The structure and evolution of centromeric transition regions within the human genome

Xinwei She[1,2*], Julie E. Horvath[1*], Zhaoshi Jiang[1,2], Ge Liu[1], Terrence S. Furey[3], Laurie Christ[1], Royden Clark[1], Tina Graves[4], Cassy L. Gulden[1], Can Alkan[1], Jeff A. Bailey[1], Cenk Sahinalp[1,5], Mariano Rocchi[6], David Haussler[3], Richard K. Wilson[4], Webb Miller[7], Stuart Schwartz[1] & Evan E. Eichler[1,2]

[1] *Department of Genetics, Center for Computational Genomics and the Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA*
[2] *Department of Genome Sciences, University of Washington School of Medicine, 1705 NE Pacific St, Seattle, Washington 98195, USA*
[3] *UCSC Genome Bioinformatics Group, Center for Biomolecular Science & Engineering, University of California, Santa Cruz, 1156 High St, Santa Cruz, California 95064, USA*
[4] *Washington University School of Medicine, Genome Sequencing Center, 4444 Forest Park Boulevard, St Louis, Missouri 63108, USA*
[5] *School of Computing Science, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada*
[6] *Sezione di Genetica, DAPEG, University of Bari, Via Amendola 165/A 70126 Bari, Italy*
[7] *Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, Pennsylvania 16802, USA*

* These authors contributed equally to this work

An understanding of how centromeric transition regions are organized is a critical aspect of chromosome structure and function; however, the sequence context of these regions has been difficult to resolve on the basis of the draft genome sequence. We present a detailed analysis of the structure and assembly of all human pericentromeric regions (5 megabases). Most chromosome arms (35 out of 43) show a gradient of dwindling transcriptional diversity accompanied by an increasing number of interchromosomal duplications in proximity to the centromere. At least 30% of the centromeric transition region structure originates from euchromatic gene-containing segments of DNA that were duplicatively transposed towards pericentromeric regions at a rate of six–seven events per million years during primate evolution. This process has led to the formation of a minimum of 28 new transcripts by exon exaptation and exon shuffling, many of which are primarily expressed in the testis. The distribution of these duplicated segments is nonrandom among pericentromeric regions, suggesting that some regions have served as preferential acceptors of euchromatic DNA.

Centromeres and the corresponding euchromatic transition regions have been described as one of the last frontiers of eukaryotic genome sequencing[1]. Before the sequencing of the human genome, the model for the organization of these regions was relatively simple (Fig. 1a)[2,3]. Although a more complex organization has begun to become apparent (Fig. 1b), the true sequence nature of these transition regions remained poorly characterized during the initial draft of the human genome, due in part to the paucity of unique mapping reagents near centromeres and artefacts associated with sequence and assembly of duplicated sequences[4–6]. During the final phases of the 'completion' of the human genome, considerable resources were directed towards improving the quality of these problematic areas (see Methods and Supplementary Methods).

Centromeric transition regions and assembly quality

We assessed the completeness of centromeric transition regions within the finished genome (build34, July 2003) by a series of experimental and computational analyses (Fig. 2; see also Supplementary Tables 1–4). First, we analysed the sequence composition of each of the 43 targeted pericentromeric regions. We found that 29 out of 43 (67.4%) of these show a minimum of 10 kilobases (kb) of satellite sequence positioned within the most proximal location of each chromosome arm (Fig. 2). Seven show a near-perfect match with higher-order alpha-satellite DNA (Supplementary Table 2 and Supplementary Methods). These proximal sequence features are consistent with centromere DNA structure (see Fig. 1). As expected, an abundance of duplicated segments is observed in close proximity to alpha-satellite DNA (<5 Mb). Gaps in the sequence assembly are particularly prevalent in these areas and show the strongest association with segmental duplications (76 out of 78 pericentromeric gaps are flanked by segmental duplications). Using a fluorescence *in situ* hybridization (FISH)-based assay to assess the multi-site distribution of segmental duplications, we estimate that 26.7% (82 out of 307 signals) of pericentromeric duplications are absent (Methods; see also Supplementary Tables 5, 6). A second, sequence-based assay calculates that approximately 34% of sequence-tagged sites cannot be identified within the current finished genome assembly (Supplementary Table 7). We conservatively estimate that ~4 Mb of satellite-rich sequence and ~6.5 Mb of highly duplicated material remain to be sequenced as part of these transition regions. This is in addition to the estimated 200 Mb of missing sequence that constitutes heterochromatic and acrocentric portions of the human genome.

It should be noted that this clone-order-based assembly (build 34) differs significantly from whole-genome shotgun sequence assembly (WGSA) of the human genome[7]. We analysed a recently published WGSA of the human genome and found that in this assembly an additional 19% (38.2 Mb) of the pericentromeric sequence is not assembled (24 Mb), not assigned (11.3 Mb) or misassigned (2.3 Mb). We estimate that more than 40% of the duplicated sequence presently assembled within build 34 might be incorrectly mapped by this WGSA assembly. The clone-order-based assembly of the human genome, therefore, provides one of the first detailed views of the organization of centromeric transition regions within mammalian genomes. As most of the eukaryotic genome projects have now adopted WGSA strategies, it is unlikely that such regions will be readily resolved in the future unless targeted efforts are undertaken.

