32). The GC content of the entire sequence is 43.18%. However, it rises steadily from 39.48% in the proximal 200 kb (38.34% after removal of all tandem repeats) to 54.04% in the most distal 100 kb (Fig. 1B), consistent with the general observation that G-dark bands are GC poor relative to G-light bands (36). In total, 48.46% of the entire sequence is accounted for by interspersed repeats, with SINES, LINES and long terminal repeats (LTRs) accounting for 10.60, 24.28 and 10.87% of the sequence, respectively (Fig. 1B and C). These
figures are typical of human autosomal DNA (37). There is a slight excess of LINEs in the proximal region (28.04% of the first 720 kb is LINE related, compared with the genome-wide average of 20.6% for autosomes with an equivalent GC content) and a more distinct excess of LTR elements [17.3% of the sequence from 200 to 560 kb is LTR related compared with the genome average of 7.8% for autosomes of the same GC content (37)]. Again, these observations appear consistent with the known characteristics of G-dark bands (38). However, the frequency of SINEs is highest within the proximal region which is inconsistent with generalizations based on cytogenetic classifications (38). It is also clear from Figure 1C that LINE and LTR elements are non-randomly distributed, with the existence of 15–45 kb sequence tracts where a single element type predominates.

**Proximal 10q11 is gene poor**

The principal gene-related features identified within the sequence are shown in Figure 1D and Tables 1 and 2. Because gene-related sequences are known to transpose into and between pericentric locations (18,21–24), the 10q11 sequence is likely to contain paralogous copies of human genes/sequence tracts present at other genomic locations (a paralogue being defined here as one of a set of homologous genes/sequence tracts within a single species). As a result, we have been conservative in our definition of what constitutes a
Table 1. Genes, ESTs and confirmed transcripts within 10q11

<table>
<thead>
<tr>
<th>Position in sequence</th>
<th>Gene name/GenBank acc. no./Unigene cluster</th>
<th>ESTs/Spliced ESTs</th>
<th>EST alignment to 10q sequence (%)</th>
<th>Expressiona</th>
<th>Classification (repeats present)</th>
</tr>
</thead>
<tbody>
<tr>
<td>175 778–175 372</td>
<td>AI797613</td>
<td>1 No</td>
<td>100</td>
<td>H B Pl Li M K Pn</td>
<td>EST</td>
</tr>
<tr>
<td>215 058–215 342</td>
<td>AA350311</td>
<td>1 No</td>
<td>99.6</td>
<td></td>
<td>EST</td>
</tr>
<tr>
<td>216 723–216 505</td>
<td>AA350425</td>
<td>1 No</td>
<td>99.5</td>
<td></td>
<td>EST (L1MA3)</td>
</tr>
<tr>
<td>289 697–289 988</td>
<td>HS967220</td>
<td>1 No</td>
<td>99.5</td>
<td></td>
<td>EST (AluSc)</td>
</tr>
<tr>
<td>290 668–290 877</td>
<td>AA724580</td>
<td>1 No</td>
<td>99.5</td>
<td></td>
<td>EST</td>
</tr>
<tr>
<td>292 042–292 484</td>
<td>AA846774</td>
<td>1 No</td>
<td>99.4</td>
<td></td>
<td>EST</td>
</tr>
<tr>
<td>305 014–306 543</td>
<td>HS1251874</td>
<td>1 No</td>
<td>99.7</td>
<td></td>
<td>EST (L2)</td>
</tr>
<tr>
<td>397 054–397 344</td>
<td>NS6650/Hs193883</td>
<td>3 No</td>
<td>98.2–98.4</td>
<td></td>
<td>EST</td>
</tr>
<tr>
<td>398 282–399 270</td>
<td>AA011715/Hs217610</td>
<td>8 No</td>
<td>98–99.6</td>
<td></td>
<td>EST (AluSc/LIM4)</td>
</tr>
<tr>
<td>401 307–406 028</td>
<td>ZNF37B/AI524643</td>
<td>1 Yes</td>
<td>99.8</td>
<td></td>
<td>Pseudogene</td>
</tr>
<tr>
<td>434 919–434 498</td>
<td>AL119437</td>
<td>1 No</td>
<td>99.3</td>
<td>+</td>
<td>Transcript</td>
</tr>
<tr>
<td>455 240–455 431</td>
<td>AI769609/Hs156453</td>
<td>3 No</td>
<td>99.4–99.8</td>
<td>+</td>
<td>Transcript (L1MC2)</td>
</tr>
<tr>
<td>473 256–474 456</td>
<td>AA936524/Hs267204</td>
<td>3 No</td>
<td>98.6–100</td>
<td>+</td>
<td>Transcript (L1MC4/MER115)</td>
</tr>
<tr>
<td>479 488–479 958</td>
<td>R70655</td>
<td>2 No</td>
<td>98.8, 100</td>
<td></td>
<td>Transcript (MER45B)</td>
</tr>
<tr>
<td>520 850–475 875</td>
<td>ZNF11/33B/Hs72991*</td>
<td>88 Yes</td>
<td>96–100</td>
<td>+</td>
<td>Gene</td>
</tr>
<tr>
<td>666 675–713 708</td>
<td>KIAA0187/Hs117980b</td>
<td>≥150 Yes</td>
<td>96–100</td>
<td>+</td>
<td>Gene</td>
</tr>
<tr>
<td>959 238–1 010 390</td>
<td>RET/Hs241572</td>
<td>86 Yes</td>
<td>96–100</td>
<td>+</td>
<td>Gene</td>
</tr>
</tbody>
</table>

Features were identified by using BLAST to query the Swissprot, TREMBL and dbEST databases. They have been classified as follows: genes, ESTs >96% coinciding with either *ab initio* gene prediction or protein similarity; transcripts, ESTs >96% where transcription has been confirmed by RT–PCR but there is no associated *ab initio* gene prediction or protein similarity; ESTs, ESTs >96% where transcription has not been confirmed by RT–PCR and there is no associated *ab initio* gene prediction or protein similarity; gene fragments, protein similarity with intact ORF, no ESTs >96%; pseudogenes, protein similarity with disrupted ORF. Only ESTs with >96% identity over ~80% of their length to the 10q11 sequence are shown. Due to the duplicative nature of this region, the range of percentage identities that the ESTs in each cluster share with the 10q11 genomic sequence is shown. The results of RT–PCR experiments are also shown. The cDNAs analysed were derived from the following adult tissues: H, heart; B, brain; Pl, placenta; Lu, lung; Li, liver; M, skeletal muscle; K, kidney; Pn, pancreas.

