# Molecular structure and evolution of an alpha satellite/ non-alpha satellite junction at 16p11

**Juliann E. Horvath, Luigi Viggiano[1], Brendan J. Loftus[2], Mark D. Adams[2], Nicoletta Archidiacono[1], Mariano Rocchi[2] and Evan E. Eichler+**

Department of Genetics and Center for Human Genetics, Case Western Reserve School of Medicine and University Hospitals of Cleveland, Cleveland, OH 44106, USA, [1]Instituto di Genetica, Via Amendola 165/A, 70126 Bari, Italy and [2]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received 15 September 1999; Revised and Accepted 27 October 1999

DDBJ/EMBL/GenBank accession nos AC002307, AC002038, AF182004–AF182010, AF183323–AF183377, AF183380

**We have determined the detailed molecular structure and evolution of an alpha satellite junction from human chromosome 16p11. The analysis reveals that the alpha satellite sequence bordering the transition lacks higher-order structure and that the non-alpha satellite portion consists of a mosaic of duplicated segments of complex evolutionary origin. The 16p11 junction was formed recently (5–10 million years ago) by the duplication and transposition of genomic segments from Xq28 and 4q24. Once this mosaic structure was formed, a larger complex was spread among multiple pericentromeric regions. This resulted in the formation of large (>62 kb) paralogous segments that share a high degree (~97%) of sequence similarity. Both phylogenetic and comparative analyses indicate that these pericentromeric-directed duplications occurred around the time of the divergence of the human, gorilla and chimpanzee lineages, resulting in the subtle restructuring of the primate genome among these species. The available data suggest that such chimeric structures are a general property of several different human chromosomes near their alpha satellite junctions.**

## INTRODUCTION

Alpha satellite DNA constitutes an estimated 3–5% of all human genomic material (1). It is the only repetitive motif associated with centromere function (2) and has been strongly implicated in facilitating proper meiotic and mitotic chromosomal segregation. Alpha satellite is composed of a 171 bp repeating monomer unit which itself may be organized into several discrete higher-order repeat or array structures (3–5). The ubiquity of the repeat at primary points of constriction among all human and primate chromosomes has been taken as indirect evidence of its functional role. Although alpha satellite DNAs were among the first repeat motifs discovered, large-scale sequence analysis of such regions has proceeded slowly. With few exceptions (6), these regions remain among the most poorly mapped areas of the human genome. Due to the repetitive nature of alpha satellite repeats and difficulties associated with mapping, cloning and sequencing DNA within its vicinity, centromeric and pericentromeric DNA has been generally excluded as potential targets of the Human Genome Project (7). Relatively few physical landmarks exist which demarcate the transition of non-alpha satellite and alpha satellite DNA. The biological importance of such regions in chromosome segregation and their demarcation as points of functional transition between euchromatin and heterochromatin, however, necessitate additional scrutiny.

Recently another property of pericentromeric regions has emerged which indicates that regions near the periphery of human centromeres have been preferential sites for gene duplication. A series of recent gene duplications, often incomplete in structure, have been documented for numerous chromosomal regions (1p12, 1q12, 2p11, 9p11, 10p11, 13q11, 14q11, 15q11, 16p11, 17p11, 18p11, 18q11, 20p, 20q, 21q11 and 22q11) (8–18). Although gene duplications have been reported to map near centromeric regions, the molecular relationship between alpha satellite DNA and pericentromeric duplications has not been explored. Preliminary data from several independent studies indicate that evolution of such regions is amazingly complex. An evolutionary survey of the pericentromeric region of chromosome 10, for example, revealed that markers within the region were often absent, duplicated or rearranged (6). The data suggest that such regions are subject to a remarkable degree of evolutionary turnover, even among closely related primate species.

We present a detailed structural analysis of a 160 kb transition from alpha satellite to non-alpha satellite on human chromosome 16p11. Comparative sequencing and cytogenetic studies reveal that the non-alpha satellite portion is composed of duplicated genic/genomic segments that originated from 4q24 and Xq28. These mosaic structures were duplicated and distributed to other pericentromeric regions. This has resulted in large blocks of paralogy (>62 kb) with a high degree of sequence similarity (96.9%) among non-homologous chromosomes. Differences in the copy number and distribution of these duplicated segments were observed among the genomes of man and the great apes, indicating that such macromuta-

+To whom correspondence should be addressed at: Department of Genetics, Case Western Reserve University, BRB720, 10900 Euclid Avenue, Cleveland, OH 44106, USA. Tel: +1 216 368 4883; Fax: +1 216 368 3432; Email: eee@po.cwru.edu

tional events have been a source of chromosomal variation. The data suggest extensive interchromosomal exchange at the periphery of alpha satellite DNA and support a two-step model for pericentromeric-directed duplications within the hominoid genome.

## RESULTS

### Identification of an alpha satellite boundary

Previously (11), we had determined that a 9.7 kb genomic segment from Xq28, containing exons 7–10 of the adrenoleukodystrophy (*ALD*) gene, had been duplicated to the pericentromeric regions of human chromosomes 2p11, 10p11, 16p11 and 22q11. The available fluorescence *in situ* hybridization (FISH) and sequence-tagged site (STS)-mapping data from 16p11 had placed the duplicated segment in close proximity to the chromosome 16 centromere indicating that this segment was the most proximally located 'unique' sequence probe within 16p11 (11). To determine whether the duplicated *ALD* segment might demarcate the transition between alpha and non-alpha satellite sequence, a human bacterial artificial chromosome (BAC) library and a chromosome 16-specific cosmid library were probed independently with the duplicated *ALD* segment and a generic alpha satellite repeat probe (see Materials and Methods). Four clones were identified (BAC37914, BAC30582, c366a10 and c341b10) which cohybridized to both the *ALD* and alpha satellite repeat probes. End-sequence analysis of each clone revealed that only one end of the insert contained sequence typical of alpha satellite repeat whereas the opposite end occurred within non-alpha satellite DNA. These results suggested that each of these clones spanned an alpha/non-alpha satellite DNA junction. In order to confirm the chromosome 16 origin of the BACs, paralogous *ALD* sequence variants specific for the chromosome 16 duplicated segment were compared with the corresponding homologous sequences from the BACs. These were found to be identical to the chromosome 16 sequence signatures, therefore confirming their map assignment. Finally, the map location of the clones with respect to the centromere-associated higher-order chromosome 16 alpha satellite array was tested by extended chromatin analysis using chromosome 16 somatic cell hybrid nuclear spreads. *In situ* hybridization with the probes in conjunction with D16Z2 alpha satellite (19) repeat consistently showed that the clones originated from the periphery of the higher-order alpha satellite repeat domain, although some variation in copy number and organization has been detected among different individuals (data not shown).