Duplication organization

We assessed a variety of sequence properties (repeat content, duplication features, exon density, per cent G + C composition, and so on) as a function of distance from each putative centromere (Fig. 3; see also Supplementary Fig. 1a, b). Two significant features were noted: reduced gene density and increased duplication content. On the basis of our global analysis of segmental duplications (E.E.E., manuscript in preparation)[6,8], we have developed a working model for human pericentromeric organization (Fig. 1b). The majority of human pericentromeric DNA (29 out of 43 chromosome arms) shows evidence of blocks of segmental duplication located in close proximity to centromeric satellite DNA. Within a 5-Mb window of the centromere, we found that 22.7% (47.2 out of 207.9 Mb) of the bases are duplicated (sequence identity and length thresholds of >90% and >1 kb, respectively). Pericentromeric regions account for 31.1% of all duplicated bases (47.2 out of 152 Mb) and nearly 33.2% of all pairwise alignments (8,384 out of 25,239) for the entire human genome. Simulations confirm that segmental duplications are significantly ($P < 0.0001$) enriched (six–sevenfold) near centromeres when compared with a random genome model.

The pericentromeric enrichment for segmental duplications is most markedly seen for interchromosomal pairwise alignments where 5,357 out of 14,860 (36.0%) of all duplications between chromosomes occur within the first 5 Mb of the centromere. The proportion of interchromosomal duplications is most pronounced the closer the proximity to the centromere, where a clear gradient effect is observed (Fig. 3). Within the first 500 kb, interchromosomal duplications outnumber intrachromosomal duplications 6 to 1. The proportion of interchromosomal to intrachromosomal pairwise alignments drops from 2.6 to 1.66 as 2 Mb and 5 Mb pericentromeric regions are considered, respectively. By 4.5 Mb, there is a noticeable decline in all duplications. Centromeric satellite sequences were significantly ($P < 0.001$) enriched precisely at the integration sites of segmental duplication, suggesting that satellite sequences have had a role in this process of non-homologous interchromosomal exchange[9]. In addition to these sequences, various classes of low complexity and simple repeat sequences mapped within 500 base pairs (bp) of the duplication boundaries (Supplementary Table 8). We observed a correlation ($r^2 = 0.4509$) between the number of such repeat elements and the number of pericentromeric duplications.

Although pericentromeric regions are, in general, enriched for duplication, there is considerable variability. Three groups may be distinguished. Pericentromeric regions where the duplication content is below the genome average (<5.2%) (5p11, 4q11, 19q11, 18q11, 8p11, Xp11, 6q11 and 16q11) show a relatively sharp transition between unique and alpha-satellite DNA[10–12]. Sixteen pericentromeric regions (1p11, 3p11, 3q11, Xp11, 4p11, 5q11, 8q11, 17q11, 12p11, 19p11, 11q11, 12q11, 14q11, 20q11, 20p11 and Yq11) show an intermediate level of duplication between the genome and pericentromeric average (5.2–32.2%). Nineteen pericentromeric regions (10q11, 16p11, Yp11, 13q11, 2q11, 6p11, 7q11, 11p11, 21q11, 22q11, 10p11, 18p11, 17p11, 7p11, 1q11–1q12, 2p11, 15q11, 9q11 and 9p11) show extensive zones of duplication ranging from 500 kb to 5.5 Mb in length (Fig. 2, see also Supplementary Table 1c–f).

Specific constellations of pericentromeric regions share a greater number of, and in general have larger, segmental duplications. Regions 16p11, 15q11, 2p11, 7p11, 7q11 and 22q11 define one of the largest cohorts with approximately 22.7% (10.7 out of 47.2 Mb)

