Genomic Sequence and Transcriptional Profile of the Boundary Between Pericentromeric Satellites and Genes on Human Chromosome Arm 10p

Jane Guy,[1] Tom Hearn,[1,6] Moira Crosier,[1] Jonathan Mudge,[1] Luigi Viggiano,[2] Dirk Koczan,[3] Hans-Jurgen Thiesen,[3] Jeffrey A. Bailey,[4] Julie E. Horvath,[4] Evan E. Eichler,[4] Mark E. Earthrowl,[5] Panos Deloukas,[5] Lisa French,[5] Jane Rogers,[5] David Bentley,[5] and Michael S. Jackson[1,7]

[1]The Institute of Human Genetics, The International Centre for Life, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 3BZ, UK; [2]DAPEG, Sezione di Genetica, Universita' di Bari, Bari 70126, Italy; [3]Institute of Immunology, University of Rostock, Rostock 18055, Germany; [4]Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA; [5]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Contiguous finished sequence from highly duplicated pericentromeric regions of human chromosomes is needed if we are to understand the role of pericentromeric instability in disease, and in gene and karyotype evolution. Here, we have constructed a BAC contig spanning the transition from pericentromeric satellites to genes on the short arm of human chromosome 10, and used this to generate 1.4 Mb of finished genomic sequence. Combining RT-PCR, in silico gene prediction, and paralogy analysis, we can identify two domains within the sequence. The proximal 600 kb consists of satellite-rich pericentromerically duplicated DNA which is transcript poor, containing only three unspliced transcripts. In contrast, the distal 850 kb contains four known genes (*ZNF248*, *ZNF25*, *ZNF33A*, and *ZNF37A*) and up to 32 additional transcripts of unknown function. This distal region also contains seven out of the eight intrachromosomal duplications within the sequence, including the p arm copy of the ~250-kb duplication which gave rise to *ZNF33A* and *ZNF33B*. By sequencing orthologs of the duplicated *ZNF33* genes we have established that *ZNF33A* has diverged significantly at residues critical for DNA binding but *ZNF33B* has not, indicating that *ZNF33B* has remained constrained by selection for ancestral gene function. These results provide further evidence of gene formation within intrachromosomal duplications, but indicate that recent interchromosomal duplications at this centromere have involved transcriptionally inert, satellite rich DNA, which is likely to be heterochromatic. This suggests that any novel gene structures formed by these interchromosomal events would require relocation to a more open chromatin environment to be expressed.

[Supplemental material is available online at www.genome.org and also at http://www.ncl.ac.uk/ihg/10p11.htm. The sequence data from this study have been submitted to EMBL under accession nos. AL391686, AL161931, AL133350, AL121927, AL132657, AL135791, AL132659, AL117337, AL117339, AL132658, AL133217, AL133216, AJ245587, AJ245588, AJ251655, AJ275023–AJ275036, AJ250940–AJ250950, AJ275024–AJ275036, AJ492195, AJ492196, AJ491691–AJ491697. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: W. Amos.]

Despite the publication of the draft sequence (IHGSC 2001; Venter et al. 2001) there are still significant obstacles to be overcome if the sequence map of the human genome is to be completed and accurately annotated. One of the most pressing of these obstacles is the poor quality of data close to human centromeres and telomeres where it has proved particularly difficult to construct clone-based maps. Recently, there has been a concerted effort to anchor telomeres to sequence data using half-YAC clones which has resulted in the successful integration of 32 telomeres within the working draft (Reithman et al. 2001). In contrast, pericentromeric satellites and the DNA flanking them remain poorly represented within finished human sequence.

Pericentromeric satellites, which include the classical satellites and β satellite, can form homogeneous arrays several megabases in length and one or more of these arrays is linked, at the level of cytogenetic resolution, to the centromeric α satellite on most human chromosomes (for review, see Lee et al. 1997). Although pericentromeric satellites have no known function, the fact that repetitive sequences can repress transcription in a wide variety of eukaryotes (Henikoff 1990; Milot et al. 1996), and that transcriptional inactivation can involve physical relocation of sequences to centromeric domains (An-

[6]Present address: Division of Human Genetics, Southampton University, The Duthie Building, Tremona Road, Southampton SO16 6YD, UK.
[7]Corresponding author.
E-MAIL mjackson@ncl.ac.uk; FAX +44 191 241 8666.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.644503. Article published online before print in January 2003.

drulis et al. 1998; Garrick et al. 1998; Francastel et al. 1999), suggests that they play a role in the modulation of gene expression. It is also now clear that sequences close to these satellites are unusual in that they are prone to duplication both within and between chromosomes, and it is this which has hampered mapping efforts. For instance, analysis of the human draft sequence has established that pericentromeric DNA is enriched 3.1-fold for intrachromosomal duplications, and 4.5-fold for interchromosomal duplications (Bailey et al. 2001), while analysis of chromosome 22 has established that more than 50% of interchromosomal duplications on this chromosome are located in the most centromeric 1.5 Mb of sequence (Bailey et al. 2002).

Many of the intrachromosomal duplications in pericentromeric regions are members of chromosome-specific low copy repeat families (LCRs), some of which are responsible for common microdeletion/duplication syndromes (for review, see Stankiewicz and Lupski 2002). These appear to have formed relatively recently during primate evolution (Keller et al. 1999; Park et al. 2002). In contrast, pericentromeric interchromosomal duplications characterized to date have no associated pathology. However, these duplications are of evolutionary interest for several reasons. First, they evolve by an unusual two-step process termed pericentromeric directed duplication (Eichler et al. 1997), where initial duplication into a centromere proximal location is followed by repeat-mediated duplication between chromosomes (Horvath et al. 2000b; Luitjen et al. 2000). Consistent with this, the distribution of some duplications correlates with the distribution of specific repeats including α satellite superfamily 2 (Regnier et al. 1997), CAGGG repeats (Eichler et al. 1999) and satellite 3 (Guy et al. 2000). Second, repeated rounds of duplication have created tracts of pericentromeric sequence where exons from different genes have been juxtaposed, which has led to the suggestion that these regions are breeding grounds of biological novelty (Eichler et al. 1997; 1999) and may have contributed to the increased complexity of the human proteome relative to other sequenced eukaryotes (Eichler 2001).

To assess the impact of pericentromeric sequence movement upon transcriptional activity, and investigate the relationship between sequence organization and chromatin state, sequence data from rigorously assembled contigs, together with experimental verification of in silico predicted gene structures is required. While whole chromosome sequences have included pericentromeric satellites as endpoints of contig construction (Dunham et al. 1999; Hattori et al. 2000; Deloukas et al. 2001), they have not included experimental verification of linked putative genes. To date, only two analyses of pericentromeric sequence organization have also investigated transcriptional potential. An analysis of 1 Mb of 10q11 between satellite 3 arrays and the *RET* proto-oncogene (Guy et al. 2000) led to the suggestion that two distinct pericentromeric sequence domains exist on this chromosome arm—a transcriptionally inert proximal domain rich in satellites and interchromosomal duplications, and a distal, transcriptionally active domain rich in intrachromosomal duplications. Analysis of the Cat Eye syndrome critical region, which lies distal of satellites on the long arm of chromosome 22, also identified a proximal domain rich in interchromosomally duplicated DNA which is transcript poor (Footz et al. 2001).

To further investigate the association between interchromosomal duplication, satellite sequences, and transcription, we have extended our analyses to the short arm of human chromosome 10. In this report, we describe the contig construction, sequence generation, and transcriptional analysis of ~1.4 Mb of DNA linking pericentromeric satellites to genes on 10p11. The proximal ~560 kb of sequence consists of satellite arrays linked by interchromosomally duplicated DNA with little evidence of transcriptional activity. The distal ~890 kb contains over 30 transcripts, including a cluster of four ZNF genes, together with all but one of the intrachromosomal duplications identified within the sequence. These observations are strikingly similar to the sequence organization in 10q11, and support the domain model of pericentromeric sequence organization already proposed for this chromosome (Guy et al. 2000). Furthermore, they indicate that the vast majority of recent pericentromeric directed duplication events on chromosome 10 are transcriptionally inert due, presumably, to the repressive effects of linked satellite sequences. These results have important implications both for the role of these rearrangements in gene formation, and for the completion of chromosome-wide sequencing efforts.

RESULTS

Contig Construction and Sequence Acquisition

A BAC contig linking pericentromeric satellites to genes in 10p11 was constructed by screening publicly available libraries with markers from an existing 10p11 YAC contig (Jackson et al. 1996; 1999; see Methods). Restriction fragment and sequence-based analyses of somatic cell hybrids (Tunnacliffe et al. 1994) and 10p11 YACs (Jackson et al. 1996) were used to distinguish 10p11 clones from paralogs on 10q (e.g., 14-kb 10p11-specific fragment in BACs 59J12 and 508N22; Fig. 1A) and from paralogs on other chromosomes (e.g., 5-kb 10p-specific fragment in BAC 291L22, Fig. 1B). Southern analyses also confirmed that the BAC contig extended into the satellite 3 sequences present within the existing YAC contig (comigrating BAC and YAC fragments in Fig. 1C), and probably extend beyond it into more proximal satellite sequences (fragments specific to BAC 291L22; Fig. 1C). A further contig was seeded following the identification of clones containing an *ALD* paralog from 10p11 (bA453N3; Horvath et al. 2000a), and was also found to contain satellite 3 sequences (data not shown). The relative position of the two contigs within the PFGE map could be established because all satellite 3 sequences on 10p11 are known to be confined to the 3-Mb *Bss*HII fragment which contains the centromeric 2.2-Mb α satellite array, *D10Z1* (See legend to Fig. 1D). The fact that the satellite 3 arrays on this chromosome arm are approximately 150 kb in size (Jackson et al. 1996) made it unlikely that further walking using BACs which terminate in satellite sequence would close the gap between the contigs. As a result, mapping efforts were suspended and the tiling path shown in Figure 1D was chosen from the full-depth contig (see http://www.ncl.ac.uk/ihg/10p11.htm) and sequenced.