### Sequence organization of the alpha satellite boundary

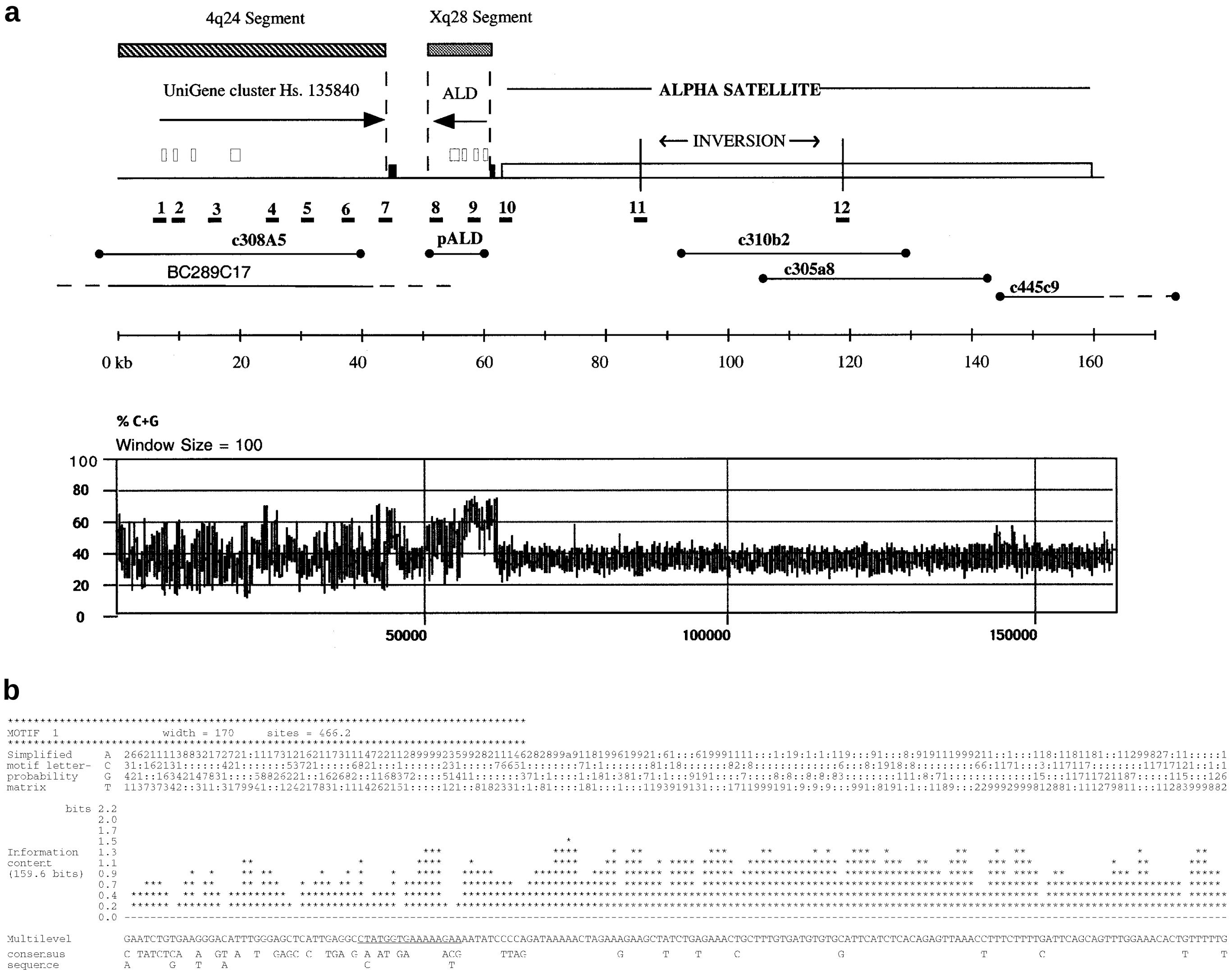
The insert (161 580 bp) of one of the chromosome 16p11 BACs (clone 37914) that spanned an alpha/non-alpha satellite boundary was sequenced in its entirety (GenBank accession no. AC002307). Repeatmasker analysis revealed that the sequenced clone consisted of ~92 kb of alpha satellite repeat (coordinates 64 536–161 580 of GenBank accession no. AC002307). The tract of alpha satellite DNA was relatively homogeneous, with only four retroposed elements (L1PA5, L1PA2, MER9 and L1PA2) disrupting its continuity. A large (33 kb) alpha satellite inversion was identified (Fig. 1a) and confirmed by PCR amplification and sequencing directly from somatic cell hybrid DNA (see Materials and Methods). The alpha satellite portion of the sequenced insert was subjected to multiple expectation–maximization for motif elicitation (MEME) analysis (20). A multilevel consensus sequence was generated based on the alignment of 466 individual monomer units (Fig. 1b). It showed 80.7% sequence identity to a previously characterized human alpha satellite consensus (X07685) (21). Inter-monomeric comparisons of alpha satellite revealed a non-uniform degree of conservation along the length repeat (with positions 55–170 showing the greatest degree of sequence conservation). The canonical sequence of the CENP B box, common to most higher-order alpha satellite arrays, was not found. Instead, the canonical binding site for the pJα binding protein (22) was located at this position. Dot-matrix and Miropeats analysis (23,24) of the sequence failed to show any evidence of higher-order repeat structure nor could recurrent higher-order restriction patterns be identified by *in silico* digestion of the alpha satellite segment (WebCutter Analysis). These data suggest that the 37914 alpha satellite may be classified as belonging to suprachromosomal family 4 (4). Phylogenetic analysis of 37914 alpha satellite consensus revealed no close evolutionary relationship to previously characterized monomer consensus sequences (J1, J2, D1, D2, W1-W5 and M1) (data not shown).

### Validation of sequence

Due to reports of the potentially unstable nature of genomic clones harboring alpha satellite DNA (25–27), the fidelity of the sequence organization was assessed in a variety of ways. First, PCR assays were developed at sites of alpha satellite inversion and transition (Fig. 1a, products 10–12). Products were amplified and directly sequenced from three human control DNAs and a chromosome 16 somatic cell hybrid (GM11000). No differences in PCR length or sequence were observed, indicating that the junctions were *bona fide*. Secondly, genomic BAC clones that originated from a different library source, and therefore a different individual, were identified (RPCI-11 445e5 and 347o23) which corresponded to the 37914 sequence. Comparative (*Eco*RI and *Not*I) restriction digests and Southern analyses of these BAC clones revealed no significant restriction fragment length polymorphisms, indicating that these allelic copies are virtually identical. As a final assessment of the validity of the sequence, an underlying scaffold map of chromosome 16 cosmids was developed based on a series of hybridizations with probes derived from 37914 (Fig. 1a). A subset of these cosmids was selected for T7 and SP6 end-sequence analysis. Both the sequence and its position with respect to the insert of each clone was in agreement with the sequence organization of 37914, verifying that the sequence represented the true genomic organization of this region.

### Complex interchromosomal duplications demarcate the transition

Once the sequence organization of the region had been verified, the non-alpha satellite portion (coordinates 1–64 521) of the insert was systematically examined for the presence of recently duplicated segments. The sequence was masked for common primate repeat elements (Repeatmasker v2.0) and BLASTN nucleotide sequence similarity searches were



```
********************************************************************************
MOTIF  1                width = 170     sites = 466.2
********************************************************************************
Simplified      A  2662111138832172721:111731216211731114722112899992359928211146282899a9118199619921:61:::619991111:::1:19:1:1:119:::91:::8:919111999211::1:::118:1181181::11299827:11:::::1
motif letter-   C  31:162131:::::421::::::53721::::6821:::1::::::231::::76651:::::::71:1:::::::81:18::::::82:8::::::::::::6:::8:1918:8:::::::::66:1171:::3:117117::::::::11717121::1:1
probability     G  421::16342147831:::58826221::162682::1168372:::51411::::::371:1::::1:181:381:71:1:::9191:::7:::8:8::8:8:83::::::::111:8:71:::::::::::::::15:::11711721187:::::115:::126
matrix          T  113737342::311:3179941::124217831:1114262151::::121::8182331:1:81:::181:::1:::1193919131:::1711999191:9:9:9:::991:8191:1::1189:::229992999812881:111279811:::11283999882

Multilevel         GAATCTGTGAAGGGACATTTGGGAGCTCATTGAGGCCTATGGTGAAAAAGAAAATATCCCCAGATAAAAACTAGAAAGAAGCTATCTGAGAAACTGCTTTGTGATGTGTGCATTCATCTCACAGAGTTAAACCTTTCTTTTGATTCAGCAGTTTGGAAACACTGTTTTTG
consensus          C TATCTCA  A GT A  T GAGC C  TGA G A AT GA    ACG     TTAG             G          G     T    T    C              G                    T       C                    T     T
sequence           A      G   T  A                    C           T
```

**Figure 1.** Organization of the 16p11 αλπηα satellite junction. (**a**) A schematic diagram depicting the organization of the 16p11 alpha satellite junction as determined by analysis of clone 37914 (AC002307). The 4q24 and Xq28 paralogy domains (shown as hatched and shaded boxes), the corresponding exonic regions (open boxes) and the direction of transcription based on the ancestral loci (horizontal arrows) are depicted above the line. Numbers (1–12) immediately below the line correspond to the positions of monochromosomal somatic cell hybrid PCR assays developed against the reference sequence (see below). The positions of cosmid clones (e.g. c308a5), long-range PCR products (pALD) and BAC clones (BC289C17) are indicated. BC289C17 corresponds to the ancestral 4q24 locus. All other probes are derived from chromosome 16-specific reagents. The coordinate start and end points of each clone were verified by end-sequence analysis of each probe. Shown below the figure is the % GC composition calculated over a 100 bp sliding window. (**b**) A multilevel consensus sequence of the alpha satellite portion (~92 kb) of clone 37914. The consensus was generated from 466 alpha satellite motifs using MEME software analysis which provides a relative measure of the degree of conservation for each base pair position in the consensus motif (reported as information bits). The observed frequency of each base pair in the multilevel consensus is summarized in the simplified motif letter-probability matrix. A letter 'a' within the motif letter-probability matrix denotes conservation approaching 100%. The underlined region in the multilevel consensus sequence corresponds to the pJα binding site.

performed. The size of the duplicated segments and the degree of sequence similarity were determined by a combination of dot-matrix, Miropeat and GAP alignment analyses between target and query sequences (Table 1). BLAST analysis identified two putative genic regions that showed evidence of intron/exon structure when compared with cDNA source material. These included: the expected duplication of four exons/introns of the Xq28 *ALD* gene and a four exon segment of a putative gene (UniGene cluster Hs 135840) which we assigned to chromosome 4q24 (Fig. 1a). In addition, a remarkably complex series of interchromosomal and intrachromosomal duplications were identified from diverse regions of the human genome. The largest duplication detected was >62 kb in length and showed ~97% sequence similarity between the 2p11 and 16p11 paralogous segments (Fig. 2). The paralogy included the entire non-alpha satellite portion of 37914, terminating precisely at the alpha satellite transition. Interestingly, the degree of sequence similarity among the various pericentromeric duplicated segments was virtually identical (96–97%) (Table 1), suggesting that these interchromosomal exchanges occurred at a similar timepoint during evolution (see below).