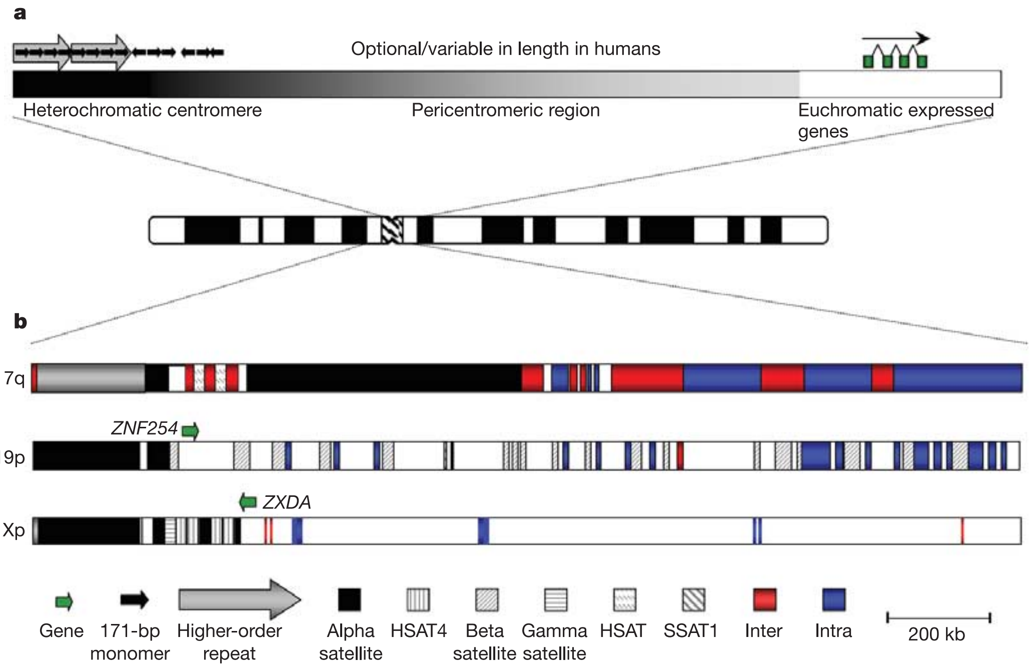


Figure 1 Models of centromeric transition regions. **a**, Pre-genome sequence model of pericentromeric organization: tandem reiterations of higher-order alpha-satellite DNA constitute larger array structures whose precise composition is diagnostic for a particular chromosome[35]. Blocks of alpha-satellite DNA lacking higher-order structure as well as other pericentromeric satellite DNA sequences map to the periphery[2,11,36,37]. In some cases, such as 9q12, 16q12 and 1q12, these peripheral satellite DNAs became sufficiently large to warrant their own cytological designations known as a secondary constriction[36–38]. **b**, Models of pericentromeric organization based on three sequenced chromosomes[11,26,39] showing various degrees of duplication content and interstitial satellite content[40]. Chromosome 7q represents a high level of segmental duplication whereas 19p represents an intermediate level and Xp a low level of segmental duplication.

of all duplicated bases shared between these six regions (Fig. 4; see also Supplementary Fig. 2). 18p11/21q11 and 9p11/20q11 define two smaller, more ancient, associations (Fig. 4). This distinction between quiescent and active regions of pericentromeric duplication is generally supported by our detailed FISH analyses of pericentromeric regions (Supplementary Tables 5, 6) and is, therefore, unlikely to be an artefact of missing sequence. We cannot exclude the possibility, however, that such regions may have become active in different primate lineages, as our analysis is based largely on examination of the contemporary human genome structure. In addition to the 43 pericentromeric regions near active centromeres, three other ancestral centromeres (2q21, 9q12 and 15q25.6) have been described[13–15] that were inactivated by chromosomal rearrangements during the human lineage of evolution. Each of these regions was marked by an abundance of pericentromeric duplications (Supplementary Table 1c, d). As suggested earlier[13], euchromatic regions characterized by an abundance of pericentromeric duplications may provide an evolutionary footprint of ancestral primate centromeres that have subsequently disappeared as a result of chromosomal fusion events. We identified five additional regions of the human genome (3p12, 3q21, 7q11, 13q12 and 22q11) that were marked by the presence of satellite sequences and an abundance of pericentromeric duplications.

Our detection of segmental duplications is based on arbitrary sequence identity and length thresholds (>90% and >1 kb). To address the question of whether the paucity of segmental duplications for certain chromosomes might simply be a consequence of these criteria, we performed a genome-wide analysis to detect more divergent segmental duplications (>250 bp and >75% sequence identity) (Supplementary Fig. 3). For those pericentromeric regions that harbour extensive duplication, the transition region between unique and duplicated regions has remained relatively precise even when the placement of more divergent duplications is considered. Among the duplication-quiescent centromeres (regions 5p11, 3q11, 4q11, 18q11 and 6q11) virtually no additional duplications were detected near the centromere, indicating that these regions have not been targets of segmental duplication during the entire course of human–primate evolution (Supplementary Fig. 3). Notably, for 8p11, 16q11 and 19q11, we observed small patches (<200 kb) of segmental duplication (<90% identity) that extended distally from the centromere. One interpretation may be that these regions were once capable of accepting duplications but subsequently became quiescent during the last 40 million years of chromosome evolution, as the segmental duplications all show >10% sequence divergence[16]. Another explanation for quiescent pericentromeric regions may be that chromosome rearrangement[17] or centromere-repositioning events have uncoupled active centromeres and zones of pericentromeric duplication such that certain regions now appear quiescent. The pericentric inversion of 18q11 specifically within the human lineage[14] and the recent emergence of the chromosome 6 centromere[18] might explain the dearth of pericentromeric duplications for these chromosome arms.