---

gene, requiring the presence of both spliced expressed sequence tags (ESTs) with >96% identity to the sequence and either similarity to a known protein or association with an *ab initio* gene prediction. Using these criteria there are only 13 genes and 17 ESTs/EST clusters that meet these criteria. This limited expression contrasts sharply with *KIAA0187* and *ZNF11/33B* which are expressed in all tissues analysed. Examples of RT–PCR experiments are presented in Figure 2. In addition to these gene features, there are a further 13 pseudogenes and gene fragments within the sequence that are not associated with ESTs and have not been analysed by RT–PCR. These are presented in Table 2 and include HOX, ZNF and immunoglobulin lambda-related sequences.
The proximal 200 kb is rich in satellite repeats

The position and periodicities of tandem arrays >300 bp in total length are presented in Table 3 and Figure 1D. There are four 5–10 kb repeat arrays spaced ∼55–60 kb apart within the proximal 200 kb of sequence: two satellite II arrays, a GC-rich array with a 6 bp periodicity (the CAGGG repeat (24)) and a satellite III array. Interspersed between these repeats are nine shorter arrays including two TTCC tetranucleotides, two CT-rich repeats with 13 and 18 bp periodicities based on the pentamer CATT and four complex GC-rich satellites with periodicities of 22–23 bp which are related to the GCTTTTGGC repeat implicated in the transposition of ALD paralogues (21). There are also seven alpha satellite monomers abutting directly onto the satellite III array which exhibit ∼10–15% divergence from chromosome 10-specific alphoid sequences (39), and one short array of the HSREPS522 repeat (40). However, the most striking aspect of these repeats is their distribution: no array longer than 500 bp is present in the distal 825 kb of sequence.

Tracts of intrachromosomal duplicated sequence lie distal to the satellite repeats

The pericentromeric region of chromosome 10 contains two large genomic duplications with one copy of each duplication on each arm (25,32). Comparisons with unfinished sequence from 10p11 (http://webace.sanger.ac.uk:80/cgi-bin/display) indicates that the proximal duplication is ∼250 kb long and extends from ∼278 to 528 kb, whereas the more distal duplication is ∼90 kb long and extends from ∼762 to ∼853 kb (Fig. 1D, duplications 1 and 4). These figures are consistent with previous PEGE estimates (32). However, there is also an ∼11.3 kb non-processed gene fragment related to RSU1 and an ∼20 kb non-processed pseudogene fragment related to IFB12 (Fig. 1D, duplications 2 and 3). As RSU1 and IFB12 have been mapped to 10p13 and 10p12, respectively, it is likely that the identity between these cDNAs and the 10q11 genomic sequence is a result of one or two further intrachromosomal duplication events. This is supported by the fact that a FISH probe from this region hybridizes both to 10q11 and to 10p13 on human metaphases [y738F9-R (32)]. Thus, a minimum of three independent events appear to have resulted in the intrachromosomal duplication of >370 kb of sequence within a 650 kb region immediately distal of the satellite repeats (Fig. 1D).

Satellite arrays are separated by pericentromerically duplicated sequences

To identify all known paralogous sequences within the human genome we queried the non-redundant EMBL (EMBL-NR) database with a masked version of the 10q11 sequence (see Materials and Methods). We discarded all hits <2 kb long, together with three hits which consisted almost entirely of L1 elements and may have been formed by transposition rather than pericentromeric duplication (41). A total of 57 hits within 36 clones or contigs were identified which varied in length from ∼2.3 to ∼56 kb. The distribution of these paralogues relative to the 10q11 sequence is shown graphically in Figure 3A

Table 2. Pseudogenes and gene fragments with no associated ESTs

<table>
<thead>
<tr>
<th>Position in sequence</th>
<th>Protein similarity/e value</th>
<th>GenBank accession no.</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>84 254–84 547</td>
<td>Ig kappa LC variable region/9e-38</td>
<td>CAA45916</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>91 440–91 339</td>
<td>Naaldase 11 protein/5e-15</td>
<td>Q04609</td>
<td>Gene fragment</td>
</tr>
<tr>
<td>108 754–109 098</td>
<td>Ret finger protein/6e-14</td>
<td>Q9Y577</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>109 277–109 432</td>
<td>NADH-ubiquinone oxidoreductase chain 1/5e-13</td>
<td>Q35723</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>180 580–178 493</td>
<td>Poly(A) binding protein/2e-10</td>
<td>P11940</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>220 757–217 572</td>
<td>ZNF91-like/0</td>
<td>P28160</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>315 613–293 985</td>
<td>63.5 kDa CaEl hypothesical</td>
<td>P34624</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>437 727–437 933</td>
<td>Chloride cond. Inducer protein precursor/3e-10</td>
<td>Q61835</td>
<td>Gene fragment</td>
</tr>
<tr>
<td>448 964–449 767</td>
<td>Unknown protein hsp co/5e13</td>
<td>Q9Y262</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>596 883–576 677</td>
<td>IFB12/0</td>
<td>O60494</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>608 207–619 465</td>
<td>RSU-1/7e-41</td>
<td>Q15404</td>
<td>Gene fragment</td>
</tr>
<tr>
<td>630 215–629 973</td>
<td>Retinal homeobox protein 2/2e-9</td>
<td>O42357</td>
<td>Gene fragment</td>
</tr>
</tbody>
</table>

Only protein matches with an e value of >10-9 are shown together with one accession number for each feature. For definition of gene-related features see footnotes to Table 1.
The majority of clones containing sequences paralogous to 10q11 map to the pericentromeric regions of other human chromosomes including 2p11, 16p11.2, 19p12–q12, 21q11.1 and 22q11.1–11.2 (Table 4), suggesting that the sequence identity is a result of pericentromeric-directed duplication events. This is confirmed by comparative FISH analyses which demonstrate that these sequences map to different pericentromeric locations in different primate species (32, and http://bioserver.uniba.it/fish/Cytogenetics/10q11-examples/10q11.html). However, hits to the EMBL-NR database are not randomly distributed within the 10q11 sequence. A total of 32 of the 57 paralogous domains match sequences within the proximal 250 kb of sequence (Fig. 3A). Furthermore, these are longer (mean = 16.89 kb), and show higher sequence identity to 10q11 (mean = 94.3%), than paralogues of more distal sequences (means = 7.98 kb and 82.43%, respectively). This clustering of paralogues with high sequence identity to 10q11 within the proximal 250 kb suggests that the distribution of these sequences may be related to the satellite repeats in this part of the sequence.