To assay the extent of duplication of this region throughout the human genome, nine PCR assays were developed based on the non-alpha satellite portion of 37914 and used to screen a

**Table 1.** Database sequence similarity

| Target | Map | L | m | Indel | K | %sim with indels | Coordinates Q | Coordinates T |
|---|---|---|---|---|---|---|---|---|
| Genomic | | | | | | | | |
| AC002038 | 2p11 | 60268 | 58531 | 152 | 0.029 ± 0.001 | 0.969 ± 0.001 | 1–61 389 | 47 856–109 695 |
| AC002038 | 2p11 | 2940 | 2855 | 5 | 0.029 ± 0.003 | 0.969 ± 0.003 | 61 582–64 521 | 140 119–143 065 |
| U52111 | Xq28 | 9698 | 9270 | 28 | 0.046 ± 0.002 | 0.953 ± 0.002 | 51 358–61 150 | 63 156–53 400 |
| AC002041 | 16p11 | 2957 | 2872 | 7 | 0.029 ± 0.003 | 0.970 ± 0.003 | 61 560–64 521 | 123 953–126 919 |
| AL031601[a] | 10q11 | 31473 | 30650 | 84 | 0.027 ± 0.001 | 0.971 ± 0.001 | 1–37 150 | 220 691–252 655 |
| AL031601[a] | 10q11 | 2954 | 2859 | 6 | 0.033 ± 0.003 | 0.966 ± 0.003 | 61 558–64 521 | 134 610–131 652 |
| AC006359[a] | 2p11[b] | 2950 | 2844 | 6 | 0.037 ± 0.004 | 0.962 ± 0.004 | 61 570–64 521 | 31 012–28 057 |
| Genic | | | | | | | | |
| AI027746 | 4q24 | 391 | 379 | 1 | 0.031 ± 0.009 | 0.967 ± 0.009 | 9638–9681, 13 501–1 3612, 18 809–19 043 | 394–351, 350–239, 238–4 |
| AA393779 | 4q24 | 226 | 220 | 1 | 0.027 ± 0.011 | 0.969 ± 0.011 | 8593–8700, 9532–9650 | 226–333, 334–451 |
| AI188518 | 4q24 | 700 | 678 | 2 | 0.032 ± 0.007 | 0.966 ± 0.007 | 42174–42351, 42645–43174 | 1–179, 180–704 |
| ALD4557300 | Xq28 | 1551 | 1470 | 4 | 0.054 ± 0.006 | 0.945 ± 0.006 | 56 923–58 116, 58 262–58 390, 58 540–58 627, 60 927–61 072 | 3604–2379, 2378–2253, 2252–2165, 2164–2019 |

A summary of highly significant sequence homologies to the non-alpha satellite portion of 16p11 BAC, AC002307, within the NR, HTGS and EST sequence databases (BLASTN v2.0.9). The table summarizes the target region identified, the map location of that sequence, the length in bp (L) of sequence compared using ALIGN software, the number of sequence matches (m), the number of insertion deletions (indel) and the % sequence similarity (%sim). The coordinates of query 37914 and target sequence are provided (coordinates Q and T). Genic segments refer to hits in the EST database which have non-processed 'intron/exon' structure.

[a]Working draft sequence in which the query coordinates will likely change over time.

[b]Putative map location.

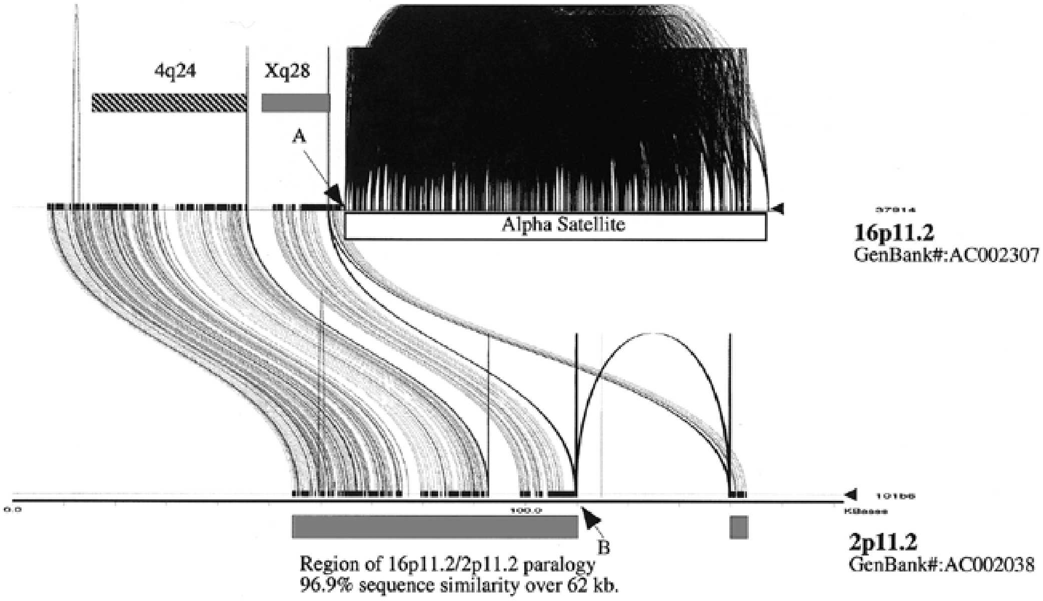


**Figure 2.** Duplications of larger segments between 16p11.2 and 2p11.2 pericentromeric regions. Miropeat analysis was performed on 323 kb of sequence from 16p11.2 BAC clone 37914 (AC002307) and 2p11.2 BAC clone 101B6 (AC002038). Miropeats identifies regions of sequence similarity and displays this similarity information graphically in the positional context of the sequence (vertical line) as black bars delineated by joining lines between the two sequences (http://www.genome.ou.edu/miropeats.html ). Arcs distinguish internal repeat sequences. The analysis was performed using 37914 and 101B6 sequences which had been masked for common repeat elements. As a result, breaks in the sequence similarity are indicated due to the masking of Alu, LINES etc. Not all alpha satellite sequence homology was removed using the repeat-masked alpha satellite consensus. Many of these intrachromosomal repeat sequences are still identified by Miropeats as multiple arcs within 37914. Sequence alignment (ALIGN) of the duplicated segment revealed 96.9% sequence identity over >62 kb of compared 16p11.2/2p11.2 sequence. This larger complex includes a 43 kb segment derived from 4q24 and 9.7 kb of adrenoleukodystrophy sequence originally duplicated from Xq28. A and B demarcate junctions of duplication which have been verified by PCR and direct sequencing from monochromosomal hybrids of chromosomes 16 and 2, respectively.

panel of monochromosomal somatic cell hybrid DNAs. The PCR products were directly sequenced, the sequences aligned and the average pairwise nucleotide similarity calculated (Table 2). All seven of these paralogous STSs (pairs 1–7) amplified more than one chromosome. The sequenced products confirmed the high degree (~96–97%) of sequence simi-