Euchromatic colonization of human pericentromeric DNA

To understand further the evolutionary dynamics of pericentromeric DNA, we targeted one chromosome, 2p11, for a more systematic analysis. We constructed the first sequence contig (737 kb) representing an autosomal transition from euchromatin to higher-order alpha-satellite DNA (Fig. 5). We validated the organization of the region by Southern, paralogous sequence-tagged site (STS) content and extended fibre-FISH analysis (Supplementary Methods and J.E.H., unpublished data). It should be pointed out that sequence closure in this region required extensive sequence redundancy as well as considerably more validation than other relatively ‘unique’ regions of the human genome owing to the presence of highly identical duplications as well as large-scale structural variation among different chromosomal haplotypes. Our initial sequence analysis showed that 91% of the 737-kb region consisted of segmental duplications.

Two types of duplication alignment were distinguished within 2p11. Fifty-seven per cent of the duplicated bases mapped to interstitial euchromatin located outside pericentromeric regions (Methods) whereas 34% of the duplications were part of alignments that mapped exclusively between pericentromeric regions. We classified these as ancestral duplicons and pericentromeric interspersed repeats, respectively[12,19]. Only 2% of the 2p11 sequence was devoid of duplications whereas the remaining alignments (~7%) mapped between pericentromeric and subtelomeric regions or the Y chromosome. Both the ancestral duplicons and pericentromeric repeats were distributed among multiple pericentromeric regions, most often as part of larger pairwise alignments. The pattern of duplication indicated a mosaic organization that had been formed by the duplicative transposition of at least 13 different ancestral euchromatic regions followed by secondary rounds of pericentromeric duplication[12,20–22]. To validate this euchromatic origin of the ancestral duplicons, we investigated nine of these regions in more detail. STS were designed within each duplicon and hybridized to genomic libraries from orang-utan (CHORI-251) and baboon (RPCI-41). All positive non-human primate clones were end-sequenced and mapped to the human genome by sequence similarity searches. In 7 out of 9 cases, the baboon data were consistent with a single, non-duplicated locus that mapped to a non-pericentromeric region of the human genome (Supplementary Table 9). In 6 out of 9 cases, the orang-utan clones mapped to the same locus, suggesting that these sites had been distributed to the pericentromeric region of human chromosomes relatively recently during evolution (<14 million years ago). Finally, mouse–human synteny analysis showed collinearity of syntenic anchors extending from the duplicated region into genomic sequence that flanked the putative ancestral segment but was not duplicated (Supplementary Fig. 4). These results unambiguously confirmed these sites as ancestral donor regions and provided directionality to the duplication events. An analysis of 15q11 similarly identified at least 16 different ancestrally donated euchromatic regions (Supplementary Fig. 5).

On the basis of our analysis of 2p11 and 15q11 (Locke, D. P. *et al.*, manuscript in preparation) as well as detailed studies of other human pericentromeric regions[8,12,20–28], we sought to identify the ancestral origin of all pericentromeric duplications that had emerged within the last 35–40 million years of human evolution (<10% sequence divergence). An ancestral locus was considered if it met three criteria: (1) it is not located within 5 Mb of the centromeric DNA; (2) most of the pairwise alignments underlying the locus map to pericentromeric regions of the human genome; and (3) mouse conserved synteny extends beyond the duplication alignment as determined by BLASTZ comparisons (http://www.genome.ucsc.edu) (Supplementary Methods). We analysed 8,343 pericentromeric duplication alignments and identified 271 (741 pairwise alignments) regions that met these criteria (Supplementary Table 10). These ancestral duplications correspond to 29.4% of all pericentromeric duplications (13.9 out of 47.2 Mb) within a 5-Mb window of the centromere. The putative ancestral donor loci ranged in length from 1 kb to 586 kb (average = 39.4 kb; median = 9.1 kb) and were duplicated on average to 2.73 (741/271) different pericentromeric regions. A total of 109 out of 271 (40%) of these donor sites contained intron–exon structure, suggesting that this process had been responsible for the mobilization of entire genes or partial gene fragments.

These data indicate that at least 30% of human pericentromeric duplications originated as transposed euchromatic sequence that was dispersed towards centromeric regions during hominoid chromosome evolution. By count or total number of duplicated bases, most of the ancestral duplications showed 94–97% sequence identity (Supplementary Fig. 6). We observed a marked reduction (twofold by count and tenfold by number of duplicated bases) for ancestral duplications that showed >98% sequence identity. Comparative and phylogenetic data suggest a continuum of events with a particular burst of activity after the separation of the Old World monkey species but before the radiation of the great-ape species (0.02 to 0.05 substitutions per site; 10–25 million years ago). Subsequent pericentromeric–pericentromeric duplications differentially distributed blocks within specific great-ape lineages leading to quantitative and qualitative differences[8,12,20–22,29].

Whereas donor loci appear to be randomly distributed, the pericentromeric dispersal was not uniform (Table 1). Several pericentromeric regions are significantly enriched ($P < 0.0012$), indicating that these particular regions have been preferential acceptors of duplicatively transposed material whereas others may have been protected or may appear quiescent owing to recent large-scale deletion or rearrangement. Four pericentromeric regions alone (7q11, 16p11, 15q11, and 17p11) account for 37.4% of the ancestral duplication alignments (277 out of 741) (Table 1). The extent of pericentromeric duplications among other non-human primates and mammalian organisms has only begun to be addressed[30–32]. Our analysis, however, predicts that the current sequence architecture of many human centromeric transition regions is a derived property where syntenic relationships rapidly decay.