Paralogous sequences often terminate at or within tandem repeats

To investigate the association of paralogous sequence with satellite repeats in more detail we looked for sequence features at the termini of all 57 paralogous domains identified within the EMBL-NR database (see Materials and Methods). A total of 20 domains terminate within interspersed repeats at one or both ends (data not shown). However, as ~48% of the 10q11 sequence is accounted for by interspersed repeats (Fig. 1B), the significance of this is unclear. Only four domains have one or both ends precisely defined by a repetitive element, implying a possible mechanistic involvement of the repeat in their formation. What is more striking is that 24 paralogous domains terminate at or within tandem repeats.

Table 3. Tandem repeats >300 bp in 10q11

<table>
<thead>
<tr>
<th>Position in 10q11 sequence</th>
<th>Shortest high scoring periodicity</th>
<th>No. of repeats</th>
<th>% matches</th>
<th>% indels</th>
<th>Consensus length of other periodicities (bp)</th>
<th>Repeat sequence/repeat name</th>
</tr>
</thead>
<tbody>
<tr>
<td>13–5307</td>
<td>5</td>
<td>1018.2</td>
<td>53</td>
<td>11</td>
<td>26/49</td>
<td>CGATT (81)/satellite 2</td>
</tr>
<tr>
<td>9120–9580</td>
<td>4</td>
<td>117.5</td>
<td>62</td>
<td>11</td>
<td>10/15</td>
<td>TTCC</td>
</tr>
<tr>
<td>9787–12 156</td>
<td>22</td>
<td>71.5</td>
<td>53</td>
<td>15</td>
<td>16/38</td>
<td>CGGGGCGAAAAAGCCACAGCGG</td>
</tr>
<tr>
<td>20 295–20 910</td>
<td>4</td>
<td>154.8</td>
<td>65</td>
<td>10</td>
<td>–</td>
<td>TTCC</td>
</tr>
<tr>
<td>44 906–45 315</td>
<td>5</td>
<td>90.6</td>
<td>67</td>
<td>25</td>
<td>13/18</td>
<td>CATTT</td>
</tr>
<tr>
<td>48 463–49 648</td>
<td>22</td>
<td>43.5</td>
<td>54</td>
<td>13</td>
<td>32/41</td>
<td>CGGGGCGAAAAAGCCCGCCGGG</td>
</tr>
<tr>
<td>64 590–70 924</td>
<td>5</td>
<td>1041.6</td>
<td>56</td>
<td>13</td>
<td>25/49</td>
<td>CGAT (81)/satellite 2</td>
</tr>
<tr>
<td>75 813–76 762</td>
<td>23</td>
<td>42.4</td>
<td>66</td>
<td>14</td>
<td>40</td>
<td>GGGGGCGAAAAGCCGCGCGCGG</td>
</tr>
<tr>
<td>124 720–127 605</td>
<td>5</td>
<td>617.8</td>
<td>66</td>
<td>18</td>
<td>13/18</td>
<td>CATTT</td>
</tr>
<tr>
<td>128 609–134 578</td>
<td>6</td>
<td>617.2</td>
<td>55</td>
<td>13</td>
<td>11/17</td>
<td>CAGGCG [CAGGG repeat (24)]</td>
</tr>
<tr>
<td>141 686–142 468</td>
<td>23</td>
<td>33.2</td>
<td>60</td>
<td>16</td>
<td>41</td>
<td>GGGGGCGAAAAGCCGCGCGCGG</td>
</tr>
<tr>
<td>193 874–203 965</td>
<td>5</td>
<td>1000</td>
<td>69</td>
<td>0</td>
<td>10/15 etc.</td>
<td>CCATT/satellite 3</td>
</tr>
<tr>
<td>203 966–205 198</td>
<td>171</td>
<td>7.2</td>
<td>79</td>
<td>1</td>
<td>342</td>
<td>Alpha satellite</td>
</tr>
<tr>
<td>573 491–573 807</td>
<td>45</td>
<td>7.2</td>
<td>61</td>
<td>8</td>
<td>–</td>
<td>HSREP522^3</td>
</tr>
<tr>
<td>580 108–580 410</td>
<td>3</td>
<td>102.3</td>
<td>63</td>
<td>5</td>
<td>15</td>
<td>TTC</td>
</tr>
<tr>
<td>614 522–614 880</td>
<td>2</td>
<td>190.5</td>
<td>66</td>
<td>15</td>
<td>9/11</td>
<td>CA</td>
</tr>
<tr>
<td>849 259–849 910^8</td>
<td>2</td>
<td>59/180^9</td>
<td>76/69</td>
<td>13/4^9</td>
<td>–</td>
<td>TA</td>
</tr>
<tr>
<td>860 745–861 173</td>
<td>2</td>
<td>215.5</td>
<td>56</td>
<td>6</td>
<td>–</td>
<td>CA</td>
</tr>
<tr>
<td>969 977–970 659</td>
<td>3</td>
<td>227</td>
<td>58</td>
<td>1</td>
<td>–</td>
<td>GTG</td>
</tr>
</tbody>
</table>

*a*The consensus shown is the reverse/complement of the 10q11 sequence.