**Table 2.** Paralogous STS PCR

| Primer pair | Oligo name | Sequence | Position in 37914 | Chromosomal duplications | % similarity | GenBank accession nos |
|---|---|---|---|---|---|---|
| 1 | 101-74 | CTGTATCAATCACTGCTGTGCTCAG | 8866–8842 | 16, 2, 4, 10, 22 | 96.4 ± 1.1 | AF183346–AF183350 |
| | 101-110 | TCTTCATGAACCACCTAGATTTGC | 8505–8528 | | | |
| 2 | 101-78 | CTAGTATCAGAGATGTGGCAGAAG | 9307–9330 | 16, 2, 4, 10, 13, 15, Y, 22 | 96.8 ± 0.9 | AF183323– AF183326, AF183366–AF183369 |
| | 101-79 | CAACCAGAATGAGGGGATTTCCTA | 9745–9768 | | | |
| 3 | 101-10 | TATCAAGCTGGTTCCAGGAACTGG | 17064–17087 | 16, 2, 4, 10, 15, Y, 22 | 96.1 ± 0.8 | AF182004–AF182010 |
| | 101-38 | GTACTGAACATGATCCAGTGTGCTG | 17718–11742 | | | |
| 4 | 101-116 | AACTCCTGGTGTTATGAGGGCAAC | 18505–18528 | 16, 2, 4, 10, 22 | 96.9 ± 0.8 | AF183352–AF183356 |
| | 101-118 | AAGAAGTAGGCAGATGATGACAGG | 19055–19032 | | | |
| 5 | 101-11 | CACTTGGTACAATCACCAATGCAAAG | 22829–22854 | 16, 2, 4, 10, 13, 15, 22 | 97.6 ± 0.8 | AF183339–AF183345 |
| | 101-39 | GGAAGCTGTGAAGAAGCTGGTCTC | 22375–22398 | | | |
| 6 | 101-14 | TGGCTGATCTGTCTGACAACAGTG | 37576–37599 | 16, 2, 4, 10, 22 | 96.9 ± 0.7 | AF183333–AF183337 |
| | 101-41 | CAACACCTAGTTGGCCATATAGTCC | 38347–38371 | | | |
| 7 | 101-42 | CAAACAGCTTTGGATCCATAGCCAC | 42801–42825 | 16, 2, 4, 10, 22 | 97.3 ± 0.9 | AF183327–AF183331 |
| | 101-81 | AGTTTCCTGCCTGGGATGGTTCAC | 43152–43175 | | | |
| 8 | 83191 | CACCCGCAGCACCTGGATGTCAGC | 52052–52075 | 16, 2, 10, 22, X | 96.6 ± 1.4 | AF183371–AF183375 |
| | 83192 | TCACAGGCTAGTGGACATGGCAGAC | 51851–51875 | | | |
| 9 | 101-85 | CCTTGTGTGACCAGGTGATCTACC | 61027–61050 | 16, 2, 10, 22, X | 95.4 ± 1.0 | AF183358–AF183362 |
| | 101-86 | ACAGTAGCCATCACTGCACACATG | 60414–60437 | | | |
| 10 | 37914-1 | CATCACAAAGCAGTTTCTCAGAGAGC | 64779–64804 | 16 | NA | AF183377 |
| | 37914-3 | TCCACAACATGGTTTACTTCCAAG | 64349–64372 | | | |
| 11 | 37914a2 | GGCTTTGTGATATGTGCATTCATC | 86785–86808 | 16, 7 | 97.5 ± 0.2 | AF183364–AF183365 |
| | 37914a4 | AAAGCTTTCTGAGAAGCTGCTTTG | 86169–86192 | | | |
| 12 | 37914a5 | GTCTCTTCTTGTTTTTAAGCTGGG | 119607–119625 | 16 | NA | AF183380 |
| | 37914a6 | CAAATATCCCTTTGCAGATCCTTC | 119957–119980 | | | |

This table summarizes the PCR assays used in somatic cell hybrid analysis to investigate the distribution and degree of paralogy among non-homologous chromosomes. The position of each PCR assay with respect to the 37914 reference sequence is summarized in Figure 1. Oligonucleotides beginning with '101' were designed to chromosome 2 BAC sequence (GenBank accession no. AC002038). Chromosomal duplications denote monochromosomal hybrids which were positive for this PCR assay. All hybrid products were sequenced and the degree of sequence similarity calculated as the average of pairwise sequence alignments. Sequences have been deposited into GenBank under accession nos AF182004–AF182009, AF183323–AF183331, AF183333–AF183337, AF183346–AF183350, AF183352–AF183356, AF183358–AF183362, AF183364–AF183369, AF183371–AF183375, AF183377, AF183380.

larity predicted by database searches. Based on the distribution of these paralogous STSs, two distinct domains of duplication could be discerned within the non-alpha satellite segment: a region (~43 kb in length) paralogous to chromosome 4 and a region (9.7 kb in length) paralogous to the X chromosome. These data, taken together with the database sequence similarity searches, indicate that the two duplication domains form part of a larger duplication complex (>62 kb), at least for chromosomes 16p11 and 2p11. Interestingly, these patterns of paralogy do not appear to extend into the alpha satellite portion of 16p11. In fact, the only STS that was specific to chromosome 16 was designed to span across the alpha satellite/non-alpha satellite transition, indicating that this structure is unique to this chromosome.

## Evolution of the alpha satellite junction

In order to evaluate the evolutionary origin of this region, two complementary sets of experiments were undertaken. First, comparative FISH analysis was performed against metaphase chromosomal spreads representing four hominoid species (*Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla* and *Pongo pygmaeus*) and one representative cercopithecoid *(Macaca fascicularis*). Several different subcloned portions of the chromosome 16 region (Table 3 and Fig. 1a) were used as FISH probes. Our analysis revealed that the different duplication domains bordering the alpha satellite domain were of diverse evolutionary origin. The proximal ~43 kb portion of the region, represented by subclone c308a5, cross-hybridized to the syntenic region of 4q24 among all five species examined (Fig. 3a). Both the chromosome 16 cosmid, c308a5, and a clone representing the human 4q24 locus (BC289C17) cross-hybridized only to the human syntenic region of 4q24 among the two most distantly related primates, indicating that this locus represented the most likely ancestral location of this duplication. Similarly, analysis of *ALD* revealed that this 9.7 kb segment cross-hybridized consistently to the distal end of Xq28, defining the ancestral origin of this segment (Fig. 3b). In both of these cases, pericentromeric localizations of these probes were identified only in man and the African apes (*Gorilla–Pan*). Among these species, both quantitative and qualitative hybridization differences were observed in the distribution of FISH signals. In contrast to these regions,

**Table 3.** Summary of comparative FISH analysis

| Species | Probe | | |
|---|---|---|---|
| | 308a5/289c17[a] | ALD9.7 | 445c9 |
| *Homo sapiens* | 4q24, 16p11, 10q11, 22q11, 2p11, 14p11, Yq11 | Xq28, 16p11, 22q11, 10q11, 2p11 | 16p11 |
| *Pan troglodytes* | 4q24, 16p11, 10q11, 22q11 | Xq28, 16p11, 22q11, 10p11 | 16p11 |
| *Gorilla gorilla* | 4q24, 16p11, 10p11 | Xq28, 16p11 (2×)[b], 10p11, 2p11, 14q11 | 16p11 |
| *Pongo pygmaeus* | 4q24 | Xq28 | 16p11 |
| *Macaca mulatta* | 4q24 | Xq28 | 16p11[c] |

A summary of FISH cross-hybridization for three different probes representing the 4q24 duplicated segment (308a5/289c17), the Xq28 duplicated segment (pALD9.7) and the monomeric alpha satellite segment (445c9). Chromosomal band assignments are written with respect to the corresponding human chromosomal assignments and do not refer to the specific chromosomal nomenclature for each species.

[a] 289C17 and 308a5 represent the 4q24 and 16p11 duplicated segment, respectively.

[b] Two distinguishable metaphase signals were observed on each homologue.

[c] Centromeric signals were detected among multiple chromosomes within this species including phylogenetic group XVI.

probes derived from the alpha satellite repeat portion (c445c9 and c305a8) hybridized to the pericentromeric region of phylogenetic chromosome 16 (28) among all species examined (Table 3), indicating that the diverged alpha satellite sequence within 16p11 represents a more ancient structural property of this particular chromosome.

As a second approach to investigate the evolutionary history of the region, phylogenetic analyses were performed on aligned sequences derived from each of the human paralogues and an orang-utan species as a defined outgroup. Phylogeny construction (using both neighbour joining and maximum parsimony methods) for both the 4q24 and Xq28 duplications (Fig. 3c and d) resulted in tree topologies which were consistent with the comparative FISH results. The data suggest that the duplication of the Xq28 and 4q24 genic segments occurred after the separation of orang-utan from the *Homo–Pan–Gorilla* clade [<15 million years ago (mya)] (29,30). Following this initial transposition event to the pericentromeric regions of human chromosomes, secondary events were responsible for the spreading of these paralogous segments among other pericentromeric regions (chromosomes 2, 10, 16 and 22). A clear separation of these two events is supported in both phylogenetic trees by nearly 100% of all bootstrap replicates (data not shown). Using an estimate of evolutionary distance (Kimura two-parameter model) between orthologous 4q24 and Xq28 orang-utan and human sequences and a divergence time of ~14 mya (29,30) between these two species (see Materials and Methods), we estimate that the initial transposition of the 4q24 and Xq28 segments to a pericentromeric region occurred 6.7 ± 2.3 and 6.3 ± 3.0 mya, respectively. Similarly, we have estimated that the subsequent pericentromeric swapping of these segments occurred 3.9 ± 1.2 and 3.5 ± 2.0 mya. These data suggest that the evolutionary timing of transposition and pericentromeric swapping of the 4q24 and Xq28 segments may have been closely synchronized.