Pericentromeric transcripts

For most species, pericentromeric regions are generally regarded as transcriptionally poor[1,33]. We measured transcript and gene density as a function of distance from each centromere using annotated known genes, Refseq genes and spliced expressed sequence tags (ESTs) (Fig. 3b; see also Supplementary Table 11a, b). Gene and exon density gradually increase as distance from the centromere increases. A noticeable reduction in exon density was observed within 2 Mb of the centromere when compared to the genome

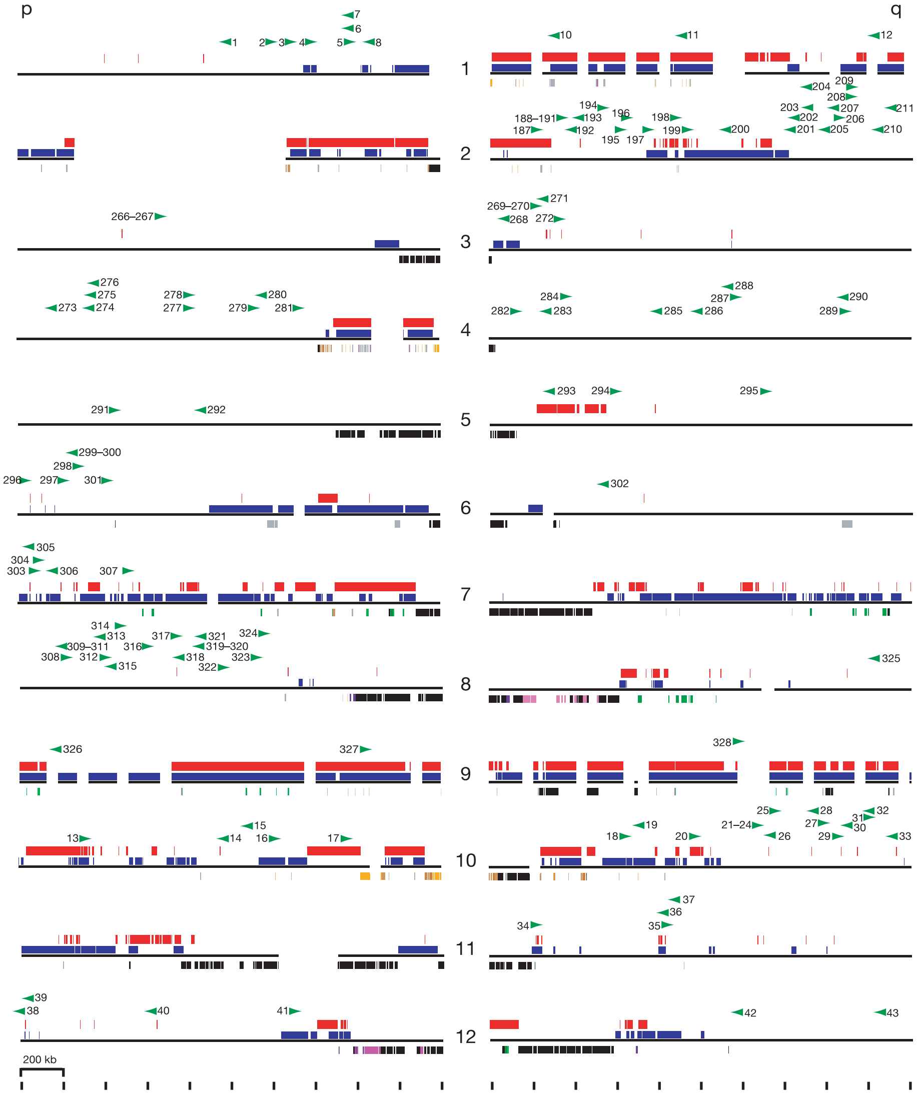
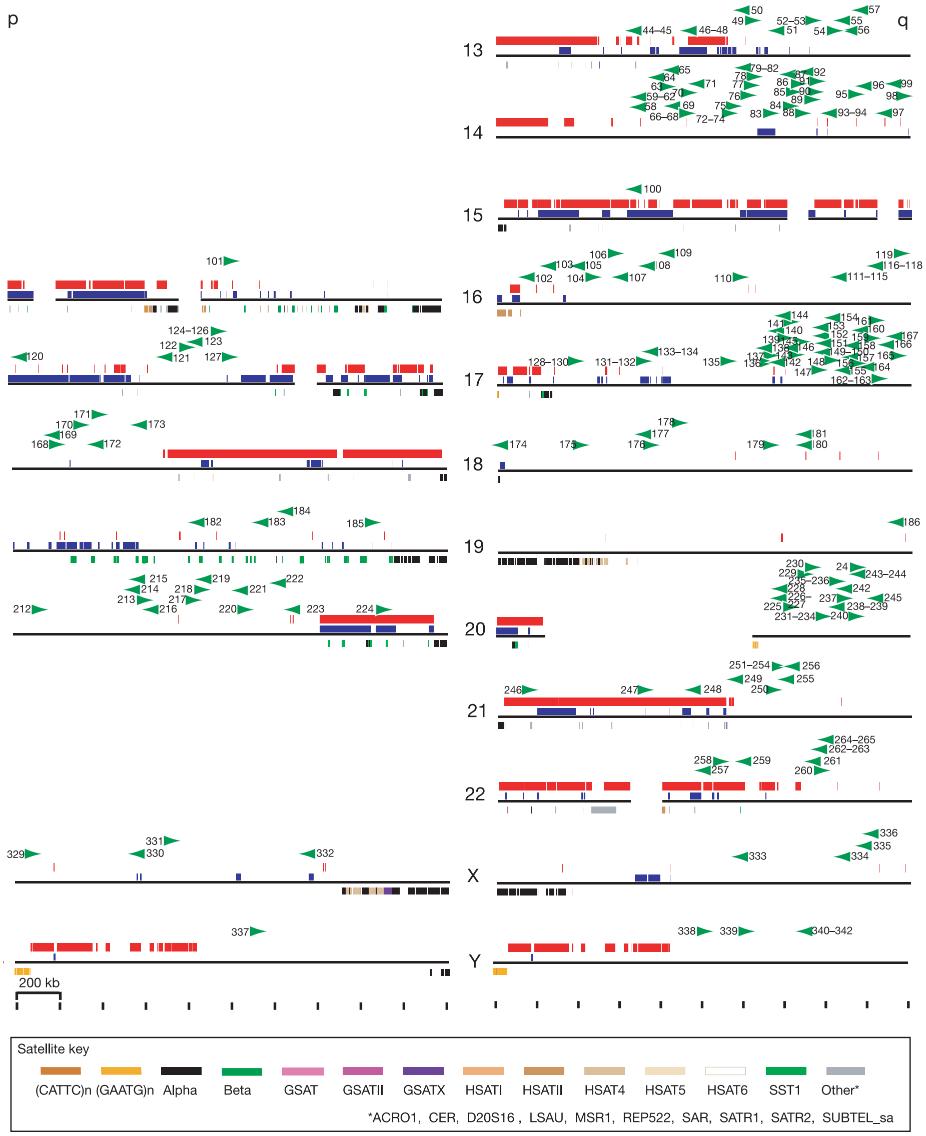