A total of 1,451,141 bp of finished sequence data were generated (contigs of 1,165,101 bp and 286,040 bp, see Methods) which extends ~250-kb proximal and 1 Mb distal of the most telomeric array of classical satellite sequence in 10p11.1. The larger contig spans the p arm cluster of the duplicated ZNF genes which flank the centromere of this chromosome (Tunnacliffe et al. 1993). The GC content of the entire sequence is 40.40%, and this remains relatively constant throughout both contigs, reaching peaks of 45%–46% close to the satellite 3 arrays. This is in contrast to the equivalent region in 10q11, where a steady increase from ~35%—~55% is

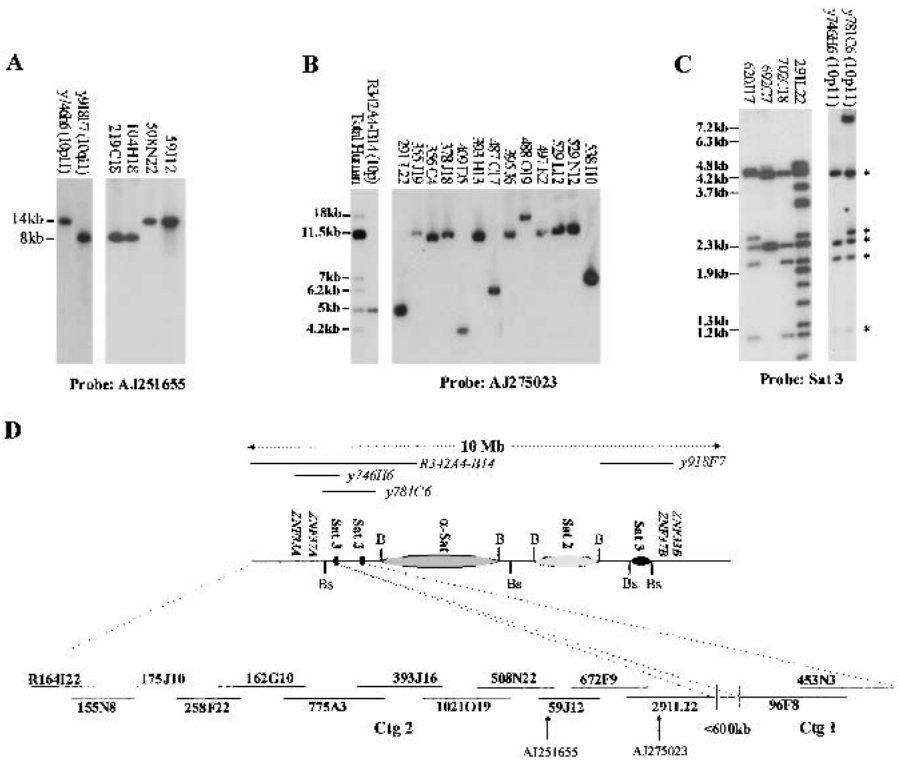

Figure 1 Construction of 10p11 contigs. (*A–C*) Southern hybridization analyses used to assess chromosomal/regional origin of BACs. The map position/sequence content of reference YACs (y746H6, y918F7, and y781C6) and a somatic cell hybrid (R342A4-B14) is given in parenthesis, and fragment sizes are given in kilobase pairs (kb). All DNAs were digested with *Eco*R1. The 10p11 probes are defined by their accession numbers and the satellite 3 probe used was pHS5 (Cooke and Hindley 1979). The regional localization of all BACs sequenced was confirmed using these resources in combination with FISH analyses (data not shown). (*D*) BAC tiling path in relation to known pericentromeric satellite organization. The approximate position of *Bss*HII (Bs) and *Bam*H1 (B) sites which define the satellite arrays (Jackson et al. 1993), and the duplicated ZNF genes which flank the centromere (Tunnacliffe et al. 1993), are shown. The content of YACs and somatic cell hybrid used in panels (*A–C*) is also indicated above the satellite map. All BACs in both contigs are from the RPCI11 library, with the exception of R164I22, which was obtained from Research Genetics (see methods). The orientation of contig 1 and the estimate of the gap size between the contigs is based on sequence identity, number, size, and known position of satellite arrays (see text). The full depth contig from which tiling path 2 was chosen can be viewed at http://www.ncl.ac.uk/ihg/10p11.htm.

observed moving from centromeric satellites to the *RET* proto-oncogene. (Guy et al. 2000). In total, 60.56% of the sequence is accounted for by satellites and interspersed repeats, with satellites/simple repeats, SINEs, LINES, and LTRs accounting for 11.33%, 14.03%, 24.66%, and 8.32% of the sequence, respectively. The figures for interspersed repeats are typical for human autosomal DNA (IHGSC 2001, Venter et al. 2001).

Proximal 10p11 Is Gene Poor

The principle gene-related features identified within the sequence are shown in Figure 2A and Table 1. There is unequivocal evidence for the existence of four genes in the sequence, *ZNF33A*, *ZNF37A*, *ZNF25*, and *ZNF248*, all of which have been identified previously (Tunnacliffe et al. 1993). They are all ZNF genes of the C_2H_2 KRAB subfamily and all have CpG islands at their 5′ ends. Northern analyses using probes from the linker-coding region or 3′ UTR of each gene confirmed that all are expressed in a wide range of adult tissues, with multiple transcripts detected for *ZNF248*, *ZNF33A*, and *ZNF37A* (data not shown). The encoded proteins are 579 (ZNF248), 456 (ZNF25), 810 (ZNF33A), and 561 (ZNF37A) amino acids in length, with predicted molecular weights of 67.1, 53.5, 94.4, and 65.4 kD, respectively, while the KRAB domains of these proteins show good agreement with a KRAB consensus sequence (Agata et al. 1999), strongly suggesting that they are able to repress transcription via interaction with TIF1β. The zinc finger domains contain seven (*ZNF248*) to 16 (*ZNF33A*) tandemly arrayed Krüppel-type zinc fingers (see http://www.ncl.ac.uk/ihg/10p11.htm for details of gene structures and KRAB domain alignment).

In addition to the ZNF genes, there are 43 unassigned ESTs or EST clusters which share >96% identity to the 10p11 sequence (Table 2). However, none of these are associated with large open reading frames or ab initio gene predictions (with the exception of pseudogenes such as *HSD17B7*ψ), and ten contain interspersed repeats. Furthermore, alignment to related sequence within the human draft (see Methods) indicated that 11 ESTs with high identity to the sequence are derived not from 10p11, but from paralogous loci on other chromosomes. Strikingly, all 11 of these ESTs are related to the proximal half of the sequence (Table 1), leaving only three bona fide ESTs within the first 600 kb (AL045218, AW499533, and AW854054) compared to 36 Genes/ESTs in the distal 850 kb. This marked variation in transcriptional activity between the proximal and distal 10p11 sequence mirrors the pattern previously established for the equivalent region of 10q11 (Guy et al. 2000).

To confirm transcription of these ESTs, 15 were analyzed by RT-PCR using cDNAs derived from six adult and four fetal human tissues. The results of these analyses are summarized in Table 2 and examples of RT-PCR analyses are shown in Figure 3. Ten primer pairs (~30 kb in Ctg. 1, ~133 kb, ~351kb, ~385 kb, 754 kb, 785 kb, 785 kb, 1062 kb, 1075 kb, and 1099 kb in Ctg. 2) produced amplification from one or more cDNA, indicating diverse expression patterns, although transcription could not be confirmed for 5 loci. It is possible that some of these ESTs represent the 3′ termini of further genes, or functional noncoding RNAs. However, the lack of gene structures associated with these ESTs, the fact that only one is spliced (AA927631, ~350 kb Ctg 2), and the low number of ESTs in each cluster (many represented by a single EST; Table 1), suggests that many are likely to represent leaky or aberrant transcription.

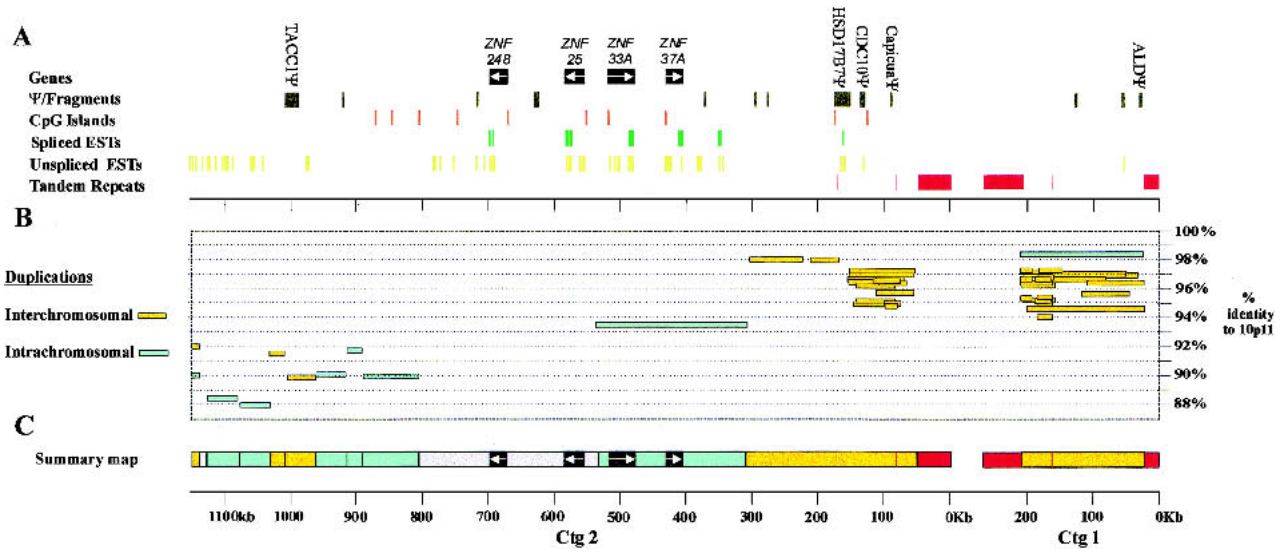

Figure 2 Overview of 10p11 sequence data. (*A*) Principle features of sequence: The position of known genes (Table 1) and pseudogene fragments (http://www.ncl.ac.uk/ihg/10p11.htm), CpG islands, spliced and unspliced ESTs (Table 1), and tandem repeats (Table 3). The orientation of gene transcription is indicated with white arrows. (*B*). Paralogs of the 10p11 sequence within the human draft. These have been divided into interchromosomal (yellow) and intrachromosomal (blue). The position of each independent BLAST hit greater than 5 kb in length within the EMBL-NR and EMBL-HTG databases are represented by a horizontal line. Both the extent of the BLAST hit within the 10p11 sequence (x-axis) and the % identity to the 10p11 sequence (y-axis) is indicated (see Methods). Deletions and insertions within individual sequence matches are not shown. (*C*) Summary map showing genes, satellite repeats, and duplications for reference with later Figures. Because satellite repeats anchor this map, the physical scale of each contig is indicated in 100 kb intervals moving away from the centromere towards 10pter (*right* to *left*). For details of sequence analysis see Methods. The sequence composition of the ~27–600 kb gap between contigs 1 and 2 is unknown.