## DISCUSSION

### Alpha satellite monomeric organization

Alpha satellite repeats devoid of higher-order structure were a predicted property of centromeric junctions based on early models of restricted unequal crossing-over near the boundaries of the tandem repetitive DNA (31). Sequence analyses of such peripheral DNA (32,33) have generally confirmed this prediction, although in most cases the length of alpha satellite DNA sequenced has been relatively short, making it difficult to completely eliminate the possibility of higher-order structure. Our detailed analysis of ~92 kb of contiguous alpha satellite DNA from chromosome 16 failed to detect any evidence of multimeric arrays. Interestingly, the monomeric repeat sequences characterized in this study show no closer phylogenetic relationship to the chromosome 16 specific higher-order alpha satellite sequence (D16Z2) (19) than to any other chromosome. The greatest sequence similarity was observed with other members of the alpha satellite suprachromosomal 4 (4). MAST database sequence similarity searches of the 466 alpha satellite monomers found the highest conservation $[P(N) = 5\ e^{-52}]$ with non-multimeric alpha satellite previously characterized from a junction from chromosome 7 (32). Comparative FISH analysis (Table 3) indicates that this monomeric alpha satellite structure is found in association with the centromeric region of chromosome 16 among all primates examined (Table 3). Such conservation contrasts sharply with similar analyses of higher-order repeat alpha satellite segments whose chromosome specificity appears to have evolved much more recently (34,35). In this regard, it is important to note that the 92 kb of alpha satellite sequence is virtually devoid of CENP-B box motifs. Instead, the consensus sequence at this site matches nearly perfectly to the previously reported pJα binding site (22). CENPB box sequence motifs are believed to have emerged at the time of the divergence of man, chimpanzee and gorilla (34). This is confirmed by the identification of only pJα binding-site variants among lower primates and old world monkey alpha satellite sequence (36). Taken in total, these data suggest that the alpha satellite domain described in this study did not simply originate as a consequence of degradation at the boundary of the chromosome 16 multimeric array, but that it is a relatively ancient structure which has existed at this chromosome location for the last 20 million years of human/primate evolution.

### The mosaic organization of the 16p11 junction

Analysis of the non-alpha satellite component located at the periphery of 16p11 alpha satellite reveals that the region has

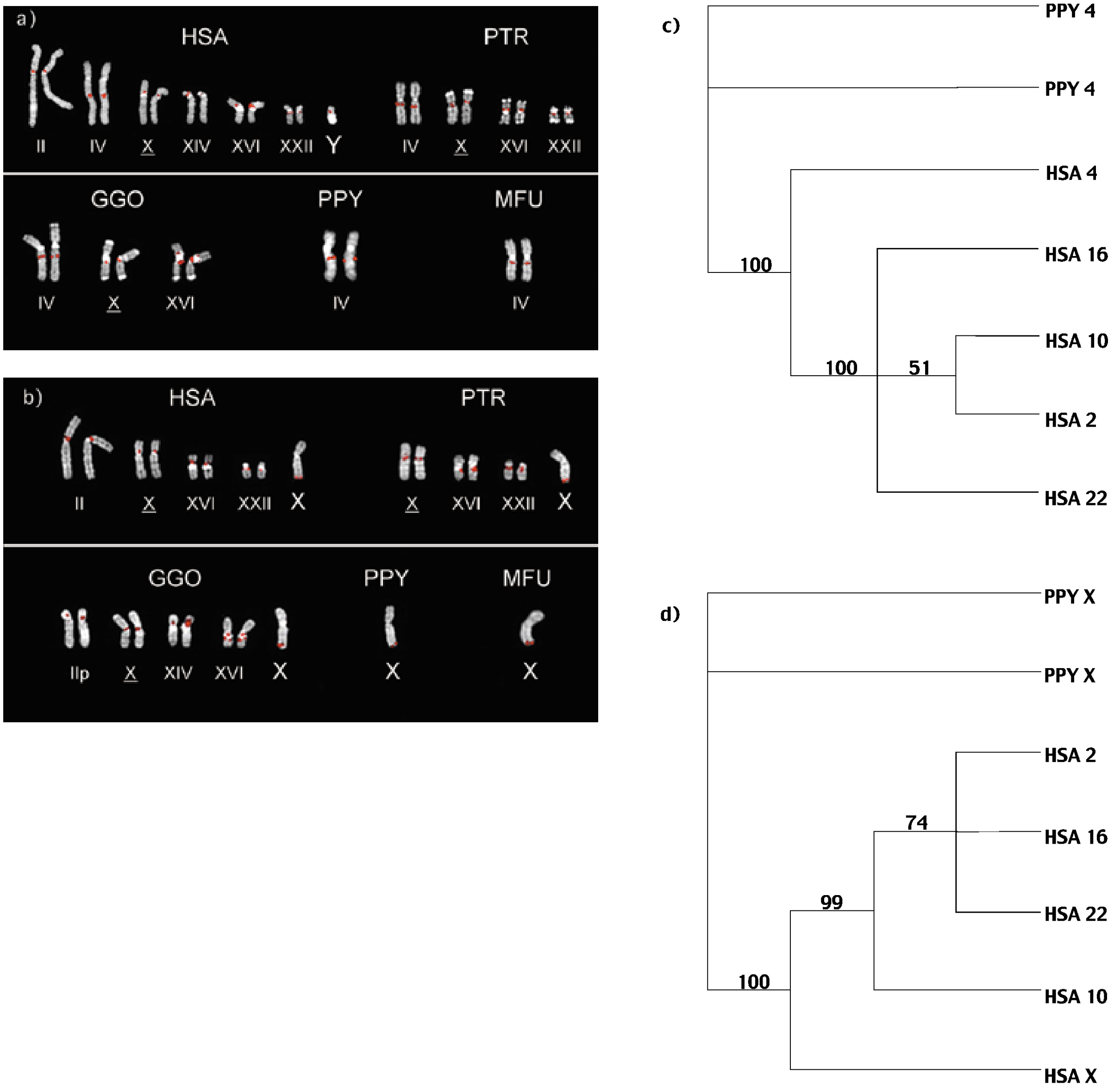


**Figure 3.** Phylogenetic analysis of the 16p11 pericentromeric region. Comparative FISH analysis of (**a**) BC289C17 (representing the 4q24 duplicated segment) and (**b**) pALD [representing the Xq28 duplicated segment against metaphase chromosomal preparations from human (HSA), common chimpanzee (PTR), gorilla (GGO), orang-utan (PPY) and macaque (MMU)]. Chromosomes are designated using phylogenetic group assignments (roman numerals). As a test of hybridization specificity, FISH analyses were performed with probes representative of both duplicated and ancestral loci. Pericentromeric signals were only observed among human, chimpanzee and gorilla chromosomes. Among macaque and orang-utan metaphases, human probes derived from duplicated and ancestral regions hybridized only to regions syntenic to human Xq28 and 4q24 loci, respectively. Phylogenetic tree construction based on maximum parsimony analysis (PAUP) of 2192 bp aligned sequences from the (**c**) 4q24 and 604 bp of aligned (**d**) Xq28 duplicons. Orang-utan sequence taxa (accession nos AF182010, AF183332, AF183338, AF183351, AF183357, AF183363, AF183370, AF183376) from the corresponding regions of 4q24 and Xq28 (designated PPY 4 and PPY X) were specified as an outgroup within each analysis. The number of bootstrap replicates supporting each branchpoint in the tree are indicated.

been the target of remarkable mutational dynamism over the last 10 million years. The available data indicate that the 16p11 junction was formed as a result of at least two distinct evolutionary events (Fig. 4). The first step involved the duplication and transposition of genomic segments from 4q24 and Xq28 to the pericentromeric region of an ancestral hominoid chromosome ~6 mya. Although the precise termini of the 4q24 junction have not yet been determined, it is interesting that the inserted 4q24 and Xq28 sequences occur in close proximity to GC-rich repeat sequences (Fig. 2, indicated by the vertical lines). Similar sequences have been postulated to play a role in the movement of genomic material to the pericentromeric regions of human chromosomes (10,11,37,38). After the formation of this complex mosaic structure, we propose that a much larger segment (>62 kb), which included the original Xq28 and 4q24 duplicons, was spread among the pericentromeric regions of several primate chromosomes, most notably 16p11, 22q11, 2p11 and 10p11 (Fig. 2). The rapid dispersal of these segments among pericentromeric regions would account for the multiplicity of genic paralogues which have been