Figure 2 Pericentromeric architecture. The first 2 Mb on either side (p and q) of the centromere are shown for each chromosome. The centromere position was defined as the most proximal base pair for each chromosome arm based on the finished genome assembly (July 2003). Intrachromosomal (blue) and interchromosomal (red) duplications (>90% sequence identity, >1 kb in size) are shown above the line. Gaps (>50 kb in size) are shown as breaks in the black line. Centromeric satellite repeat composition is shown below the horizontal line according to the key. Significant blocks (>10 kb) of alpha-satellite or other centromeric satellite DNA is observed for 29 out of 43 pericentromeric regions. Acrocentric arms were not targeted as part of the human genome project and are therefore not shown. See Supplementary Fig. 3 for the distribution of pairwise alignments over a larger pericentromeric region (10 Mb), as well as a wider range of alignment divergence (0–0.25 substitutions per bp). Approximate locations of known genes are depicted by green arrows indicating the direction of transcription (see Supplementary Table 10a for corresponding gene names).

average (4.22 genes per Mb versus 7.3 genes per Mb). To test whether this reduction was significant, we randomly reassigned pericentromeric regions in the human genome and assessed exon content at 2 and 5 Mb. A significant reduction was observed at 2 Mb ($P = 0.0008$) but not at 5 Mb ($P = 0.07$).

Although transcriptional activity within pericentromeric regions is uniformly reduced, the transposition of gene-rich euchromatic segments and the rapid evolutionary turnover of such regions creates the potential for the formation of new transcripts[24,25]. We identified a total of 28 genes/messenger RNAs that had been completely duplicated within pericentromeric regions and for which there was evidence of transcription (as determined by best EST placement) (Supplementary Table 12 and Fig. 7). In addition to complete gene duplications, two other types of transcript innovations have been noted within pericentromeric regions: 'fusion' transcripts formed by the splicing of exons from two different duplicon modules, and 'exapted' transcripts which acquire one or more exons outside of the ancestral duplicated region. We identified a total of 11 fusion and 17 exapted transcripts representing 28 novel transcript clusters. Eleven of these (Supplementary Table 12, Supplementary Fig. 7) were associated with predicted genes with open reading frames and may therefore represent emerging genes.