Satellite Arrays Are Confined to the Proximal 500 kb of Sequence

Chromosome 10 is one of the few human metacentric chromosomes where the relative position of centromeric satellite arrays has been accurately established (Jackson et al. 1993). To integrate the sequence data with this information, RepeatMasker (http://repeatmasker.genome.washington.edu) and Tandem Repeat Finder (Benson 1999) were used to characterize repeats within the two contigs (Fig. 2A and Table 3). The termini of contig 1 are defined by satellite 3 arrays which are ~25 kb and ~78 kb in length, while the proximal termini of contig 2 is defined by a ~45 kb array of the same satellite. Since it is known that there are two ~150-kb arrays of satellite 3 in 10p11 (Jackson et al. 1993) one satellite block from each contig is likely to be derived from the same array. Consistent with this, the 78-kb array in contig 1 and the 45-kb array in contig 2 have similar higher order periodicities (Table 3). This allows us to tentatively infer the orientation of contig 1 (as shown) and suggests that the gap between the two contigs may be as little as ~27 kb. There are only four other tandem repeats >300 bp in length within the sequence: A CATTT repeat (similar to those identified in 10q11; Guy et al. 2000), a highly homogeneous GC-rich repeat with a 49 bp periodicity which is found in many subtelomeric regions, and a repeat consisting of three tandem copies of a 121 bp sequence which includes an Alu fragment. The most striking aspect of these repeats is their distribution; they are confined to the proximal third of the 10p11 sequence, with no tandem repeats in the distal 930 kb of sequence. This is also very similar to the pattern observed in the equivalent ~1 Mb of 10q11 (Guy et al. 2000).

Eighty Percent of the 10p11 Sequence Is Duplicated Elsewhere in the Genome

Pericentromeric regions contain an excess of duplicated DNA, the organization and behavior of which remains poorly understood (Eichler 2001). To identify all human sequences paralogous to 10p11, we queried the nonredundant (EMBL-NR) and High Throughput Genomic (HTG) divisions of EMBL with masked 10p11 sequence (see Methods). This identified 52 independent paralogs of the 10p11 sequence >5 kb in length (Fig. 2B) and established that essentially the entire sequence, with the exception of a ~280-kb tract containing the *ZNF25* and *ZNF248* genes (positions 527698–808106 of contig 2), has been recently duplicated within the human genome. An analysis of these paralogs, which share 87.9%–98.3% identity to 10p11 sequence, is shown in Figure 4. The majority of paralogs of sequences in contig 1 have been mapped to pericentromeric locations within the human draft and share between 95.1% and 98.3% identity to 10p11 (Fig. 4A). This suggests that they have evolved almost exclusively through recent satellite-associated pericentromeric-directed duplication events. Consistent with this, one or both termini of identity between 10p11 and the majority of paralogs falls within the satellite 3 sequence or the short CATTT repeat at position 168–170 kb (Fig. 4A). Furthermore, all paralogs are linked to satellite sequences such as α satellite, β satellite, or hsrep522 either within the same BAC clone or within the human draft (data not shown). Despite this, four of the paralogs cannot be accurately placed within the draft (within AC024036, AC069172, AC024972, and AC010791) and map positions must be viewed as provisional given the frequency of false overlaps and incorrect assignments within draft data

Table 1. ESTs and Genes Within the 10p11 Sequence

Region of identity (contig1), contig2	GenBank acc. no/ unigene cluster (No.)	% match	Features (repeat/*gene name*)	Genome origin (if not 10q11)
(30269-30688)	AL045218	99		
(65078-69657)	AI797613/Hs135840 (6)	98		4q24 (ap001860)
(70352-70426)	AI740992/Hs133165 (3)	97		4q24 (ap001860)
(184338-183720)	AW974557	98		2q11 (al445993)
48769-49073	AI284091	98	L1	1q42.11 (al365438)
63525-72216	AI971943	97		1p36.33 (ac004908)
85121-84555	BE145230	97	AluJo	Multiple, Subtelomeric.
133098-132535	AW499533	100		
140359-140839	AA195187	98	AluSp	4q26 (ac022702)
173552-151404	AW854054/Hs187579 (75)[a]	98	*HSD17B7Ψ*	
178360-177773	AI799915/Hs15248 (164)	99		1q23.3 (AC069037)
179950-180507	AI927669/Hs42392 (39)	99		1q23.3 (AL441926)
277316-277660	AA593504/Hs162587 (1)	97		1q43 (al360271)
294267-294626	AW015485/Hs341696 (1)	97		10p11.2 (al390956)
349457-349181	AW856442	100	HAL1	
351871-365253	AA927631/Hs340030 (1)	99		
352513-352875	AW072278	100	L1PA8	
379608-379900	AI637955/Hs224979 (4)	99	L2	
382137-382486	AA680406/Hs126913 (4)	99		
385502-385916	AW301129/Hs318978 (1)	100		
435627-406440	X69115/Hs54488 (10)	100	*ZNF37A*	
519524-469848	X68687/Hs70617	100	*ZNF11/33A*	
553075-580039	AK056452	100	*ZNF25*	
672343-700930	AJ492196	100	*ZNF248*	
709987-710637	BE387652/Hs57553 (181)[a]	98	*TLK2Ψ*	
700012-700440	HS562273	100		
754102-754795	BE378519	99		
785464-785056	AI248257/Hs149302 (1)	100		
785930-785515	AI991440	100		
771258-770961	AA923150/Hs148281 (1)	100	L1MC4	
1062408-1062674	AA744917	100		
1073813-1074078	BE064203	100		
1075595-1075848	BE063630	99		
1076792-1076582	AI393188	100		
1099274-1099737	AA921809/Hs132449 (3)	99		
1100666-1101084	BE063346	99		
1101637-1101938	AA725605/Hs293102 (2)	100	AluJ	
1111674-1111962	AI902319	100		
1126559-1126868	BE008091	100		
1132040-1132647	AI905942	100	MER5A + 5B	
1133889-1134201	BE069326	100		
1135128-1135707	BE061064	100	MIR	
1138066-1139146	BE072345	100	L2	
1139874-1140174	AI906585	100		
1145135-1145244	BE072230	100	MLT1H1	
1147376-1147687	AA632089	100		
1149924-1150010	AW175720	100	MIR	

ESTs were identified by using BLAST to query the Swissprot, TREMBL, Unigene, and dbEST databases. ESTs were defined as genes if they coincided with either ab initio gene predictions or protein similarity and an intact ORF. ESTs were defined as pseudogenes if they coincided with protein similarity and a disrupted ORF relative to the known protein. Details of gene fragments (similarity to part of known protein, no ESTs) can be found at http://www.ncl.ac.uk/ihg/10p11.htm. No consistent ab initio gene predictions were obtained in the absence of ESTs. The position of each feature within contig 1 (parenthesis) and 2 are shown, together with accession number and unigene cluster information. The % identity of the ESTs to 10p11 are also shown (% match over >80% of EST length). AA927631 (351871-365252) is the only anonymous EST which is spliced.

[a]Only two ESTs within the *TLK2* cluster (BE367652, AV706880) are derived from the 10p11 pseudogene and only four ESTs from the *HSD17B7* cluster (BG199997, BG189163, AI351558, BG182213) are derived from the 10p11 pseudogene.

(Bailey et al. 2001; Katsanis et al. 2001; Christian et al. 2002).

Capicua Pseudogenes Have Undergone Chromosome 7 and Telomere Specific Duplications

In contrast to paralogs of contig 1, most paralogs in the proximal region of contig 2 (which includes the *HSD17β*, *CDC10*, and *Capicua* pseudogenes) map to subtelomeric regions of the genome (e.g., 1q43, 19p13.3, 6p25.3, 4q26). Several pericentromeric and interstitial paralogs also exist, including five loci which have been mapped to chromosome 7 (Fig. 4B). Most of these paralogs terminate at the satellite 3 sequence, or close to the 49 bp repeat at position ~80 kb (Table 3 and Fig. 4B), implicating these repeats in the duplication process. Because of the unusual genomic distribution of these paralogs, the dynamics of duplication was analyzed further by constructing

Table 2. RT-PCR Analysis of 10p11 Genes and ESTs

Position in (contig1), contig2	GenBank acc. no/ unigene cluster	Expression									
		K	L	B	H	T	Li	FL	FB	FH	FK
(30269-30668)	AL045218	−	−	−	−	−	−	−	−	+	−
133098-132535	AW499533	+	+	+	−	+	−	−	−	+	−
351871-365253	AA927631/Hs340030	−	−	−	−	−	−	+	+	−	+
382137-382486	AA680406/Hs126913	−	−	−	−	−	−	−	−	−	−
385502-385916	AW301129/Hs318978	+	−	+	−	+	−	+	+	−	+
435627-406440	*ZNF37A*	+	+	+	+	+	+	+	+	+	+
519524-469848	*ZNF33A*	+	+	+	+	+	+	+	+	+	+
553075-580039	*ZNF25*	+	+	+	+	+	+	+	+	+	+
672343-700930	*ZNF248*	+	+	+	+	+	+	+	+	+	+
700012-700440	HS562273	−	−	−	−	−	−	−	−	−	−
754102-754795	BE378519	+	−	+	−	+	+	−	−	−	−
785464-785056	AI248257/Hs149302	+	−	+	−	+	+	−	−	+	−
785930-785515	AI991440	+	−	+	−	+	+	−	−	−	−
1062408-1062674	AA744917	−	−	−	−	−	−	−	−	+	−
1073813-1074078	BE064203	−	−	−	−	−	−	−	−	−	−
1075595-1075848	BE063630	+	−	−	−	+	−	−	−	−	−
1076792-1076582	AI393188	−	−	−	−	−	−	−	−	−	−
1099274-1099737	AA921809/Hs132449	−	−	−	−	+	−	−	−	−	−
1100666-1101084	BE063346	−	−	−	−	−	−	−	−	−	−

+ expression, − no expression. The cDNAs were derived from the following tissues: K—kidney, L—lung, B—brain, H—heart; T—testis, Li—Liver, FL—fetal lung, FB—fetal brain; FH—fetal heart, FK—fetal kidney. We have confirmed that all transcripts are derived from chromosome 10 by using the appropriate RT-PCR primers to analyse a monochromosomal somatic cell hybrid panel (not shown). Chromosomal origin was not confirmed for ESTs/EST clusters where expression was not analysed by RT-PCR. Primer sequences can be found at http://www.ncl.ac.uk/ihg/10p11.htm.

a maximum likelihood tree from a 3293-bp alignment of 15 paralogs of the *Capicua* pseudogene (Fig. 4C). All five chromosome 7 paralogs form a distinct clade with relatively long terminal branches (0.009–0.016 substitutions/site; Fig. 4C), while five of the subtelomeric loci (11p15.5, 19p13.3, two loci in 6q25.3, and 20q13.33) form a separate clade with much shorter internal branches (0.002–0.003 substitutions/site). This suggests that two distinct duplication processes have had a major influence on the distribution of this pseudogene family—a series of intrachromosomal events on chromosome 7, followed by more recent interchromosomal events between subtelomeric regions. If we assume a neutral subsitution rate of 1.5×10^{-9} to 2.0×10^{-9} per site per year (Miyamoto et al. 1987; Sakoyama et al. 1987), the branch lengths suggest that the most recent subtelomeric interchromosomal duplications have occurred as recently as 1.0–1.3 Myr ago, whereas the chromosome 7 loci have been involved in exclusively intrachromosomal events for at least the last 13–17 Myr.