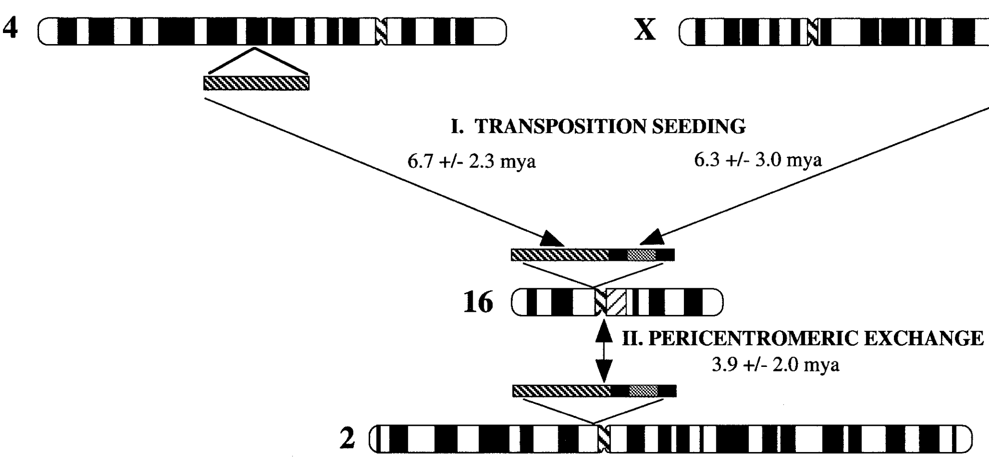


**Figure 4.** Model of interchromosomal pericentromeric duplications. A two-step model for the occurrence of duplicated segments within the pericentromeric regions of human chromosomes is proposed: (I) initial duplication and transposition of genomic segments from 4q24 and Xq28 to a pericentromeric region ~6 mya, followed by (II) the dispersal of this larger complex among multiple pericentromeric regions, ~4 mya. The target chromosome of the initial transposition is not known.

observed among several non-homologous chromosomes (11–13,39). In addition to these events, both the FISH and PCR data (Tables 2 and 3) suggest more complex events for other pericentromeric regions such as 14p11 and Yq11. It is possible that these duplications represent separate, more restricted pericentromeric exchanges among non-homologous chromosomes. Alternatively, these smaller paralogous segments may have formed as a result of secondary deletion events which further reduced the size of the paralogous segment after the initial duplication event between non-homologous pericentromeric regions.

Both the comparative and phylogenetic analyses support a model in which pericentromeric swapping of larger segments occurred near or shortly after the time of the divergence of man and the greater apes (~5 mya). It is surprising that evolutionary estimates for the arrival of the 4q24 and Xq28 modular components within the pericentromeric region is similar (6.4 and 6.6 mya, respectively). This estimate is consistent with the comparative FISH analysis, which indicates that the transpositions occurred after the divergence of man and the orang-utan but before the divergence of the chimpanzee/gorilla lineages. Such temporal synchronization might suggest that the genome of our ancestral hominoid was particularly unstable 6 mya and/or that this specific pericentromeric region accepted non-homologous segments over only a narrow window of evolutionary time. Differences in the cytogenetic distribution of these segments (Table 3) among different primate chromosomes are likely the consequence of secondary duplication or deletion events specific to each primate lineage (6). Although the mechanism for the amplification of the mosaic structure among non-homologous pericentromeric regions is unknown, it is interesting that the monomeric alpha satellite repeat sequence precisely demarcates the duplication boundary of the larger segment between chromosomes 2p11 and 16p11 (Fig. 2). FISH analysis with subclones representing the alpha satellite and non-alpha satellite portions clearly indicate that these two segments have not been subjected to the same level of genomic duplication (Table 3). These observations might suggest that the alpha satellite sequence provided an important mechanistic barrier in the process of 'pericentromeric' spreading or that the larger 4q24–Xq28 duplication was targeted to the alpha satellite boundary during the non-homologous exchange events. In this regard, it will be interesting to determine how long the proximity of the alpha satellite and non-alpha satellite segments has been maintained.

## Pericentromeric exchange

Rapid evolutionary turnover within centromeric regions is neither uncommon nor unexpected (40). Several different mutational processes have been invoked to account for the complex organization of alpha satellite, including gene conversion, interchromosomal exchange and unequal crossing-over (40). Three findings from this analysis, however, are unexpected. First, the boundaries of interchromosomal exchange have been shown to extend to the non-alpha satellite portion located at the periphery of the alpha satellite domain. Second, the material that has been shuttled among the pericentromeric regions is composed of genic segments originating from distinctly non-pericentromeric regions of the hominoid genome. Third, the degree of sequence similarity is high (~97%), indicating that such events occurred more recently (~5 mya) than the origin of suprachromosomal families which was thought to have predated the divergence of man and the great apes (>5 mya) (41). It has been suggested by several groups that the spread of suprachromosomal alpha satellite subfamilies may underlie the origin of duplications of non-alpha satellite located at the periphery (12). The paralogy domains identified in this study, however, clearly do not correlate with the defined suprachromosomal stratification (4). These pieces, therefore, would unlikely have been moved as a consequence of events which spread suprachromosomal families during great ape evolution. It is also unclear how the duplicated segments are organized with respect to other satellite repeat families which have been identified by low resolution techniques at the periphery of higher-order alpha satellite DNA (2). Perhaps mutational processes independent from higher-order repeat alpha satellite turnover are required to explain the preferred integration and rapid dispersal of these segments among pericentromeric regions of human chromosomes. Further high-resolution molecular and cytogenetic analyses are required to resolve the complex organization and evolution of these regions.

The presence of large paralogous segments within pericentromeric regions of non-homologous chromosomes suggests recent interchromosomal exchanges of genetic material. Such transfers of alpha satellite sequences have been postulated to account for the presence of virtually indistinguishable alpha satellite suprachromosomal family sequence on chromosomes 1, 5 and 19 (42,43). Similarly, the homogeneous nature of sequence among the short arms of the acrocentric chromosomes (chromosomes 13, 14, 15, 21 and 22) has been taken as evidence of frequent interchromosomal exchange, possibly due to the close association of these chromosomal segments within the nucleolus (44,45). It is possible that the presence of large duplicated segments within chromosomes 2p11, 10p11, 16p11 and 22q11 facilitates non-homologous associations or that such illegitimate associations in the recent evolution of our species promoted these exchanges. The functional significance of such large paralogous segments located at the junction of alpha satellite DNA remains obscure. One possibility may be that these transposed and duplicated 'genic' segments with their associated 'euchromatic-binding sites' for transcription and splicing may serve effectively as a buffer between euchromatin and heterochromatin associated proteins. Such transition zones might alleviate the potential deleterious position effects of *bona fide* genes residing at the periphery of alpha satellite DNA. Of course, from the evolutionary perspective such malleable regions of the genome could also confer distinct selective advantages to an evolving species. For example, the juxtaposition of different genic segments from diverse regions of the genome within a new chromosomal context could allow for the formation of new genes with potentially new functions. Alternatively the rapid genomic diversification of such regions might affect homologous chromosome pairing, providing a means for meiotic disruption and post-mating genetic barrier between two incipient species/populations. Direct comparative sequencing of orthologous segments from closely related primate species will ultimately be necessary to evaluate the evolutionary plasticity of these and other dynamic regions of the human genome.

## MATERIALS AND METHODS

### Hybridization

A chromosome 16-specific cosmid library (LA16NC02) and two total-genomic libraries (RPCI-11 and CIT-HSP) were probed with PCR amplified inserts representing the Xq28 *ALD* duplication and chromosome 16 alpha satellite sequence. Two degenerate oligonucleotides, 117934 [5′-TTT(CA)-(TA)TT(CG)AGCAGT(GT)TTGAAACACTC] and 117935 [5′-GTGAG(AG)(CT)GAATG(TC)ACACA(TG)CAC], designed to conserved portions of aligned alpha satellite sequence were used to generate by PCR a generic alpha satellite ladder (171 bp) from chromosome 16 monochromosomal somatic cell hybrid DNA (NIGMS GM10567). DNA probes specific to the duplicated portion of Xq28 *ALD* were generated by long-range PCR as previously described (11) (Table 1, PCR products 7, 12 and 16). All probes were radiolabeled and hybridized as described (11) with the exception that the probe was blocked with Cot1 DNA for 1 h at 65°C prior to hybridization.