We selected 16 distinct pericentromeric genes, mRNA and/or ESTs where there was evidence of either exon fusion or exaptation for further expression analysis. We specifically designed polymerase chain reaction with reverse transcription (RT–PCR) assays at the site of fusion/exaptation and tested a panel of eight tissues (Supplementary Table 13). Almost all assays (15 out of 16) amplified complementary DNA from the testis and more than half showed evidence of transcription (9 out of 16) from the ovary. Interestingly, 7 of the assays were exclusively expressed from the testis. These data suggest that germline tissues are much more likely to express novel



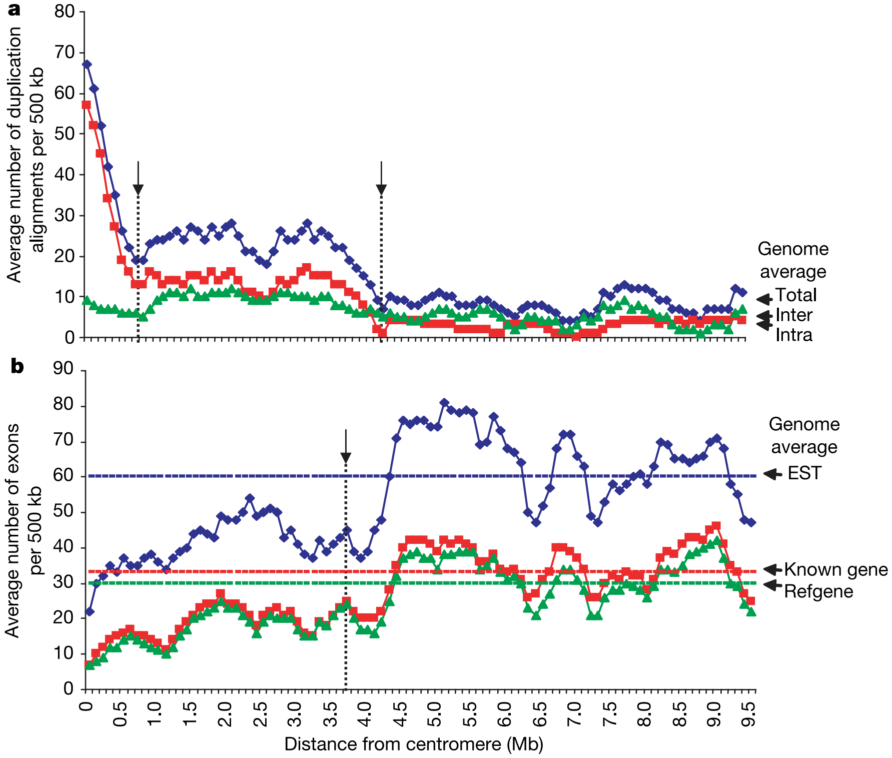


Figure 3 Sequence properties of centromeric transition regions. We computed a series of sequence properties (duplication content and exon density) in 500-kb windows (100-kb increments) for the first 10 Mb of each human chromosome arm beyond the centromere. The figures are based on the average for 30 pericentromeric regions where a large block of satellite sequence has been identified at the most proximal position. **a**, Plot of the average number of duplication alignments (blue diamonds, total) for both interchromosomal (red squares) and intrachromosomal (green triangles) duplications. Significant changes in interchromosomal duplication and overall duplication content are noted at ~1 and 4.5 Mb, respectively. **b**, The average number of exons based on analysis of ESTs (blue diamonds), known genes (red squares) and RefSeq genes (green triangles) are shown. Each EST, known gene and RefSeq gene is placed uniquely based on highest sequence similarity scores. A significant decrease ($P = 0.008$) in exon density is observed at 2 Mb but not 5 Mb ($P = 0.07$). ESTs or genes with multiple tied placements are counted only once. See Supplementary Fig. 1 for other sequence properties and a breakdown by individual chromosome.

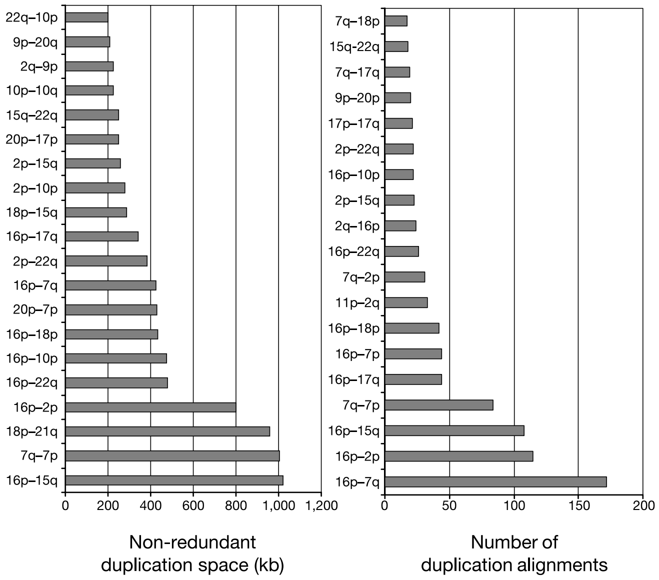


Figure 4 Cohorts of pericentromeric duplication. The histograms show the pericentromeric regions by cytogenetic location that are most preferentially associated by duplication. The top 20 regions are arranged based on the largest amount of shared duplicated sequence within 5 Mb of the centromere. Pericentromeric regions still represent work in progress. As additional sequence is generated, the rank order for specific pericentromeric regions may change.

pericentromeric transcripts than any other human tissue. Different packaging constraints of pericentromeric chromatin in germline tissue may contribute to this effect.