Although the general patterns of duplication are clear from this analysis, the precise origin of the 10p11 sequence cannot be inferred as low bootstrap values fail to resolve the relationship between the 10p11, Yq11.23, and 4q26 sequences. Despite this, published comparative FISH analyses using phage clones from 10p11 have provided evidence of sequence duplication both to and from 10p11. Two probes approximately 30 kb distal of the *Capicua* pseudogene (~110 kb in contig 2) hybridized to multiple sites in Great Ape chromosomes including regions syntenic to human 10p11, 7cen, and 9ptel, but gave a single hybridization signal in a region syntenic to HSA 7 cen in three Old World monkey species, consistent with sequence dissemination from a pericentromeric chromosome 7 progenitor (probes 746Y1.20 and 746Y1.27; Jackson et al. 1999). The >94.8%-identity shared between the *Capicua* paralogs is consistent with duplication after the ape/Old World monkey divergence (Li et al. 1996). In contrast, a more distal probe, derived from position 210 kb of contig 2, gave two hybridization signals equivalent to 10p11 and 1qtel in human, chimp and gorilla, but only a single 1qtel signal in orangutan and macaque (probe 746Y1.6; Jackson et al. 1999). This suggests that a duplication of material from 1qtel to 10p11 has occurred after the divergence of orangutans from other apes, consistent with the ~98% sequence identity shared between the 10p11 and 1qtel sequences (AL583845 in Fig. 4B). However, it is currently unclear if the physical separation of the paralogs on 1q (1q43.3 and 1q24.2; Fig. 4B) is the result of two independent 1:10 duplications, or of rearrangement within chromosome 1 subsequent to a single 1:10 duplication.

Distal 10p11 Is Enriched for Intrachromosomal Duplications

In contrast to the interchromosomal paralogs of the proximal sequence, the distal 850 kb contains seven out of the eight paralogs which are present in other regions of chromosome 10. (Fig. 2B,4D). Out of ~460 kb of duplicated DNA within this region, only ~35 kb shares high identity to sequence on other chromosomes. Two of the intrachromosomal paralogs have been partially characterized previously (Tunnacliffe et al. 1993; Jackson et al. 1996). The larger of these contains the *ZNF33A* and *ZNF37A* genes, is 228970 bp long, extends from nucleotides 298727–527697 of contig 2, and shares 93.4% identity to 10q11. The smaller extends from nucleotides 808107–895076 and shares 89.9% identity to 10q11. The different sequence identities to 10q11, together with the fact that these duplications are in different orientations relative to their 10q11 counterparts, supports the hypothesis that these have been created by two independent events (Jackson et al.

Table 3. Tandem Repeats >300 bp in 10q11

Position in 10p11 sequence	Shortest high scoring periodicity	No. of repeats	% matches	% indels	Consensus length of other periodicities	Repeat sequ repeat na
(1-25435)	5bp	1001.4	70	0	10bp/26/98bp	GGAAT Sate
(168190-170696)	5bp	487.8	66	21	13bp/18/23bp	CATTT
(207701-285997)	5bp	1000.4	53	15	10bp/15bp etc.	GGAAT Sate
1-46495	5bp	995.8	69	4	10bp/15bp etc.	GGAAT Sate
79169-80644	49bp	30.3	92	1	—	
170593-170958	121bp	3	87	2	—	

Tandem repeats were identified with Tandem Repeat Finder using both stringent and relaxed search parameters (Benson 1999). Because repeat arrays are not homogeneous, a number of details are shown. The position in the sequence (column 1) indicates the extent of the entire array. The No. of repeats, % match, and % indels (columns 3–5) refer to the region defined by the shortest high scoring periodicity and not the entire array. Significant higher order periodicities are also shown, with the highest scoring periodicity being underlined and embolden. The 121bp repeat includes a 36bp Alu fragment. The sequence of the 49bp repeat is: GGCCGGTGTGAGGCAAGGGGCTCACGCTGACCTCTGTCAGCGT GGGAGG.

1999). In addition, a further five smaller intrachromosomal paralogs related to sequence in the distal ~350 kb region have been identified (Fig. 2B). These paralogs map to 10p11.2, approximately 1–2 Mb telomeric of the sequence presented here, and to 10p13–14 (Fig. 4D). Three interchromosomal paralogs were also identified in this region (17q11.2, 5p14.3, and 8p11.2). Interestingly, there is a gradient of sequence identity moving away from the centromeric satellites. The paralogs within the proximal 520 kb share 95.1%–98.3% sequence identity to 10p11, whereas the predominantly chromosome 10 specific paralogs related to the distal 850 kb share 87.9%–93.4% identity to 10p11, indicating that the most recent duplication events have been interchromosomal events involving sequences flanking the satellites. Similar gradients in sequence identity have been observed previously in the pericentromeric regions of 10q (Guy et al. 2000), 22q (Bailey et al. 2002), and at the telomere of Xq (Ciccodicola et al. 2000), suggesting that this may be a genome wide phenomenon.

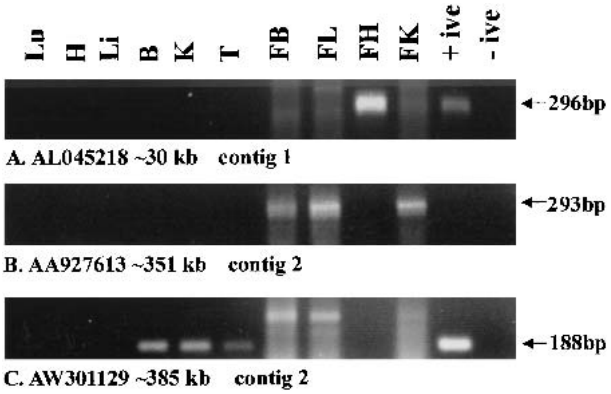

Figure 3 RT-PCR analysis of 10p11 ESTs. The tissue source of each cDNA used is shown above the panels (see Table 2 footnote). The EST accession number is given together with position in the sequence (contig 1 in parenthesis). The expected PCR product size is indicated with an arrow shown to the *right* of each panel. The (+ ive) control lane shown in all experimental panels is genomic DNA. No amplification is observed in the AA927631 genomic DNA control (*B*) as the expected size for this spliced EST is 13,248bp. The sequences amplified in (*A*) and (*C*) are contiguous with genomic DNA. The origin of the larger-than-expected amplification product in Fetal Brain and Fetal Heart (*C*) is unknown. Each experiment was performed in duplicate (not shown).

Intrachromosomal Duplication Has Created One Primate Specific Gene

While there is no evidence of recent gene formation within the interchromosomal duplications, several lines of evidence indicate that both the *ZNF33A* and *ZNF33B* genes, which lie within the ~230 kb 10p11 and 10q11 intrachromosomal duplications respectively, are transcribed and translated. These include the existence of numerous ESTs specific for each gene (Guy et al. 2000), intact open reading frames, and conservation of both the core residues and spacing of the KRAB and zinc finger domains in both predicted proteins (Tunnacliffe et al. 1993). Notably, seven of the sixteen zinc fingers differ by substitutions at positions predicted to be important for base recognition (i.e., positions −1, 2, 3, or 6, relative to the N-terminus of the α helix; data not shown), suggesting that this duplication has given rise to a primate specific gene, and that one or both of these proteins has diverged to recognize a new target sequence.

The ZNFA/B duplication occurred at around the same time as the split between the human and Old World monkey lineages (~20–30 Mya), although there is no evidence of positive selection as the ratio of nonsynonymous to synonymous substitutions between the two complete coding sequences is 0.725 (Hearn 2000). To further investigate the evolution of *ZNF33A* and *ZNF33B*, we generated sequence data from the orthologous gene in species that diverged from the human lineage before the ZNFA/B duplication occurred. PCR primers for the zinc finger-coding region were tested on total genomic DNA of pygmy marmoset, slender loris, pig, and pilot whale. Single products of the expected size were amplified from all species and sequenced. Conceptual translations of the sequences obtained are aligned with the human protein sequences (hZNF33A and hZNF33B) shown in Figure 5. This analysis shows that the gene fragment amplified from each species is more similar to hZNF33B than to hZNF33A. Within the region compared, the human proteins differ at 11 positions, and at ten of these positions all homologues analyzed agree with the sequence of hZNF33B. This indicates that following the ZNFA/B duplication, *ZNF33B* remained under the selective constraints imposed on the ancestral *ZNF33* gene,