^b^This repeat is interrupted by a partial L1PB2 element at position 849 381–849 556; the two figures given refer to the arrays proximal and distal to this element, respectively.
terminate in a tandem repeat at one or both ends. Of these, 21 share sequence identity with the proximal 250 kb and are shown in detail in Figure 3B. In 10 cases, one terminus is defined by a tandem repeat on chromosome 10 only (Fig. 3B, vertical bars). In a further eight cases, one terminus falls within a tandem repeat which is present on both chromosomes, whereas in three cases both termini fall within two different tandem repeat families (alignments 1C, 5B and 9). No one repeat is exclusively found at the termini, with paralogues terminating within the 22 bp GC-rich repeat, the 4 bp TTCC repeat, and the 5 bp classical satellites II and III. Furthermore, in alignment 6, one terminus falls within a GC-rich repeat on chromosome 10 and a CT-rich repeat on chromosome 21 (Fig. 3B). In several cases a tandem repeat is present within a domain of paralogy (1E, 2, 4D, 3B and 11).

**Identification of a boundary between multi-copy pericentromeric sequences and chromosome-specific DNA**

Because of the limited pericentromeric sequence data within the EMBL-NR database we needed to establish whether the skewed distribution of paralogous sequences (Fig. 3) is due to ascertainment bias. To do this we developed a panel of 74 sequence tagged sites (STSs) and established the distribution of 10q11-related sequences within the human genome using a monochromosomal somatic cell hybrid panel (see Materials and Methods). A number of paralogous STSs identified in this
The accession number, chromosomal assignment (where known), nucleotide range of alignment, alignment size and percentage identity to 10q11 is shown for each high scoring match >2 kb in length to the EMBL-NR database (see Materials and Methods). Where more than one alignment has been identified within a single clone/sequence they have been treated separately unless the results from Repeatmasker indicated that the indel of an interspersed repeat is responsible for the discontinuity between alignments.

<table>
<thead>
<tr>
<th>GenBank accession no. of clone with sequence identity to 10q11</th>
<th>Chromosomal assignment/map position of clone</th>
<th>Nucleotide range of sequence match within clone</th>
<th>Nucleotide range of sequence match within 10q11 contig</th>
<th>Alignment size (bp)</th>
<th>Identity (%)</th>
<th>No. of termini defined by tandem repeats</th>
<th>No. of termini defined by interspersed repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ac002041</td>
<td>16p11.2</td>
<td>(A)</td>
<td>89 731–99 723</td>
<td>12 525</td>
<td>96.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2. d87003</td>
<td>22q11.2</td>
<td>(A)</td>
<td>10 063–11 772</td>
<td>12 780</td>
<td>93.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3. ac002038</td>
<td>2p11</td>
<td>(A)</td>
<td>25 035–24 951</td>
<td>13 297</td>
<td>97.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4. ac006352</td>
<td>7</td>
<td>(A)</td>
<td>50 182–57 794</td>
<td>77 433</td>
<td>92.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5. ac006359</td>
<td>7</td>
<td>(A)</td>
<td>63 949–57 468</td>
<td>64 928</td>
<td>96.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6. AF239321</td>
<td>21q11.1</td>
<td>(A)</td>
<td>37 583–28 506</td>
<td>51 012</td>
<td>97.4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7. U91320</td>
<td>16p13.11</td>
<td>(A)</td>
<td>29 372–22 88</td>
<td>29 298</td>
<td>92.4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>8. AJ06996</td>
<td>21q22.1</td>
<td>(A)</td>
<td>34 606–32 357</td>
<td>22 660</td>
<td>96.3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9. ap005453–544</td>
<td>22q11.2</td>
<td>(A)</td>
<td>14 654–19 595</td>
<td>14 974</td>
<td>96.1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>10. ac006458</td>
<td>22q11</td>
<td>(A)</td>
<td>111 009–156 084</td>
<td>42 178</td>
<td>97.4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11. ap005456–547</td>
<td>22q11</td>
<td>(A)</td>
<td>92 091–67 475</td>
<td>26 813</td>
<td>96.8</td>
<td>0</td>
<td>2-L1PA4</td>
</tr>
<tr>
<td>12. ac002572</td>
<td>21q11</td>
<td>(A)</td>
<td>81 000–78 348</td>
<td>26 962</td>
<td>89.1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>13. ap000544–546</td>
<td>22q11</td>
<td>(A)</td>
<td>81 791–26 024</td>
<td>51 010</td>
<td>97.4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14. ac000541–542</td>
<td>22q11</td>
<td>(B)</td>
<td>30 200–20 94                                  35 963</td>
<td>95.24</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>15. ac004026</td>
<td>19p12</td>
<td>(B)</td>
<td>20 466–35 702</td>
<td>13 930</td>
<td>75.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16. x94277</td>
<td>Xp11.3–11.4</td>
<td>(B)</td>
<td>106 232–97 517</td>
<td>10 083</td>
<td>86.6</td>
<td>0</td>
<td>4-LTR19</td>
</tr>
<tr>
<td>17. AL096791</td>
<td>Xp11.21–11.4</td>
<td>(B)</td>
<td>88 142–95 620</td>
<td>84 949</td>
<td>85.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18. ac005357</td>
<td>19q12</td>
<td>(A)</td>
<td>38 716–19 158</td>
<td>23 063</td>
<td>82.7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>19. AL008723</td>
<td>22q12–13</td>
<td>(A)</td>
<td>115 555–335 341</td>
<td>56 663</td>
<td>74.53</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20. AL022324</td>
<td>22q11                                      12.2–12.2</td>
<td>(A)</td>
<td>33 581–339 244</td>
<td>56 969</td>
<td>71.99</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21. ac002059</td>
<td>12q12</td>
<td>(A)</td>
<td>120 635–116 682</td>
<td>50 606</td>
<td>76.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22. ac000045</td>
<td>Unknown</td>
<td>(A)</td>
<td>29 495–25 437</td>
<td>53 339</td>
<td>75</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23. ac005758</td>
<td>19q12</td>
<td>(A)</td>
<td>28 173–37 561</td>
<td>10 723</td>
<td>81.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24. d87000</td>
<td>22q12                                      12.2–12.2</td>
<td>(A)</td>
<td>31 333–33 998</td>
<td>32 925</td>
<td>72.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25. AL078475</td>
<td>21q11.1</td>
<td>(A)</td>
<td>36 941–56 164</td>
<td>22 638</td>
<td>85.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26. ac004883</td>
<td>7q11.21–23</td>
<td>(A)</td>
<td>36 703–34 297</td>
<td>31 466</td>
<td>87.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27. ap000025</td>
<td>21q11.1</td>
<td>(A)</td>
<td>60 969–67 978</td>
<td>7213</td>
<td>88.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28. ac000382</td>
<td>11</td>
<td>(A)</td>
<td>97 633–108 234</td>
<td>9335</td>
<td>86.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29. d87005</td>
<td>22q12</td>
<td>(A)</td>
<td>42 921–124</td>
<td>3215</td>
<td>81.12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30. D88269</td>
<td>22q12.2</td>
<td>(A)</td>
<td>18 906–15 921</td>
<td>31 222</td>
<td>81.39</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31. ac003064</td>
<td>22q12                                      13–13</td>
<td>(B)</td>
<td>32 761–29 209</td>
<td>37 284</td>
<td>80.17</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>32. AL021937</td>
<td>22q12–13</td>
<td>(A)</td>
<td>38 839–42 289</td>
<td>36 065</td>
<td>81.66</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>33. Z83839</td>
<td>22q12–13</td>
<td>(B)</td>
<td>42 921–124</td>
<td>33 181</td>
<td>81.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>34. ac005374</td>
<td>22q11                                      1.2–12.2</td>
<td>(A)</td>
<td>36 292–823</td>
<td>47 879</td>
<td>78.65</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>35. ap0050525–526</td>
<td>22q11                                      1.2–12.2</td>
<td>(B)</td>
<td>40 940–44 461</td>
<td>37 505</td>
<td>92.59</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>36. ac005224</td>
<td>17</td>
<td>(A)</td>
<td>42 592–44 973</td>
<td>2933</td>
<td>74.22</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>37. d87003/ d87018</td>
<td>22q11</td>
<td>(A)</td>
<td>37 352–437 472</td>
<td>16 610</td>
<td>96.02</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The accession number, chromosomal assignment (where known), nucleotide range of alignment, alignment size and percentage identity to 10q11 is shown for each high scoring match >2 kb in length to the EMBL-NR database (see Materials and Methods). Where more than one alignment has been identified within a single clone/sequence they have been treated separately unless the results from Repeatmasker indicated that the indel of an interspersed repeat is responsible for the discontinuity between alignments.
way were sequenced to assess levels of divergence from the 10q11 sequence. The results of these analyses are presented in Figure 4B.