### Sequence analysis

A random shotgun library for BAC37914 was prepared in pUC18 vector, as previously described (46). Base-calling was performed using phred (47) and sequence assembly was performed using TIGR Assembler. A suite of software tools was used to assess the quality and completeness of the sequence assembly at the end of the random phase. Directed sequencing was performed as suggested by these software tools and as determined by manual inspection of contigs to close any remaining gaps (generally <12/100 kb BAC DNA), and improve the quality of weak regions. Each base was covered either on both strands or by two independent sequencing chemistries (e.g. dye-labeled terminators and dye-labeled primers).

### Monochromosomal somatic cell hybrid analysis

A monochromosomal somatic cell hybrid DNA panel (NIGMS, Human Genetic Mutant Cell Repository Mapping Panel 2) was assayed by PCR to determine the extent of paralogy and the degree of sequence similarity among duplicons. PCR amplification reactions were carried out in a final volume of 20 µl containing 0.20 mM dNTPs (Pharmacia, Piscataway, NJ), 20 pmol of each primer and 0.625 U of *Taq* polymerase in standard 1× reaction buffer (Boehringer Mannheim, Indianapolis, IN). Amplification conditions were identical for PCR primer pairs 1–2 and 5–10 (Table 2): an initial denaturation of 2 min at 95°C, followed by 35 cycles of 30 s at 95°C, 30 s at 55°C and a 45 s extension at 72°C. A final extension of 7 min was carried out at 72°C. PCR reactions with primer pairs 3 and 4 were performed as described with the exception that annealing temperatures of 65 and 60°C were used, respectively. All cycling conditions were optimized for use in a PE 9600 thermocycler (Perkin-Elmer Applied Biosystems, Norwalk, CT). Amplified PCR products were directly sequenced using both forward and reverse PCR primers and dichlororhodamine dye-terminator sequencing chemistry (Perkin-Elmer Applied Biosystems). Reactions were performed following the manufacturer's protocol with the following modifications. Prior to sequencing, 8 µl of each PCR product was treated with 1.5 U exonuclease I and 0.30 U of shrimp alkaline phosphatase to remove excess single-strand DNA and deoxynucleotide triphosphates. Reactions were incubated at 37°C for 5 min and then heat inactivated at 72°C for 15 min. Cycle sequencing reactions were performed in a total reaction volume of 8 µl: 5 µl (30–90 ng) of Exo/SAP-treated product, 20 pmol of sequencing primer, 2 µl of dichlorhodamine dye-terminator mix. All fluorescent traces were analyzed using the Applied Biosystems Model 377 DNA Sequencing System (Perkin-Elmer Applied Biosystems) and the quality of sequence data assessed with PHRED/PHRAP/CONSED software (http://genome.wustl.edu ).

### FISH

Chromosome metaphase spreads were prepared from lymphoblastoid cell lines representative of five hominoid species (*H.sapiens*, *Pan troglodytes*, *G.gorilla*, *Pongo pygmaeus*, *Hylobates lar*) and one cercopithecoid (*M.fascicularis*). Probes were nick-translated with either biotin-16-dUTP or digoxigenin-11-dUTP, and hybridized to chromosomal preparations as previously described (48). A Zeiss Axioskop epifluorescence microscope with a cooled charge-coupled device camera was used to generate digital images. Hybridizations were performed in conjunction with human whole-chromosome painting probes to determine orthologous band intervals relative to the human phylogenetic group assignments, as described by the International Standard for Cytogenetic Nomenclature (28).

### Sequence analysis

BLASTN sequence similarity searches were performed against nr (non-redundant) and htgs (high throughput genomic sequences) divisions of GenBank using, as query, repeat-masked 37914 (AC002307). Repeatmasker v2.0 (http://ftp/genome/washington.edu/cgi-bin/RepeatMasker ) in combination with BLAST and Miropeats (24) was used to determine the presence of duplicated sequence. The absence of higher-order repeat structure was verified by *in silico* restriction enzyme digest using the WebCutter utility (http://www.first-market.com/cutter/cut2.html ). Repeat-masked versions of sequences flanking each of the repeats were systematically searched for the presence of duplicated segments. Pairwise genomic sequence alignments were performed with ALIGN software (http://genome.cs.mitu/edu/align/align.html ). Sim4 software, which optimizes alignment based on known structural properties of exon/intron structure, was used for cDNA to genomic comparisons. Percent similarity was calculated as L(number of matched bases)/L + number of indels × 100. Standard error was estimated as the square root of the binomial distribution. The number of substitutions per 100 bp (K) and its associated variance was determined using Kimura's two-parameter method (49).

### Phylogenetic analysis

Phylogenetic Analysis Using Parsimony (PAUP) v4.0b2 software (50) was used to deduce the phylogenetic relationship among aligned sequences. Insertion and deletion in the alignment were counted as a fifth character state, irrespective of the size of the indel. Parsimony analysis was performed on aligned sequences using the exhaustive search option and the 50% majority-rule consensus tree was determined among all equally parsimonious trees. Orang-utan sequence was specified as the defined outgroup. Bootstrap analysis (1000 branch-and-bound replicates) provided an estimate of the confidence of each branchpoint in the phylogenetic tree. Evolutionary distances and associated standard errors between sequence taxa and within internal nodes of the phylogenetic trees were calculated using the Kimura-Nei two-parameter distance estimates (Molecular Evolutionary Genetics Analysis, MEGA v1.02) (51). Since published rates of nucleotide substitution among pseudogenic sequences may differ from locus to locus (52,53), the rate of nucleotide substitution was calculated for each paralogy domain based on orthologous sequence comparisons between human and orang-utan and an accepted divergence time of ~14 mya for the two species. The timing of all duplication events ($T = K/2r$) was calibrated using this estimate and associated standard errors for each estimate of genetic distance (29).

## ACKNOWLEDGEMENTS

We would like to thank Dr Norman Doggett for kindly providing chromosome 16-specific cosmid clones and reagents. The chromosome-specific gene libraries were constructed under the auspices of the National Laboratory Gene Library Project sponsored by the US Department of Energy. This work was supported, in part, by grants NIH/NIGMS GM58815-01 and NSF DEB9806913 and by a Basil O'Connor Scholar award to E.E.E. (FY98–0712) from the March of Dimes Birth Defects Foundation. The financial support of AIRC Telethon (grant E.672 to M.R. and E.962 to N.A.), and cofin98-MURST is gratefully acknowledged.

## REFERENCES

1. Jackson, M.S., See, C.G., Mulligan, L.M. and Lauffart, B.F. (1996) A 9.75-Mb map across the centromere of human chromosome 10. *Genomics*, **33**, 258–270.
2. Lee, C., Wevrick, R., Fisher, R.B., Ferguson-Smith, M.A. and Lin, C.C. (1997) Human centromeric DNAs. *Hum. Genet.*, **100**, 291–304.
3. Willard, H.F. and Waye, J.S. (1987) Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.*, **25**, 207–214.
4. Alexandrov, I.A., Medvedev, L.I., Mashkova, T.D., Kisselev, L.L., Romanova, L.Y. and Yurov, Y.B. (1993) Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. *Nucleic Acids Res.*, **21**, 2209–2215.
5. Willard, H.F. (1991) Evolution of alpha satellite. *Curr. Opin. Genet. Dev.*, **1**, 509–514.
6. Jackson, M.S., Rocchi, M., Thompson, G., Hearn, T., Crosier, M., Guy, J., Kirk, D., Mulligan, L., Ricco, A., Piccininni, S. *et al.* (1999) Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications, and unstable sequences with homologies to telomeric and other centromeric locations. *Hum. Mol. Genet.*, **8**, 205–215.
7. Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. and Walters, L. (1998) New goals for the U.S. Human genome project: 1998–2003. *Science*, **282**, 682–689.
8. Arnold, N., Stanyon, R., Jauch, A., O'Brien, P. and Wienberg, J. (1996) Identification of complex chromosome rearrangements in the gibbon by fluorescent *in situ* hybridization (FISH) of a human chromosome 2q specific microlibrary, yeast artificial chromosomes, and reciprocal chromosome painting. *Cytogenet. Cell Genet.*, **74**, 80–85.
9. Zachau, H. (1993) The immunoglobulin κ locus—or—what has been learned from looking closely at one-tenth of a percent of the human genome. *Gene*, **135**, 167–173.
10. Eichler, E.E., Lu, F., Shen, Y., Antonacci, R., Jurecic, V., Doggett, N.A., Moyzis, R.K., Baldini, A., Gibbs, R.A. and Nelson, D.L. (1996) Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.*, **5**, 899–912.
11. Eichler, E.E., Budarf, M.L., Rocchi, M., Deaven, L.L., Doggett, N.A., Baldini, A., Nelson, D.L. and Mohrenweiser, H.W. (1997) Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.*, **6**, 991–1002.
12. Regnier, V., Meddeb, M., Lecointre, G., Richard, F., Duverger, A., Nguyen, V.C., Dutrillaux, B., Bernheim, A. and Danglot, G. (1997) Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum. Mol. Genet.*, **6**, 9–16.
13. Zimonjic, D., Kelley, M., Rubin, J., Aaronson, S. and Popescu, N. (1997) Fluorescence *in situ* hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. *Proc. Natl Acad. Sci. USA*, **94**,11461–11465.