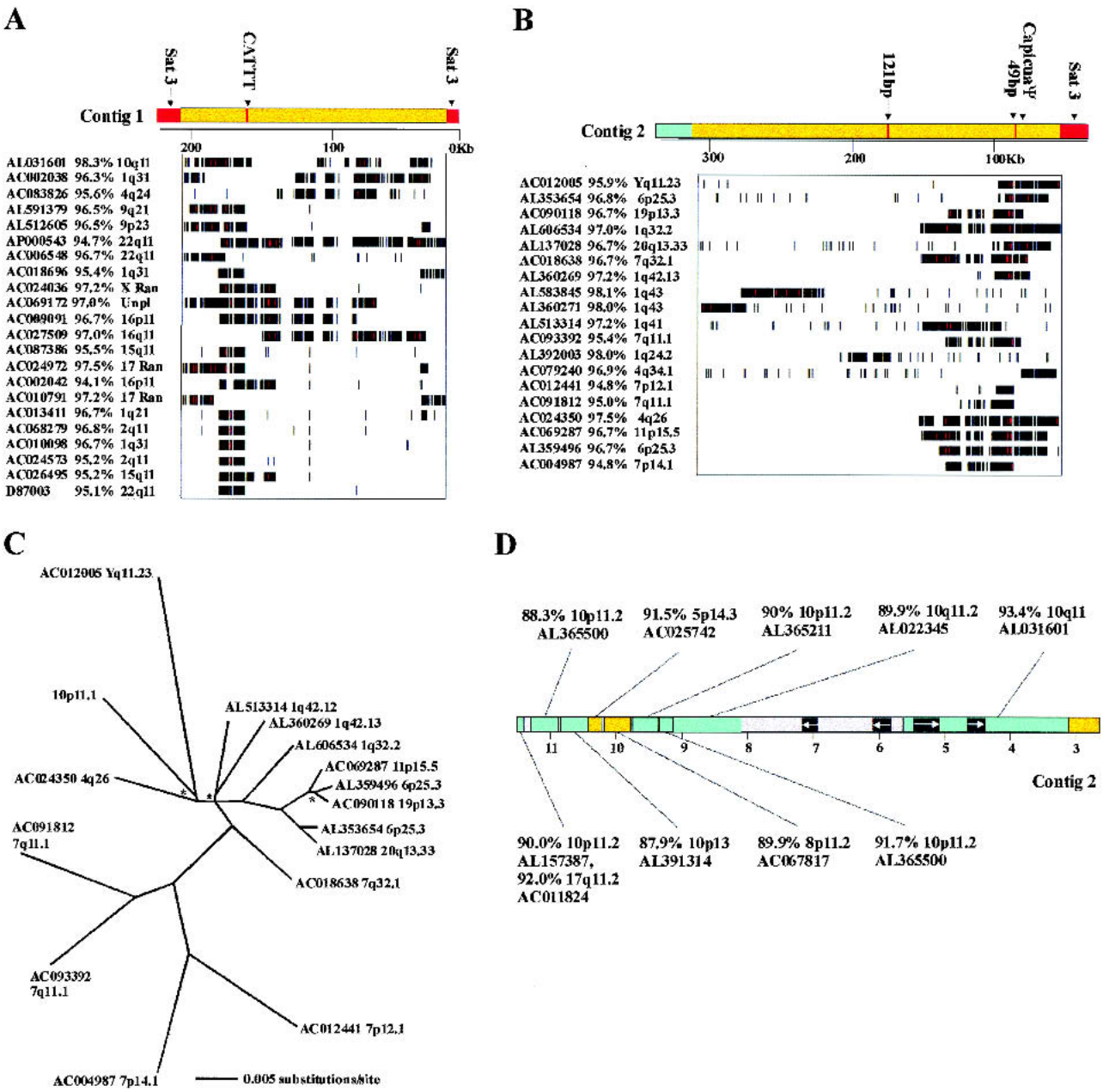

Figure 4 Genomic origin of human paralogs of the 10q11 sequence. Each independent BLAST hit greater than 5 kb in length within the EMBL-NR database is shown. Regions of high identity to the Repeatmasked 10p11 sequence are indicated by black lines. The % identity to the 10p11 sequence, the accession number, and the position within the current human draft (where known) are also indicated. The appropriate portion of the 10p11 summary map (Fig. 2C) is shown for reference, together with the identity of specific repeats (Table 3). (*A*) Paralogs of contig 1. (*B*) Paralogs of proximal ~200 kb of contig 2. (*C*) Maximum likelihood analysis of *Capicua* pseudogenes from Fig. 4B, constructed using a 3293bp alignment spanning nucleotides 74113–77345bp of contig 2. All nodes have >90% bootstrap support with the exception of the three nodes marked with an asterisk. For details of tree construction see Methods. (*D*). Paralogs of distal 900 kb of contig 2.

and that *ZNF33A* has diverged significantly and may now be capable of identifying a distinct target sequence.

DISCUSSION

We have constructed and sequenced a BAC contig which links classical satellites to genes on the short arm of human chromosome 10, and analyzed both transcriptional activity and the distribution of duplications within the sequence. In conjunction with previous work (Guy et al. 2000), the data presented here represents the first structural and transcriptional analysis of both extremities of a human centromeric region, and as such provides an opportunity to investigate higher order structure within these complex and unstable areas of the genome.

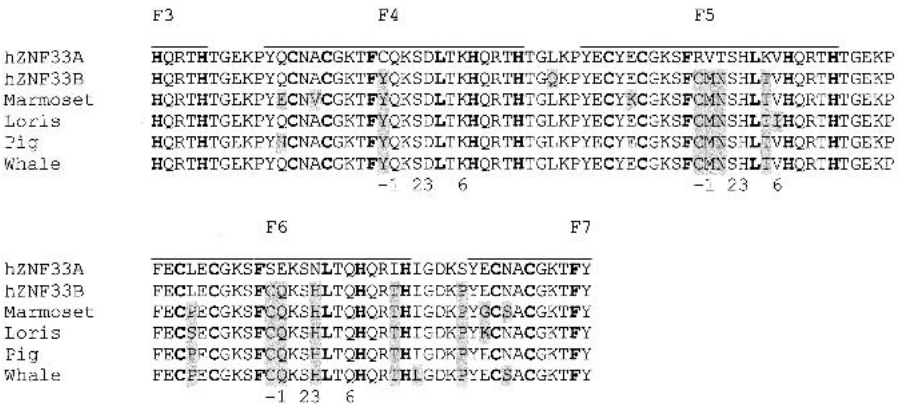

Figure 5 Evolutionary analysis of duplicated ZNF genes. Conceptual translation of *ZNF33* orthologs aligned with *ZNF33A* and *ZNF33B.* Differences to hZNF33A are highlighted in grey. Putative base-contacting residues of each zinc finger (positions −1, 2, 3, and 6 relative to the start of the α helix) are indicated below the alignment. The core residues of each zinc finger are in bold. The *ZNF33* genes were originally defined by two incomplete cDNAs, *ZNF11* and *ZNF33,* before their duplicated status was known (Thiesen 1990). *ZNF33A* and *ZNF33B* have been previously defined as *ZNF11/33A* and *ZNF11/33B* for this reason (Jackson et al 1996; 1999).

10p11 and 10q11 Have Similar Structural Features

Our previous analysis of ~1 Mb of 10q11 sequence linking classical satellites to the *RET* proto-oncogene identified two distinct sequence domains. The proximal, transcriptionally inert, domain consisted of satellites interspersed with interchromosomal pericentromeric duplications while the distal, transcriptionally active, domain contained no satellites and duplications which were principally intrachromosomal (Guy et al. 2000). The data presented here are consistent with this two-domain structure. Satellites are confined to the proximal 600 kb, which contains only ~10% of the 10p11 specific transcripts (3/32), but ~93% of the interchromosomal paralogs (40/43). In contrast, the distal 850 kb of sequence contains four genes, supporting evidence for a further 29 transcripts (albeit of questionable function), and contains seven out of the eight intrachromosomal duplications. These results confirm the association, observed in 10q11, between interchromosomally duplicated satellite-rich DNA and transcriptional inactivity, and are consistent with satellite sequences in 10p11 inducing heterochromatin formation and repressing localized transcription. The fact that transcriptional activity is observed in the distal 10p11 sequences further suggests that, as in 10q11, the transition from multicopy interchromosomally duplicated DNA to chromosome-specific duplications approximates to a heterochromatin/euchromatin boundary.

A striking difference between the 10p11 and 10q11 sequence is the large tract of 10p11 sequence with high identity to telomeres. The combination of sequence and FISH data (Jackson et al. 1999) suggests that this has involved the recent duplication of subtelomeric sequences from 1qter into 10p11. Although noteworthy, the existence of telomeric repeats within pericentromeric DNA was established before genome wide sequencing (Vocero-Akbani et al. 1996) and there are several well-characterized examples of sequence movement between pericentromeric and subtelomeric DNA (e.g., van Geel et al. 2002; Martin et al. 2002). These homologies may be of no consequence. However, the suggestion that pericentromeric-directed duplication could be involved in the repair of double strand DNA breaks (IHGSC 2001) raises the possibility that they may represent the products of the repair of broken chromosome ends. Alternatively, the duplication of DNA from recombinationally inert centromeres to recombinationally highly active telomeres could represent a novel way of removing functional genes from the cumulative effect of ratchet-like mutational processes (Muller 1964) which have been implicated in the degeneration of recombinationally inactive sequences on the Y chromosome (Charlesworth 1991), and which could place a heavy mutational burden on genes within pericentromeric sequences which are also recombinationally suppressed. Due to the incomplete nature of draft data in the telomeric and pericentromeric regions of most human chromosomes, however, any systematic analysis of this phenomenon will have to await the publication of finished human sequence.

Interchromosomal Comparison Implies Simple Heterochromatin/Euchromatin Boundaries

Although chromosome 10 is the first human chromosome where transcription at the transition between satellites and genes has been analyzed on both arms, a transcriptional analysis of the Cat Eye Syndrome critical region (CECR), distal of pericentromeric satellites on 22q11, has also been performed (Footz et al. 2001). The most proximal 400 kb of this sequence was also found to consist of transcript poor, pericentromeric-duplicated DNA. Neither of the two putative genes identified in this region (CECR7 and CECR8) contain large ORFs, while many CECR7 ESTs were found to be derived from paralogous loci on other chromosomes (Footz et al. 2001). In contrast to these results, a total of 41 genes have been annotated proximal of pericentromeric satellites within the finished sequence from chromosomes 20, 21, and 22 (Dunham et al. 1999; Hattori et al. 2000; Deloukas et al. 2001). This appears inconsistent with the satellite-induced transcriptional inactivity proposed here. However, these datasets can be reconciled with our own when it is realized that 35/41 of the annotations on these chromosomes are based solely on similarity to known proteins, or intron/exon structures in the absence of spliced EST support. The inclusion of annotations with no spliced EST support is justified for whole chromosome analyses as the identification of all possible genes is a clearly stated priority. However, we have deliberately discounted these annotations because of the high likelihood that in pericentromeric regions they represent unprocessed pseudogenes and gene fragments which have been duplicated from other loci.