All but one of the 28 STSs from the proximal 220 kb amplify PCR products of the expected size from between 3 and 14 monochromosomal hybrids (mean = 6.78). Alignment of DNA sequence generated from the PCR products amplified from these hybrids indicates that all of the paralogous sequences across the entire 220 kb are closely related to the 10q11 sequence, with paralogous STSs diverging from the 10q11 sequence by between 1.2 and 5.7%. This result is consistent with the high identities observed between paralogues of this region identified within the EMBL-NR database (Fig. 3).

In sharp contrast to the results from the proximal 220 kb, 35 of the 46 STSs distal to S214 are specific to chromosome 10 with the remaining 11 STSs amplifying a product from between three and six monochromosomal hybrids. Sequence analysis of PCR products from these STSs suggest that paralogues in this distal region may be more distantly related to the 10q11 sequence, with eight of nine paralogous STSs diverging from the 10q11 sequence by 5.0–9.6%. The small sample size makes it unclear whether this increased divergence relative to the proximal 220 kb is significant. However, it is noteworthy that three of these STSs (S528, WME31-R and S557) map within the distal cluster of paralogues identified within the EMBL-NR database which are also more diverged relative to the 10q11 sequence (Fig. 3A). A further five of the STSs from the distal region which amplify products from other chromosomes (S679, S680, S702, S705 and S713) lie within the KIAA0187 gene (666 675–713 708 bp). Paralogous fragments of this gene have already been identified within the immunoglobulin lambda and cat-eye syndrome regions on chromosome 22 (19) (Table 4, alignments 35B and 37). Thus, the results of the STS analysis are totally consistent with the BLAST results (Fig. 3) and indicate that an abrupt discontinuity exists between satellite-rich interchromosomally duplicated sequences and chromosome-specific sequences.

The genomic distribution of paralogous STSs is associated with the distribution of specific satellite families

To investigate the link between pericentromeric duplications and satellite sequences on other chromosomes we have analysed the chromosomal distribution of paralogues of the 10q11 STSs in relation to the chromosomal distribution of linked satellites. The basic distribution of the satellite 2, satellite 3 and the CAGGG repeat families, as assayed by FISH (24,42), is shown in Figure 4A. The distributions of these repeat families overlap extensively, with chromosomes 1, 2, 9, 10, 15, 16, 17 and 22 being positive for two of the three families and nine chromosomes being negative for all three families. However, what is more striking is the association between the chromosomal distribution of paralogues of STSs within the proximal 250 kb and the distribution of these satellites. Chromosomes which lack these satellites (chromosomes 3, 4, 6, 8, 11, 12, 18, 19 and X) are positive for only zero to five 10q11 STSs (mean = 1.9) compared with an average of 14 for chromosomes which contain two or more of the satellite families analysed (chromosomes 1, 2, 9, 14–17 and 22). There are some exceptions to this association; for example, chromosome 5 contains satellite 3 sequences but is negative for all but one
of the 10q11 STSs. Despite this, the association between the chromosomal distribution of each of these repeats with the distribution of paralogous STSs is highly significant using Fisher’s exact test \( P < 2 \times 10^{-4} \) for each satellite when chromosome 10 is excluded from the analysis; see Materials and Methods). No association could be detected with alpha satellite suprafamily organization (43).