14. Ritchie, R.J., Mattei, M.G. and Lalande, M. (1998) A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Hum. Mol. Genet.*, **7**, 1253–1260.
15. Potier, M., Dutriaux, A., Orti, R., Groet, J., Gibelin, N., Karadima, G., Lutfalla, G., Lynn, A., Van Broeckhoven, C., Chakravarti, A. *et al.* (1998) Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics*, **51**, 417–426.
16. Amos-Landgraf, J.M., Ji, Y., Gottlieb, W., Depinet, T., Wandstrat, A.E., Cassidy, S.B., Driscoll, D.J., Rogan, P.K., Schwartz, S. and Nicholls, R.D. (1999) Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am. J. Hum. Genet.*, **65**, 370–386.
17. Christian, S.L., Fantes, J.A., Mewbornm S.K., Huang, B. and Ledbetter, D.H. (1999) Large genomic duplicons map to sites of instability in the Prader–Willi/Angelmann syndrome chromosome region (15q11–q13). *Hum. Mol. Genet.*, **8**, 1025–1037.
18. Reiter, L.T., Murakami, T., Koeuth, T., Gibbs, R.A. and Lupski, J.R. (1997) The human COX10 gene is disrupted during homologous recombination between the 24 kb proximal and distal CMT1A-REPs. *Hum. Mol. Genet.*, **6**, 1595–1603.
19. Greig, G.M., England, S.B., Bedford, H.M. and Willard, H.F. (1989) Chromosome-specific alpha satellite DNA from the centromere of human chromosome 16. *Am. J. Hum. Genet.*, **45**, 862–872.
20. Bailey, T. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 28–36.
21. Vissel, B. and Choo, K.H. (1987) Human alpha satellite DNA—consensus sequence and conserved regions. *Nucleic Acids Res.*, **15**, 6751–6752.
22. Gaff, C., du Sart, D., Kalitsis, P., Iannello, R., Nagy, A. and Choo, K.H. (1994) A novel nuclear protein binds centromeric alpha satellite DNA. *Hum. Mol. Genet.*, **3**, 711–716.
23. Sonnhammer, E. and Durbin, R. (1995) A dot matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–GC10.
24. Parsons, J. (1995) Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.*, **11**, 615–619.
25. Warburton, D., Gersen, S., Yu, M., Jackson, C., Handelin, B. and Housman, D. (1990) Monochromosomal rodent-human hybrids from microcell fusion of human lymphoblastoid cells containing an inserted dominant selectable marker. *Genomics*, **6**, 358–366.
26. Neil, D.L., Villasante, A., Fisher, R.B., Vetrie, D., Cox, B. and Tyler-Smith, C. (1990) Structural instability of human tandemly repeated DNA sequences cloned in yeast artificial chromosome vectors. *Nucleic Acids Res.*, **18**, 1421–1428.
27. Tyler-Smith, C. and Brown, W.R. (1987) Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.*, **195**, 457–470.
28. ISCN (1985) Report of the standing committee on human cytogenetic nomenclature. *Birth Defects*, **21**, 1–117.
29. Li, W. (1997) *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
30. Goodman, M. (1999) The genomic record of humankind’s evolutionary roots. *Am. J. Hum. Genet.*, **64**, 31–39.
31. Smith, G.P. (1976) Evolution of repeated DNA sequences by unequal crossover. *Science*, **191**, 528–535.
32. Wevrick, R., Willard, V. and Willard, H. (1992) Structure of DNA near long tandem arrays of alpha satellite DNA at the centromere of human chromosome 7. *Genomics*, **14**, 912–913.
33. Mashkova, T., Oparina, N., Alexandrov, I., Zinovieva, O., Marusina, A., Yurov, Y., Lacroix, M.H. and Kisselev, L. (1998) Unequal cross-over is involved in human alpha satellite DNA rearrangements on a border of the satellite domain. *FEBS Lett.*, **441**, 451–457.
34. Haaf, T. and Willard, H. (1997) Chromosome-specific alpha-satellite DNA from the centromere of chimpanzee chromosome 4. *Chromosoma*, **106**, 226–232.
35. Archidiacono, N., Antonacci, R., Marzella, R., Finelli, P., Lonoce, A. and Rocchi, M. (1995) Comparative mapping of human alphoid sequences in great apes using fluorescence *in situ* hybridization. *Genomics*, **25**, 477–484.
36. Romanova, L.Y., Deriagin, G.V., Mashkova, T.D., Tumeneva, I.G., Mushegian, A.R., Kisselev, L.L. and Alexandrov, I.A. (1996) Evidence for selection in evolution of alpha satellite DNA: the central role of CENP-B/pJ alpha binding region. *J. Mol. Biol.*, **261**, 334–340.
37. Borden, P., Jaenichen, R. and Zachau, H. (1990) Structural features of transposed human Vκ genes and implications for the mechanism of their transpositions. *Nucleic Acids Res.*, **18**, 2101–2107.
38. Eichler, E.E., Archidiacono, N. and Rocchi, M. (1999) CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.*, **9**, 1048–1058.
39. Zimmer, F., Hameister, H., Schek, H. and Zachau, H. (1990) Transposition of human immunoglobulin V kappa genes within the same chromosome and the mechanism of their amplification. *EMBO J.*, **9**, 1535–1542.
40. Warburton, P. and Willard, H. (1996) Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In Jackson, M., Strachan, T. and Dover, G. (eds), *Human Genome Evolution*. BIOS Scientific, Guildford, UK, pp. 121–145.
41. Durfy, S.J. and Willard, H.F. (1990) Concerted evolution of primate alpha satellite DNA. Evidence for an ancestral sequence shared by gorilla and human X chromosome alpha satellite. *J. Mol. Biol.*, **216**, 555–566.
42. Baldini, A., Smith, D.I., Rocchi, M., Miller, O.J. and Miller, D.A. (1989) A human alphoid DNA clone from the *Eco*RI dimeric family: genomic and internal organization and chromosomal assignment. *Genomics*, **5**, 822–828.
43. Hulsebos, T., Schonk, D., van Dalen, I., Coerwinkel-Driessen, M., Schepens, J., Ropers, H.H. and Wieringa, B. (1988) Isolation and characterization of alphoid DNA sequences specific for the pericentric regions of chromosomes 4, 5, 9, and 19. *Cytogenet. Cell Genet.*, **47**, 144–148.
44. Gravholt, C.H., Friedrich, U., Caprani, M. and Jorgensen, A.L. (1992) Breakpoints in Robertsonian translocations are localized to satellite III DNA by fluorescence *in situ* hybridization. *Genomics*, **14**, 924–930.
45. Choo, K.H. (1990) Role of acrocentric cen-pter satellite DNA in Robertsonian translocation and chromosomal non-disjunction. *Mol. Biol. Med.*, **7**, 437–439.
46. Loftus, B., Kim, U., Sneddon, V., Kalush, F., Brandon, R., Fuhrmann, J., Mason, T., Crosby, M., Barnstead, M., Cronin, L. *et al.* (1999) Genome duplications and other features in 12 Mbp of DNA sequence from human chromosome 16p and 16q. *Genomics*, **60**, 295–308.
47. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
48. Lichter, P., Tang, C.J., Call, K., Hermanson, G., Evans, G.A., Housman, D. and Ward, D.C. (1990) High-resolution mapping of human chromosome 11 by *in situ* hybridization with cosmid clones. *Science*, **247**, 64–69.
49. Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.
50. Swofford, D.L. and Begle, D.P. (1993) *PAUP: Phylogenetic Analysis Using Parsimony*. Illinois Natural History Survey, Champaign, IL.
51. Kumar, S., Tamura, K. and Nei, M. (1993) *MEGA: Molecular Evolutionary Genetic Analysis, version 1.0*. Pennsylvania State University, University Park, PA.
52. Miyata, T. and Yasunaga, T. (1981) Rapidly evolving mouse alpha-globin-related pseudo gene and its evolutionary history. *Proc. Natl Acad. Sci. USA*, **78**, 450–453.
53. Miyamoto, M., Slightom, J. and Goodman, M. (1987) Phylogenetic relations of humans and apes from DNA sequences in the psi eta-globin region. *Science*, **238**, 369–373.