To allow a direct comparison between datasets, we have removed annotations with no spliced EST support from the most centromeric 1 Mb of chromosomes 20, 21, and 22 and present the remaining annotations in Fig. 6. Using these stringent criteria, only 6 gene structures proximal of satellite sequences on chromosomes 20, 21, and 22 remain. Two of these, members of the *TPTE* (Chen et al. 1999) and *BAGE*

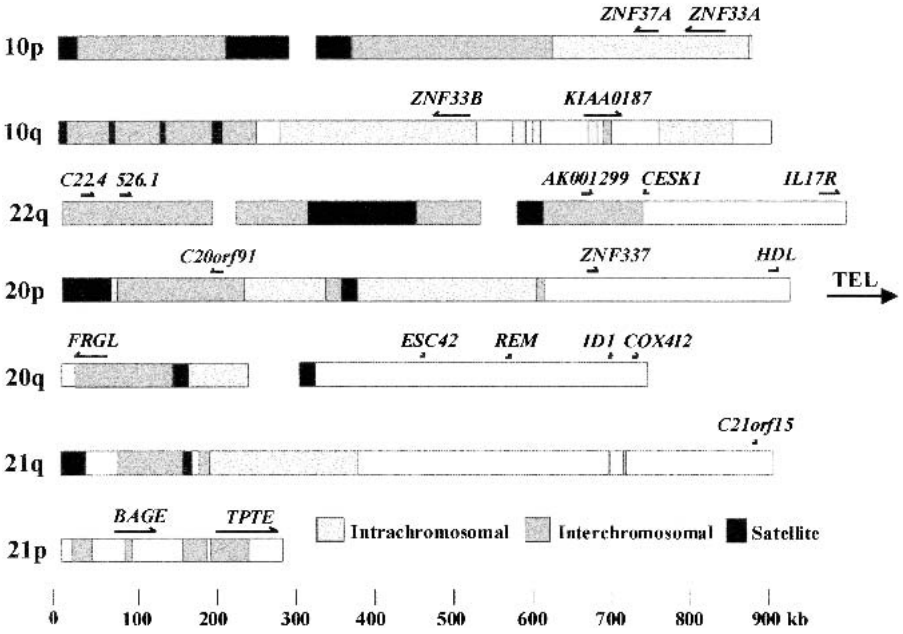

Figure 6 Comparison of 10p11 with other human pericentromeric sequences. Annotations were taken from Deloukas et al. 2001; Hattori et al. 2000; Dunham et al. 1999; Guy et al. 2000; and Bailey et al. 2002. Additional annotations from the working draft April 2001 freeze (http://genome.ucsc.edu/) and ongoing chromosome specific programs (http://hgp.gsc.riken.go.jp/data_tools/data_chr21_seq_annotation.html, http://www.sanger.ac.uk/cgi-bin/humace/SuperMap22/) were also analyzed. Only gene annotations supported by both spliced ESTs and ab initio gene predictions are shown. These were analyzed further by aligning the ESTs which define the gene to all closely related genomic sequences within the human draft. On the basis of this, *C20orf191* and *MLLT10L*, which map proximal of satellite sequences on chromosome 20 (Deloukas et al. 2001), were discounted as the ESTs which define these genes only share ~95% identity to the chromosome 20 genomic sequence. The positions of major interchromosomal and intrachromosomal duplications are shown in dark and light grey respectively. The intrachromosomal duplications on 20p and 20q have also been duplicated interchromosomally. 526.1–*ap000526.1,* c22.4–*ap000523.c22.4.*

(Boel et al. 1995) gene families, are present on the short arm of chromosome 21 and are specifically expressed in testis and testis/neoplastic tissues respectively. The precise position of these relative to centromeric satellites is unclear. None of the other four annotated genes proximal of satellites, *ap000526.1* and *ap000523.c22.4* on 22q11, and *FRGL* and *c20orf91* on 20q11, are known genes. *Ap000526.1* is a paralogous fragment of the *KIAA0187* gene defined by a single spliced EST from testis, AA725634 (Dunham et al. 1999; Crosier et al. 2002). *Ap000523.c22.4* is defined as a gene on the basis of a single 2792bp cDNA also derived from testis which has nine mismatches with the chromosome 22 sequence. Although it has numerous open reading frames, the largest is only 59 amino acids long. *FRGL* is closely related to the *FRG* gene on 4q26 and the current ENSEMBL in silico annotation (ENSG00000149531) is a 9 exon structure with the potential to encode a 179 amino acid protein truncated at both the 5′ and 3′ end relative to *FRG*. However, all ESTs which are derived from the 20q locus (identified by sequence alignment, see legend) have a frameshift mutation not predicted by the in silico annotation due to the fact that exon 4 of the pututative transcript is not present within the ESTs. Furthermore, these ESTs fall within cluster 1 of an analysis of *FRG*-related transcripts by Grewal et al. (1999), which concluded that all FRG-related transcripts are derived from pseudogenes. Finally, the only spliced EST which defines *C20orf91* (AW135038) contains a retroviral LTR and differs from the chromosome 20 sequence at six nucleotides out of 460. Thus, with the exception of *TPTE* and *BAGE* on the short arm of chromosome 21, the status of all of the gene structures proximal of pericentromeric satellites remains questionable. It is also noteworthy that a significant number of annotations close to satellites are based on expression observed in testis and/or neoplastic tissue, both of which are known to have unusual expression patterns relative to other tissues (Alizadeh et al.; Kleene 2001 2001). This has significant implications both for the potential these sequences have for gene formation, and for the closure of human sequence maps.

New Genes Are Unlikely To Be Created Proximal of Pericentromeric Satellites

Sequence exchanges within and between both subtelomeric and pericentromeric DNA have been proposed as a major source of biological novelty during the evolution of complex organisms (Trask et al. 1998; Eichler 2001). The importance of tandem and intrachromosomal duplication in the expansion and diversification of eukaryotic gene families is clear (e.g., Gu et al. 2002). A startling example of the speed of such expansion is the Morpheus gene family, which has expanded and been subject to strong positive selection within the last 25 Myr (Johnson et al. 2001). In our analysis, we have established both that a novel gene has been formed (either *ZNF33A* or *ZNF33B*) and that only one of these, *ZNF33A*, has diverged significantly since duplication. This suggests that *ZNF33B* has continued performing the ancestral function, while *ZNF33A* has evolved to identify a different target sequence. There is also a well-characterized precedent for gene creation at telomeres within other eukaryotes (van der Ploeg et al. 1992), and both the existence of highly expressed genes which have been duplicated between telomeres (Flint et al. 1997; Ciccodicola et al. 2000) and the relocation of active genes from one telomere to another (vanGeel et al. 2002) have been reported in primates.

While the role of tandem and intrachromosomal duplication in human gene evolution is overwhelming, and there is growing evidence for the role of telomeric instability, the role of pericentromeric interchromosomal duplication remains more speculative (Eichler et al. 1997; 1999; Eichler 2001). Our conclusion that sequences flanking pericentromeric satellites appear to be transcriptionally inert strongly suggests that the vast majority of pericentromerically directed duplication events will result in the creation of heterochromatic sequences. If the juxtaposition of different duplications has contributed to biological novelty it seems likely that this process would be confined to the numerous segmental dupli-

cations/rearrangements which occur outside pericentromeric domains (Bailey et al. 2001). In this respect, it is interesting that only one of the 11 transcripts created or modified as a result of recent duplications on chromosome 22 maps within pericentromeric duplications (Bailey et al. 2002). This putative gene (AK001299; Fig. 6) is not within the most recent duplications as it maps ~200 kb distal of pericentromeric satellites. It consists of 3 exons totalling 1627bp, but has a predicted ORF of only 98 amino acids. Furthermore, it is defined by only two ESTs (AK001299 and AU126316), both of which were derived from a teratocarcinoma cell line after 48 h induction with retinoic acid, raising the possibility that it is not expressed in normal tissue.

The results presented here, together with a stringent reassessment of annotations within finished pericentromeric sequence (Fig. 6) suggest that for pericentromeric-directed duplication to create new genes, an advantageous gene structure would have to be both created within a pericentromeric domain and rapidly relocated to a more open chromatin environment before mutation disrupted the open reading frame. Although extremely unlikely, there are several mechanisms which could result in precisely this course of events. First, the existence of a gradient in sequence identities close to centromeric satellites (Guy et al. 2000; Bailey et al. 2002; this study) suggests that the most recent pericentromeric-directed duplication events can displace previously duplicated sequences towards the telomere, away from the repressive effects of satellites. The high density of interchromosomal segmental duplications close to centromeres on numerous chromosomes could then provide the medium for rapid dissemination of any advantageous transcript. Second, chromosomal rearrangements during primate evolution which involve the loss of a centromere, such as chromosome fusion (Fan et al. 2002), provide an avenue through which chromatin at an inactive centromere could be rapidly remodeled, releasing novel combinations of exons from a transcriptionally repressive environment. Finally, the recent evidence that primate centromeres frequently move position within a chromosome with no signs of chromosomal rearrangement (Ventura et al. 2001) suggests that the large heterochromatic domains associated with these structures may be subject to chromatin remodeling with much higher frequency than previously thought. As a result, the indirect contribution of pericentromeric-directed duplication to gene creation remains very plausible on an evolutionary timescale.

Pericentromeric Satellites Are Logical Termini for Whole Chromosome Sequence Maps

All human sequence maps which have been presented as finished contain gaps due to the under-representation of some sequences within libraries, and the difficulty of accurately mapping highly repetitive sequences. Most of these gaps lie within 2 Mb of either a centromere or telomere (Dunham et al. 1999; Hattori et al. 2000; Deloukas et al. 2001) due to the high local density of repeats and duplications. Closure of maps in these regions can, therefore, involve significant effort. While the telomere represents a clear endpoint for sequencing efforts, the logical terminus of sequencing efforts at centromeres is not so obvious. The work presented here and elsewhere (e.g., Schueler et al. 2001) has established that contiguous sequence can be generated from the most repetitive centromeric satellites, but that it requires significant investment of resources as it is not open to the extensive mapping technologies currently employed by sequencing centers (Waterson et al. 2002).