**DISCUSSION**

We have generated ~1025 kb of contiguous genomic sequence obtained from three YAC clones which link pericentromeric satellite repeats in 10q11.1 to the RET proto-oncogene in 10q11.2. We have confirmed the fidelity of the data using PFGE (see Materials and Methods), by analysing >230 kb of overlapping sequence, and by PCR analyses of the termini of domains of paralogy in 10q11 (Fig. 3B). The global features of the sequence conform to our expectations for the transition from a centromeric G-dark band (10q11.1) to a G-light band (10q11.2). First, the GC content rises steadily from ~39.5% at the centromeric satellites to >55% at RET. A similar rise in GC content occurs within the proximal 2 Mb of the chromosome 22q sequence (19). Second, our data indicate an excess of both LINEs and retroviral LTRs within the sequence (Fig. 1B) which supports the conclusion drawn from low resolution FISH-based analyses (38) and limited sequence data (44) that human pericentric regions are rich in interspersed repeats. However, this excess is modest compared with the over-representation of transposable elements close to heterochromatin in Drosophila (9) and to the centromeres of chromosomes 2 and 4 in Arabidopsis (4,5), suggesting that regional variations in transposable element distribution may not be as marked within the human genome as they are in other eucaryotes.

**Identification of two pericentromeric sequence domains**

The most striking aspect of the sequence is the non-random distribution of satellite repeats, pericentromeric-directed duplications and intrachromosomal duplications. Satellites and other repeats are located almost exclusively within the proximal 200 kb and are separated by complex sequences which have numerous paralogues within the genome. The predominantly pericentric location of these paralogous sequences (Table 4), and comparative FISH analyses (32, and http://bioserver.uniba.it/fish/Cytogenetics/10q11-examples/10q11.html ), indicate that they have expanded in copy number due to recent pericentromeric-directed duplication events. The fact that most of these paralogues terminate at or within tandem repeats (Fig. 3B), and that few are found outside the satellite-rich region (Fig. 3A), suggests that specific repeats are directly responsible for the rapid and centromere-specific evolution of these sequences, a hypothesis supported by the correlation between the genomic distribution of paralogues of 10q11 sequences and the genomic distribution of satellites 2 and 3 and the CAGGG repeat (Fig. 4).

This sequence organization is very similar in structure and evolutionary dynamics to the sequences derived from Xq28 and 4q24 which abut onto the alpha satellite array on chromosome 16p11 (20). This, together with the identification of pericentromerically duplicated ALD- and NF1-related sequences close to satellite arrays on chromosome 22 (19) and the large number of pericentromerically duplicated sequences that have been analysed at the cytogenetic level (18), strongly suggest that tracts of rapidly evolving complex sequences may be interspersed between satellite arrays on most, if not all, human chromosomes. At the sequence level, the identification of repeats defining the termini of paralogous sequence domains (Fig. 3B) is also consistent with the presence of repeats at the termini of ALD paralogues (21) and the identification of the GC-rich CAGGG and HSREP522 repeats at the termini of pericentromeric duplications on chromosomes 16, 21 and 22 (24). The similarity of the CAGGG repeat to immunoglobulin switch regions (45) has recently led to the suggestion that a specific subset of human repeats may be responsible for intrachromosomal duplication (24). However, the wide spectrum of repeats found at the termini of paralogous domains in 10q11, including CT-rich sequences and AT-rich classical satellites (Fig. 3B), indicates that the ability to promote interchromosomal duplication may be a function of repetition per se.

In sharp contrast to sequences within the proximal 240 kb, most of the distal 780 kb of sequence is chromosome specific. However, the existence of a second, distal, pericentromeric sequence domain is implied by the clustering of >370 kb of intrachromosomally duplicated DNA within an ~600 kb region distal to the satellites. The inverted orientation of the ZNF and D10S141 duplications (32) and the fact that the RSU1 and IFB12 genes map to 10p12–10p13 suggest that these sequences have been subjected to a minimum of three independent intrachromosomal duplication events. Furthermore, the fact that these duplications do not contain repeats and are relatively stable components of primate genomes (32) indicates that they are behaviourally distinct from the more proximal interchromosomal duplications. It is possible that these have been produced by rearrangements initiated within the proximal satellites but resolved within more distal, single copy sequences. If so, the second cluster of diverged paralogues which lie between the ZNF and IFB12 duplications (Fig. 3A, 530–580 kb) could represent pericentromerically duplicated sequences that have become isolated from satellites during the formation of the ZNFB or IFB12 duplicated sequences. Data from similar duplications on chromosomes 1, 2, 9 and 21 (26–29) will establish whether distinct pericentromeric domains of intrachromosomally duplicated sequences are common features of human chromosomes.