With the attention of human genome sequencing now focusing on the closure of sequence maps it is appropriate to ask if contiguous sequence from satellite-rich pericentromeric regions is required or desirable for all chromosomes. Numerous heterochromatic loci have been defined in both *Drosophila* and *Arabidopsis* through genomic sequencing (Adams et al. 2000; The Arabidopsis Genome Initiative 2000), raising the possibility that similar loci exist within the human genome. However, the existence of *Drosophila* heterochromatic loci had already been established through genetic mapping (Gatti and Pimpinelli 1992), while the density of genes within the centromeric regions of *Arabidopsis* (~1 gene every 100 kb), and their presence within recognizable islands of unique sequence, made their identification both straightforward and cost effective. In contrast, all sequence data to date, including the work presented here, suggests that human centromeric regions are devoid of unique sequences (Hattori et al. 2000; Deloukas et al. 2001; Bailey et al. 2002), while estimates of the physical size of centromeric satellite arrays indicates that collectively they could span over 120 Mb of DNA (Lee et al. 1997). A search for human heterochromatic genes through sequencing would, therefore, be a complex and time consuming undertaking. Furthermore, it may not be necessary, since we would expect these genes to be represented within EST and cDNA databases where they could be easily identified for subsequent targeted analysis by comparison with the genomic sequence. From a gene identification point of view, therefore, blocks of pericentromeric satellite appear to represent logical minimal end points for genomic sequence generation due to diminishing return on investment.

However, while cataloguing all human genes is of central importance, it is not the sole rationale for genomic sequencing (IHGSC 2001; Venter et al. 2001). The novel duplication and homology-dependent mutational mechanisms enriched within pericentromeric regions of the genome have proved to be one of the major surprises of the human genome sequence (IHGSC 2001). If we are to fully understand the tempo and mode of these events during primate and eukaryotic evolution, it is clear that high quality reference sequences of human and of other organisms will be required. The generation of high quality data from pericentromeric regions of the human genome (Eichler 2001) and the closure of other "finished" genomes (Mardis et al. 2002) therefore remain highly desirable goals.

METHODS

STS Development, Contig Construction and BAC Sequencing

STSs used in BAC contig construction were developed from plasmid subclones derived from YACs (Jackson et al. 1999) and from the ends of BACs. YAC subclones were PCR-amplified using primers which flank the multiple cloning site and purified using Quiaquick PCR purification kits (QIAGEN) according to manufacturer's instructions except that DNAs were eluted in water. Approximately 100 ng of template was used for each sequencing reaction. All sequencing reactions were performed using an ABI PRISM BigDye cycle sequencing kit according to manufacturer's instructions (PE Applied Biosystems) and were analyzed using an ABI377 (PE Applied Biosystems). Primers were designed using PrimerSelect software (DNAstar). Sequence alignments were performed using the Megalign software package (DNAstar). High density filters of

the RCPI11 human BAC library (http://www.chori.org/bacpac/) were screened with eight probes (367M1.4, ZNF32SPC, 746l3/1, WI-16960, *ZNF37A,* 746B/B4A, 746Y1.6, 746Y1.32, and 746Y1.27 in Jackson et al. 1999 and http://www.ncl.ac.uk/ihg/10p11.htm) according to recommended protocols. A further marker, *AFM295SC3*, which contains interspersed repeats, was used to screen commercially available PCR pools (Research Genetics) according to manufacturer's recommendations. Clone overlaps were then confirmed by screening with additional markers from the region and by FISH (Jackson et al. 1999). A second contig was constructed around a chromosome 10 *ALD* paralog (453N3, Horvath et al. 2000a). Because pericentromeric regions are highly repetitive and 10p11 contains material related to 10q11 sequence, markers were routinely tested against somatic cell hybrids which contain chromosome 10p or 10q as their only human component (Tunnacliffe et al. 1994). Duplicated or repetitive markers were analyzed by a combination of restriction site analysis (Fig. 1) and direct sequencing of PCR products (Accession nos.: AJ275023–AJ275036). The tiling path was chosen following DNA fingerprinting and sequenced to greater than 99.99% accuracy as described previously (Dunham et al. 1999; IHGSC 2001).

Sequence Analysis

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 kb and 100 kb sections and analyzed using NIX (Williams et al. 1998). Interspersed and tandem repeats were identified using Repeatmasker (http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl) and Tandem Repeat Finder (Benson 1999). Similarities to existing genomic, EST, and protein sequences were identified by using repeat-masked 10p11 sequence to query Swissprot, TREMBL and EMBL-NR, and EMBL-HTG databases with the BLAST family of programs (Altschul et al. 1990; 1997). Overlap between clones was also analyzed using Gdot (an in-house dot matrix program). The GC content and distribution of interspersed repeats was established by using RepeatMasker to analyze overlapping 20 kb sequence files with a 16-kb overlap, which were generated using in-house software. Paralogous sequences within the sequence identified by BLAST were viewed using both Parasight (Bailey, unpubl.) and NIX. Only paralogs >5 kb in length were analyzed further. To remove redundancies, entries with paralogs structurally related to 10p11 (defined by Parasight and NIX) were compared to each other using Gdot and Align (Pearson and Lipman 1988). Paralogs which could not be distinguished from each other on the basis of sequence identity (taken as >99.0% identity), or on the basis of structural differences (e.g., different linked sequence), were assumed to be from the same locus and excluded from further analyses. The remaining independent paralogs were then individually compared to the 10q11 sequence using Gdot and Align. Sequence divergence was estimated using Alignscorer (Horvath et al. 2000a). This analysis may underestimate the true number of paralogs present within draft sequence.

RT-PCR Analysis

A panel of 8 cDNAs derived from adult tissues (Clontech) were analyzed according to manufacturer's recommendations. Primer information can be found at http://www.ncl.ac.uk/ihg/10p11.htm.

Phylogenetic Analysis

PAUP version 4.0b10 (Sinauer Associates) was used to construct a maximum-likelihood tree using an exhaustive search method under an HKY85 model of molecular evolution (Hasegawa et al. 1985). Estimates of the γ distribution of among-site rate variation and the proportion of invariant sites were then obtained and one round of Tree Bisection and Reconnection was performed. A neighbor-joining bootstrap of 1000 replicates was then performed using the maximum likelihood setting obtained by the above procedure. Insertions and deletions were considered missing data and excluded. Maximum parsimony analyses were also performed, and produced an almost identical topology (not shown).

Comparative Sequencing

To identify *ZNF33* orthologs, cross-species PCRs were performed using the zinc finger-coding region primers D31ZNFL (CCATAAGTCAGCCCTCACATTA) and D31ZNFR (ATGCCTGGTGAGTACTGACTTG). Pygmy marmoset and slender loris genomic DNA was extracted from tissue obtained from the Institute of Zoology, London. Pig and pilot whale total genomic DNAs were provided by Dr. W. Amos, Department of Zoology, University of Cambridge.

ACKNOWLEDGMENTS

This work was supported by grants from the Wellcome Trust (Grants 049859) and E.C (Contract BMH4-CT97–2433) to MSJ, and by a short term fellowship from EMBO (L.V) and studentships from the UK MRC (T.H) and BBSCRC (J.M). Primate and mammalian genomic DNA samples were obtained from the Institute of Zoology, London and from Dr. W. Amos, Department of Zoology, University of Cambridge.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M., Celniker, S.E., Holt, R.A., Evans, C.E., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Agata, Y., Matsuda, E., and Shimizu, A. 1999. Two novel Kruppel-associated box-containing zinc-finger proteins, KRAZ1 and KRAZ2, repress transcription through functional interaction with the corepressor KAP-1 (TIF1β/KRIP-1). *J. Biol. Chem.* **274:** 16412–16422.

Alizadeh, A.A., Ross, D.T., Perou, C.M., and van de Rijn, M. 2001. Towards a novel classification of human malignancies based on gene expression patterns. *J. Pathol.* **195:** 41–52.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Andrulis, E.D., Neiman, A.M., Zappulla, D.C., and Sternglanz, R. 1998. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature* **394:** 592–595.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11:** 1005–1017.

Bailey, J.A., Yavor, A.M., Viggiano, L., Misceo, D., Horvath, J.E., Archidiacono, N., Schwartz, S., Rocchi, M., and Eichler, E.E. 2002. Human-specific duplication and mosaic transcripts: The recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70:** 83–100.