**Identification of a boundary between heterochromatin and euchromatin**

Because the sequence presented here includes both pericentric satellite arrays and highly expressed genes, it is presumed to contain both heterochromatic and euchromatic sequences (31,32). Although the distinction between heterochromatin and euchromatin is cytological, data from other eucaryotes indicate that it can reflect basic differences in underlying sequence organization (1–5). The well established association of satellite repeats with heterochromatin (1–8) and the fact that repeats per se can induce gene silencing (14,15) strongly suggest that the proximal satellite sequences in 10q11, and the intervening duplicated sequences, would be heterochromatic in nature. We would therefore predict that they are gene poor. This appears to be the case as we can identify only a single EST cluster proximal to the satellite 3 array in addition to four pseudogenes and a single gene fragment (Tables 1 and 2).
confirmed the presence of three genes all of which have been
the transcriptional profile of large genomic regions (47,48).
prone to errors, making it is difficult to establish definitively
termination of genes within complex eucaryotic genomes is
of pericentromerically duplicated DNA which is gene poor and
sequences proximal to pericentromeric satellite arrays consist
sequence means that hybridization and PCR-based techniques
laborious task as the multicopy nature of pericentromeric
therefore, require experimental confirmation. This may be a
 bona fide 10q11 (ap000526.1). Whether these are
contains a paralogous fragment of the
short predicted ORFs (115–130 amino acids) within unspliced
EST clusters ap000528.1, ap000529.2 and ap000534.2), a further three have
specific during map construction, suggesting that this sequence
comparable in structure and content to the proximal 250 kb
of the 10q11 sequence. A total of 37 gene features have been
annotated within this 1 Mb of sequence, 30 of which are prox-
imal to a satellite II array (19, http://www.sanger.ac.uk/cgi-
bin/cwa/22cwa.pl ). As in 10q11, most of these are based on
identity to known proteins with no associated ESTs showing
>96% identity to the sequence, although in contrast to the
10q11 data there are also 11 gene features based on EST
matches. However, seven of these ESTs/EST clusters are
contiguous with genomic DNA and have no associated gene
structure (ap000523.3, ap000525.2, ap000525.7, ap000527.2,
ap000528.1, ap000529.2 and ap000534.2), a further three have
short predicted ORFs (115–130 amino acids) within unspliced
ESTs (ap000525.5, ap000525.6 and ap000527.1) whereas the
last is identified by a single spliced EST (AA725634) which
contains a paralogous fragment of the KIAA0187 gene from
10q11 (ap000526.1). Whether these are bona fide genes will,
therefore, require experimental confirmation. This may be a
laborious task as the multicopy nature of pericentromeric
sequence means that hybridization and PCR-based techniques
alone cannot provide definitive proof of transcription, a
problem already noted during transcriptional analysis of the
subtelomeric region of 4q (46). The chromosome 22 data are,
therefore, currently consistent with the interpretation that
sequences proximal to pericentromeric satellite arrays consist
of pericentromerically duplicated DNA which is gene poor and
heterochromatic.
The transcriptional analysis also indicates that sequences
which lie distal to the satellite arrays are expressed. The iden-
tification of genes within complex euchromatic genomes is
prone to errors, making it is difficult to establish definitively
the transcriptional profile of large genomic regions (47,48).
Using conservative criteria for gene identification we have
confirmed the presence of three genes all of which have been
identified previously: ZNF11/33B (49), KIAA0187 (50) and
RET (51). All other gene-related features consist of transcripts
with no clear gene structure, non-spliced ESTs, or gene frag-
ments identified through protein database searches with no
associated ESTs. Whereas we cannot formally rule out the
possibility that some, or all, of these features correspond to
bona fide genes, it is also possible that they represent a mixture
of cDNA library artefacts, short aberrant or repeat-associated
transcripts, and recently formed pseudogene fragments. Thus,
although sequences distal of the satellites are expressed, the
region appears to be gene poor.
The lack of transcripts within the satellite-rich proximal
200 kb, and the position of the most proximal confirmed tran-
script (~289 kb) and gene (ZNF11/33B, ~475 kb), suggests that
there may be a single boundary between heterochromatin and
euchromatin within the 10q11 sequence presented here. The
most likely location for this is between the satellite III array at
~200 kb and the transcript at ~289 kb. Interestingly, this
approximately defines the transition from sequences prone to
interchromosomal duplication and sequences which have been
involved in intrachromosomal duplication. However, either
FISH-based techniques (52) or chromatin immunoprecipita-
tion (53) will be required to confirm that this sequence discon-
tinuity corresponds to a boundary between chromatin types.
This will be particularly important in light of the discovery that
heterochromatin can extend several hundred kilobases beyond
the repeats that are presumed to be a pre-requisite for its forma-
tion (11).

Does pericentromeric duplication have a function?
The gene-related nature of many pericentromeric duplications
(17,20–23), together with the established role of telomere plas-
ticity in generating antigen variation in some primitive eucary-
otes (54), has led to the proposal that rapidly evolving
pericentromeric sequences may be an important source of
novel genes (17,21). The data presented here suggest that these
sequences may be heterochromatic, making it is unlikely that
pericentromeric rearrangement will lead directly to the forma-
tion of new genes. However, if gene formation is not the
primary function of pericentromerically duplicated sequences,
then what is?
The most straightforward answer to this question is that these
sequences are of no adaptive significance whatsoever and exist
due to the combined action of two well characterized
phenomena. First, centromeric satellites are known to evolve
rapidly by concerted evolution with frequent sequence
exchange between repeat arrays on non-homologous chromo-
somes (55). Second, duplications of euchromatic material into
centric heterochromatin have been reported as cytogenetically
visible variants within the human population (56, and refer-
ences therein) and may represent the first step in pericentro-
meric-directed duplication events (20). When such a chance
event occurs, the euchromatin-derived sequence may be prone
to interchromosomal transfer through exchange events which
initiate within flanking satellite sequences. Pericentromeric
duplication may, therefore, simply represent the transient inter-
mediates of satellite homogenization. More functionally based
proposals include a role in chromosome pairing [the only
established function of pericentromeric satellites outside
centromere activity is in maintaining alignment between
homologous chromosomes during achiasmate meioses in
Drosophila (57)], and a role as spacers between satellites and
genes to guard against centromeric position effects (20).

Although plausible, none of these hypotheses can account
for the lack of equivalent sequences within the heterochro-
matin of Arabidopsis (4,5) or Drosophila (9) indicating that
more elaborate possibilities warrant consideration. The identi-
fication of a growing number of proteins that bind to satellite
sequences (58,59) and the fact that the relocation of sequences
towards centromeric nuclear domains can be involved in gene
silencing (60,61) suggest that satellite sequences are involved
in complex patterns of nuclear compartmentalization (62,63)
and gene expression (64,65). It is possible that ‘pseudo-
euchromatic’ sequences between satellites help to define the
position of functionally distinct silencing domains within the
nucleus, and could even provide a means to dynamically alter
their position through changes in chromatin conformation.
Alternatively, the facts that heterochromatic sequences in
Arabidopsis and Drosophila are rich in transposons (9,10) relative to the equivalent regions in humans (Fig. 1B), and that high frequencies of element movement can significantly reduce fitness (67), raise the possibility that pericentromeric duplication represents a novel mechanism which has evolved in repeat-rich genomes to modulate transposition frequency (by providing pseudo-euchromatic pericentromeric sites where insertions will not be deleterious) and/or element density (through a continual process of deletion and renewal of recombinationally deficient pericentromeric sequences). The centromere-directed nature of the burst of retroviral transposition reported within an interspecific marsupial hybrid (68) could be accounted for by such a mechanism.

The data presented here strongly suggest that the pericentromeric sequences in 10q11 are organized into two domains based both on the presence or absence of tandem repeat arrays and on the type of rearrangement to which they are prone. Furthermore, transcriptional activity suggests that the boundary between these domains approximately defines a transition between pericentromeric heterochromatin and euchromatin. This raises the possibility that most, if not all, human chromosomes will exhibit similar organizational features. Realistic short-term aims will be to extend the sequence data presented here to the centromeric alploid array (D10Z1), to develop equivalent data from other centromeres both in man and other mammals for comparative analyses, and to experimentally confirm the existence of the abrupt changes in chromatin state implied here.

**MATERIALS AND METHODS**

**YAC subcloning, sequencing and analysis**

The restriction map of each YAC DNA preparation used for cloning was analysed by PFGE to ensure that no rearrangements occurred during culture. YAC DNA was isolated, shotgun cloned and sequenced as described previously (33,34). The restriction map of each finished sequence was compared with the existing PFGE map and found to be in agreement with the exception that four BssHII sites and two MluI sites within the PFGE map correspond to two or more closely spaced sites in CpG islands within the finished sequence (data not shown). This is presumed to be due to the low resolution of the PFGE map. The finished sequence was subjected to the standard Sanger Centre automated analyses (http://www.sanger.ac.uk/HGP/Humana/human_analysis.shtml) and imported into an AceDB database (http://www.sanger.ac.uk/HGP/Humana/ACE.shtml) to allow interactive interpretation of results. In addition, the sequence was split up into overlapping 50 and 100 kb sections and analysed using NIX (69) which provides an interactive graphical interface to simultaneously query the results of multiple gene prediction and database search programs. Ab initio gene predictions were obtained using Grai2, Fgene, GeneFinder, Fgenes, GenScan and HMMGene (70–72). Interspersed and tandem repeats were identified using Repeatmasker (http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl) and Tandem Repeat Finder (73). Similarities to existing genomic, EST and protein sequences were identified by using repeat-masked 10q11 sequence to query Swissprot, TREMBL and EMBL databases with the BLAST family of programs (74,75). The overlap between y738F9 and y214H10 was aligned using Blast 2 sequences (76) whereas cDNA sequences were aligned to genomic DNA using est_genome available through the UK HGMP Resource Centre. The GC content and distribution of interspersed repeats was established by using RepeatMasker to analyse overlapping 20 kb sequence files with a 16 kb overlap which were generated using in-house software. The tabulated summary from each analysis was then transferred to a single text file using in-house software and analysed using Microsoft Excel.

**RT–PCR analysis**

A panel of eight cDNAs derived from adult tissues (Clontech, Palo Alto, CA) were analysed according to the manufacturer’s recommendations. Primers were not generated from two ESTs due to the presence of interspersed repeats (N89814 and HS1251874). All primers used in the analysis can be viewed at http://www.ncl.ac.uk/human.genetics/rtpcr.html or obtained from the corresponding author.

**Analysis of paralogous sequences**

Each genomic clone containing paralogous sequences was compared with the 10q11 sequence using DOTTER (77). Regions of high identity were then aligned using BESTFIT (78) to estimate the percentage identity between the sequences. Where complex patterns of alignment were identified within a single clone each alignment was considered independently except when the discontinuity in alignment was clearly due to the insertion/deletion of a transposable element. Sequences longer than 32 kb were split prior to alignment with an average percentage identity being calculated from the individual alignments after correction for alignment length. The percentage identity between two sequences is defined as the number of matching nucleotides divided by the total alignment length (excluding insertions and deletions). The percentage similarity is the number of matching nucleotides divided by the sum of the number of matching nucleotides and the number of insertions and deletions. Features at the termini of each alignment (defined by BESTFIT) were identified by analysing the output of RepeatMasker and Tandem Repeat Finder. All primers used to verify the integrity of domain termini can be viewed at http://www.ncl.ac.uk/human.genetics/paralogues.html or obtained from the corresponding author.

**Somatic cell hybrid analysis**

The Human Monochromosomal Hybrid Panel available from the UK HGMP Resource Center (79) was used to establish the chromosomal distribution of 10q11 STSs. The chromosome 8 hybrid in this panel also contains the pericentromeric region of chromosome 22. Paralogous STSs amplified from this hybrid were mapped further using two additional hybrids which contain chromosomes 8 (NA1015B6) and 22 (NA10888) in isolation (80). A number of STSs amplified from the chromosome 20 hybrid (GM10478) were found to share 100% identity to the 10q11 sequence. None of these STSs amplified a product due to the presence of interspersed repeats (N89814 and HS1251874). All primers used in the analysis can be viewed at http://www.sanger.ac.uk/HGP/Humana/human_analysis.shtml or obtained from the corresponding author.
STSs amplified from chromosomes which are positive for each satellite and the number amplified from chromosomes negative for each satellite and comparing these two totals with the expected totals if a random distribution is assumed (chromosome 10 was excluded from the analysis as all STSs are derived from this chromosome). Independent analysis of each satellite families gave a $P$ value of $<2 \times 10^{-5}$ in all three cases (Fisher’s exact test). All primers used in the analysis can be viewed at http://www.ncbi.nlm.nih.gov/human/genetics sts.html or obtained from the corresponding author.

**STS sequencing**

PCR products were amplified from somatic cell hybrids (50 ng) and purified using Qiagen PCR purification kits (Qiagen, Hilden, Germany) according to the manufacturer’s instructions except that DNAs were eluted in water. Approximately 100 ng of template was used for each sequencing reaction and all sequencing reactions were performed using an ABI PRISM BigDye cycle sequencing kit according to the manufacturer’s instructions (PE Applied Biosystems, Foster City, CA) and analysed using an ABI377 (PE Applied Biosystems). Each template was sequenced in both orientations. Despite this, a significant number of templates failed to give unambiguous sequence data, suggesting that multiple paralogues of some STSs are present on some human chromosomes. High quality sequence obtained in both orientations was aligned to the 10q11 sequence and divergences calculated using the Megalign software package (DNAsstr, Madison, WI). Where more than one paralogous STS was available, each sequence was aligned to the 10q11 sequence separately and the mean divergence was taken.

**ACKNOWLEDGEMENTS**

We thank T. Strachan for critical reading of the manuscript. The financial support of the AIRC, Telethon (grant E.672), Wellcome Trust (grant 049859) and EC (contract BMH4-CT92-2433) is gratefully acknowledged, as are a short-term Wellcome Trust (grant 049859) and EC (contract BMH4-CT92-2433) molecular structure of a functional Drosophila centromere. Cell, 91, 1007–1019.