Benson, G. 1999. Tandem repeat finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.
Boel, P., Wildmann, C., Sensi, M.L., Brasseur, R., Renauld, J.C., Coulie, P., Boon, T., and van der Bruggen, P. 1995. BAGE: A new gene encoding an antigen recognized on human melanomas by cytolytic T lymphocytes. *Immunity* **2:** 167–175.
Charlesworth, B. 1991. The evolution of sex chromosomes. *Science* **251:** 1030–1033.
Chen, H., Rossier, C., Morris, M.A., Scott, H.S., Gos, A., Bairoch, A., and Antonarakis, S.E. 1999. A testis-specific gene, TPTE, encodes a putative transmembrane tyrosine phosphatase and maps to the pericentromeric region of human chromosomes 21 and 13, and to chromosomes 15, 22, and Y. *Hum. Genet.* **105:** 399–409.
Christian, S.L., McDonough, J., Liu, C-Y., Shaikh, S., Vlamakis, V., Badner, J.A., Chakravarti, A., and Gershon, E.S. 2002. An evaluation of the assembly of an approximately 15-Mb region of human chromosome 13q32-q33 linked to bipolar disorder and schizophrenia. *Genomics* **79:** 635–643.
Ciccodicola, A., D'Esposito, M., Esposito, T., Gianfrancesco, F., Migliaccio, C., Miano, M.G., Matarazzo, M.R., Vacca, M., Franze, A., Cuccurese, M., et al. 2000. Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum. Mol. Genet.* **12:** 395–401.
Cooke, H.J. and Hindley, J. 1979. Cloning of human satellite III DNA: Different components are on different chromosomes. *Nucleic Acids Res.* **6:** 3177–3197.
Crosier M., Viggiano, L., Guy, J., Misceo, D., Stones, R., Wei, W., Hearn, T., Ventura, M., Archidiacono, N., Rocchi, M., et al. 2002. Human paralogs of KIAA0187 were created through independent pericentromeric-directed and chromosome-specific duplication mechanisms. *Genome Res.* **12:** 67–80.
Deloukas, P., Matthews, L.H., Ashurst, J., Burton, J., Gilbert, J.G., Jones, M., Stavrides, G., Almeida, J.P., Babbage, A.K., Bagguley, C.L. et al. 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865–871.
Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J. et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.
Eichler, E.E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17:** 661–669.
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.
Eichler, E.E., Archidiacono, N., and Rocchi, M. 1999. CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* **9:** 1048–1058.
Fan, Y., Newman, T., Linardopoulou, E., and Trask, B.J. 2002. Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13–2q14.1 and paralogous regions. *Genome Res.* **12:** 1663–1672.
Flint, J., Thomas, K., Micklem, G., Raynham, H., Clark, K., Doggett, N.A., King, A., and Higgs, D.R. 1997. The relationship between chromosome structure and function at a human telomeric region. *Nature Genet.* **15:** 252–257.
Footz, T.K., Brinkman-Mills, P., Banting, G.S., Maier, S.A., Riazi, M.A., Bridgland, L., Hu, S., Birren, B., Minoshima, S., Shimizu, N., et al. 2001. Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: A search for candidate genes at or near the human chromosome 22 pericentromere. *Genome Res.* **11:** 1053–1070.
Francastel, C., Walters, M.C., Groudine, M., and Martin, D.I.K. 1999. A functional enhancer suppresses silencing of a transgene and prevents its localization close to centromeric heterochromatin. *Cell* **99:** 259–269.
Garrick, D., Fiering, S., Martin, D.I.K., and Whitelaw, E. 1998. Repeat-induced gene silencing in mammals. *Nat. Genet.* **18:** 56–59.
Gatti, M. and Pimpinelli, S. 1992. Functional elements in *Drosophila melanogater* heterochromatin. *Annu. Rev. Genet.* **26:** 239–275.
Grewal, P.K., van Geel, M., Frants, R.R., de Jong, P., and Hewitt, J.E. 1999. Recent amplification of the human FRG1 gene during primate evolution. *Gene* **227:** 79–88.
Gu, X., Wang, Y., and Gu, J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet.* **31:** 205–209.
Guy, J., Spalluto, C., McMurray, A., Hearn, T., Crosier, M., Viggiano, L., Miolla, V., Archidiacono, N., Rocchi, M., Scott, C., et al. 2000. Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Hum. Mol. Genet.* **9:** 2029–2042.
Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22:** 160–174.
Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K. et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311–319.
Hearn, T. 2000. "Organization, expression and evolution of Kruppel-type zinc finger genes in human chromosomal region 10p11.2-q11.2." Ph.D. thesis, University of Newcastle Upon Tyne, UK.
Henikoff, S. 1990. Position effect variegation after 60 years. *Trends Genet.* **6:** 422–426.
Horvath, J.E., Schwartz, S., and Eichler, E.E. 2000a. The mosaic structure of human pericentromeric DNA: A strategy for characterizing complex regions of the human genome. *Genome Res.* **10:** 839–852.
Horvath, J.E., Viggiano, L., Loftus, B.J., Adams, M.D., Archidiacono, N., Rocchi, M., and Eichler, E.E. 2000b. Molecular structure and evolution of an α satellite non-α satellite junction at 16p11. *Hum. Mol. Genet.* **9:** 113–123.
International Human Genome Sequencing Consortium (IHGSC). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Jackson, M.S., Slijepcevic, P., and Ponder, B.A.J. 1993. The organization of repetitive sequences in the pericentromeric region of human chromosome 10. *Nucleic Acids Res.* **21:** 5865–5874.
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.
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.
Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E.E. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413:** 514–519.
Katsanis, N., Worley, K.C., and Lupski, J. 2001. An evaluation of the draft human sequence. *Nat. Genet.* **29:** 88–91.
Keller, M.P., Seifried, B.A., and Chance, P.F. 1999. Molecular evolution of the CMT1A-REP region: A human- and chimpanzee-specific repeat. *Mol. Biol. Evol.* **16:** 1019–1026.
Kleene, K.C. 2001. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mech. Dev.* **106:** 3–23.
Lee, C., Wevrick, R., Fisher, R.B., Ferguson-Smith, M.A., and Lin, C.C. 1997. Human centromeric DNAs. *Hum. Genet.* **100:** 291–304.
Li, W.H., Ellsworth, D.L., Krushkal, J., Chang, B.H., and Hewett-Emmett, D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5:** 182–187.
Luijten, M., Wang, Y., Smith, B.T., Westerveld, A., Smink, L.J., Dunham, I., Roe, B.A., and Hulsebos, T.J. 2000. Mechanism of spreading of the highly related neurofibromatosis type 1 (NF1) pseudogenes on chromosomes 2, 14 and 22. *Eur. J. Hum. Genet.* **8:** 209–214.
Mardis, E., McPherson, J., Martienssen, R., Wilson, R.K., and McCombie, W.R. 2002. What is finished, and why does it matter. *Genome Res.* **12:** 669–671.
Martin, C.L., Wong, A., Gross, A., Chung, J., Fantes, J.A., and Ledbetter, D.H. 2002. The evolutionary origin of human subtelomeric homologies–or where the ends begin. *Am. J. Hum. Genet.*. **70:** 972–984.
Milot, E., Strouboulis, J., Trimborn, T., Wijgerde, M., deBoer, E., Langerveld, A., Tan-Un, K., Vergeer, W., Yannoutsos, N., Grosveld, F., et al. 1996. Heterochromatin effects on the frequency and duration of LCR-mediated gene transcription. *Cell* **87:** 105–114.
Miyamoto, M.M., Slightom, J.L., and Goodman, M. 1987. Phylogenetic relations of humans and African apes from DNA sequences in the ψ η-globin region. *Science* **238:** 369–373.

Muller, H.J. 1964. The relation of recombination to mutational advance. *Mutat. Res.* **1:** 2–9.
Park, S.S., Stankiewicz, P., Bi, W., Shaw, C., Lehoczky, J., Dewar, K., Birren, B., and Lupski, J.R. 2002. Structure and evolution of the Smith-Magenis Syndrome repeat gene clusters, SMS-REPs. *Genome Res.* **12:** 729–738.
Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.
Regnier, V., Meddeb, M., Lecointre, G., Richard, F., Duverger, A., VanCong, N., Dutrillaux, B., Berheim, A., and Danglot, G. 1997. Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggests a process of pericentromeric interchromosomal transposition. *Hum. Mol. Genet.* **6:** 9–16.
Reithman, H.C., Xiang, Z., Paul, S., Morse, E., Hu, X-L., Flint, J., Chi, H-C., Grady, D.L., and Moyzis, R.K. 2001. Integration of telomere sequences with the draft human genome sequence. *Nature* **409:** 948–951.
Sakoyama, Y., Hong, K.J., Byun, S.M., Hisajima, H., Ueda, S., Yaoita, Y., Hayashida, H., Miyata, T., and Honjo, T. 1987. Nucleotide sequences of immunoglobulin ϵ genes of chimpanzee and orangutan: DNA molecular clock and hominoid evolution. *Proc. Natl. Acad. Sci.* **84:** 1080–1084.
Schueler, M.G., Higgins, A.W., Rudd, M.K., Gustashaw, K., and Willard, H.F. 2001. Genomic and genetic definition of a functional human centromere. *Science* **294:** 109–115.
Stankiewicz, P. and Lupski, J.R. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18:** 74–82.
Thiesen, H.J. 1990. Multiple genes encoding zinc finger domains are expressed in human T cells. *New Biol.,* **2:** 363–374.
Trask, B.J., Massa, H., Brand-Arpon, V., Chan, K., Friedman, C., Nguyen, O.T., Eichler, E., van den Engh, G., Rouqier, S., Shizuya, H. et al. 1998. Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.* **13:** 2007–2020.
Tunnacliffe, A., Liu, L., Moore, J.K., Leversha, M.A., Jackson, M.S., Papi, L., Ferguson-Smith, M.A., Thiesen, H-J., and Ponder, B.A.J. 1993. Duplicated KOX zinc finger gene clusters flank the centromere of chromosome 10: Evidence for a pericentric inversion during primate evolution. *Nucleic Acids Res.* **21:** 1409–1417.
Tunnacliffe, A., Jackson, M.S., Gardner, E., Love, D.R., Moore, J.K., Mole, S.E., Mulligan, L.M., Graham, A., Finocchiaro, G., Orstavik, S., et al. 1994. A multiple interval physical map of the pericentromeric region of human chromosome 10. *Hum. Genet.* **93:** 313–318.
van der Ploeg, L.H., Gottesdiener, K., and Lee, M.G. 1992. Antigenic variation in African trypanosomes. *Trends Genet.* **8:** 452–457.
van Geel, M., Eichler, E.E., Beck, A.F., Shan, Z., Haaf, T., van der Maarel, S.M., Frants, R.R., and de Jong, P.J. 2002. A cascade of complex subtelomeric duplications during the evolution of the hominoid and Old World monkey genomes. *Am. J. Hum. Genet.* **70:** 269–278.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W, Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Ventura, M., Archidiacono, N., and Rocchi, M. 2001. Centromere emergence in evolution. *Genome Res.* **11:** 595–599.
Vocero-Akbani, A., Helms, C., Wang, J.C., Sanjurjo, F.J., Korte-Sarfaty, J., Veile, R.A., Liu, L., Jauch, A., Burgess, A.K., Hing, A.V. et al. 1996. Mapping human telomere regions with YAC and P1 clones: Chromosome-specific markers for 27 telomeres including 149 STSs and 24 polymorphisms for 14 proterminal regions. *Genomics* **36:** 492–506.
Waterson, R.H., Lander, E.S., and Sulston, J. 2002. On the sequencing of the human genome. *Proc. Natl. Acad. Sci.* **99:** 3712–3716.
Williams, G.W., Woollard, P.M., and Hingamp, P. 1998. NIX: A Nucleotide Identification system at the HGMP-RC. http://www.hgmp.mrc.ac.uk/NIX/

WEB SITE REFERENCES

http://www.ncl.ac.uk/ihg/10p11.htm; Supplementary information from Institute of Human Genetics, University of Newcastle.
http://repeatmasker.genome.washington.edu; RepeatMasker Web site.
http://www.chori.ori/bacpac/; Children's Hospital Oakland Research Institute, BACPAC Resource Centre.
http://www.sanger.ac.uk/HGP/Humana/human_analysis.shtml; The Wellcome Trust Sanger Institute, Human Sequence Analysis Group.
http://www.sanger.ac.uk/HGP/Humana/ACE.shtml; The Wellcome Trust Sanger Institute, Humace.
http://genome.ucsc.edu/; University of California at Santa Cruz, Bioinformatics site.
http://hgp.gsc.riken.go.jp/data_tools/data_chr21_seq_annotation.html; Riken Human Genome Research Group, Chromosome 21 annotation.
http://www.sanger.ac.uk/cgi-bin/humace/SuperMap22/; The Wellcome Trust Sanger Institute, Long range analysis of human chromosome 22.

Received July 19, 2002; accepted in revised form November 4, 2002.