In summary, our analysis indicates that most human pericentromeric regions have been subjected to a complex series of duplications during the course of evolution (Fig. 2) with a gradient effect of interchromosomal duplications biased towards the centromere. Our delineation of ancestral donor sequences allows us to estimate a lower-bound rate for this process. We identified 271 ancestral segmental duplications to 43 pericentromeric regions over an estimated 40 million years of human evolution. We calculate an effective fixation rate of six–seven transposition events per million years. Subsequent pericentromeric duplications of these ancestral loci predict that the rate of duplication among non-homologous chromosomes was at least three times higher with an average fixation rate of about 20 events per million years. We believe that these two estimates are conservative as not all ancestral segments could be identified by mouse synteny analysis. If the total number of pericentromeric duplication alignments (8,343) is used as a surrogate for duplication/rearrangement events, the estimate may be as high as 206 duplication/rearrangement events per million years. Although the rate of pericentromeric duplication has been extensive, only a few juxtapositions of ancestral cassettes have created new transcripts. On the basis of our analysis, we estimate that a novel or mosaic transcript may have emerged through pericentromeric duplication once every million years of evolution. The fate and function of such evolutionary novelties remain to be determined. □

Table 1 **Distribution of ancestral duplicons among pericentromeric regions**

	Ancestral duplicons				Pericentromeric regions (5 Mb)			
Chromosome arm	Observed	Expected (±2 s.d.)	Poisson (*P*)	Simulation	Observed	Expected (±2 s.d.)	Poisson (*P*)	Simulation
Chr 1p†	9	11.7 ± 6.7	0.0939	2,587	3	18.7 ± 7.9	0.0000	0
Chr 1q	11	10.3 ± 6.3	0.1166	4,594	13	14.2 ± 7.1	0.1042	4,314
Chr 2p†	11	9.0 ± 5.9	0.0970	2,920	41	14.4 ± 7.0	*0.0000*	*0*
Chr 2q†	11	14.8 ± 7.4	0.0698	1,904	6	18.3 ± 7.9	0.0006	2
Chr 3p†	4	8.9 ± 5.8	0.0357	547	0	19.0 ± 8.0	0.0000	0
Chr 3q†	7	10.3 ± 6.3	0.0821	1,912	5	18.9 ± 8.0	0.0001	1
Chr 4p*†	13	4.5 ± 4.1	*0.0006*	7	5	18.3 ± 8.0	0.0002	0
Chr 4q†	10	13.8 ± 7.3	0.0701	1,815	0	18.9 ± 8.0	0.0000	0
Chr 5p†	7	4.3 ± 4.1	0.0732	1,415	2	19.0 ± 8.1	0.0000	0
Chr 5q†	8	13.1 ± 7.0	0.0440	874	4	18.9 ± 8.0	0.0000	0
Chr 6p	2	5.6 ± 4.6	0.0580	779	9	18.6 ± 8.0	0.0061	79
Chr 6q†	9	10.7 ± 6.4	0.1142	3,609	5	18.7 ± 8.1	0.0001	2
Chr 7p†	4	5.4 ± 4.6	0.1600	3,666	41	18.7 ± 8.0	*0.0000*	*0*
Chr 7q†	10	9.5 ± 6.0	0.1235	4,799	61	18.4 ± 7.9	*0.0000*	*0*
Chr 8p†	1	4.0 ± 4.0	0.0733	925	0	19.0 ± 8.1	0.0000	0
Chr 8q*	0	9.7 ± 6.1	0.0001	1	7	18.5 ± 7.9	0.0013	7
Chr 9p†	3	4.0 ± 4.0	0.1954	4,212	31	14.2 ± 7.2	*0.0000*	*0*
Chr 9q†	2	7.0 ± 5.2	0.0223	272	37	15.1 ± 7.2	*0.0000*	*0*
Chr 10p	4	3.5 ± 3.7	0.1888	4,673	24	18.7 ± 8.1	0.0408	1,194
Chr 10q	15	9.1 ± 5.8	0.0208	352	20	17.8 ± 7.9	0.0776	3,185
Chr 11p	2	4.8 ± 4.3	0.0948	1,387	14	17.6 ± 7.8	0.0715	2,150
Chr 11q†	9	7.8 ± 5.5	0.1207	3,704	3	18.9 ± 8.2	0.0000	0
Chr 12p†	6	3.0 ± 3.4	0.0504	866	7	18.9 ± 8.0	0.0011	9
Chr 12q†	8	9.3 ± 6.0	0.1269	4,056	6	19.0 ± 8.0	0.0004	0
Chr 13q†	10	9.4 ± 6.0	0.1228	4,667	37	18.9 ± 8.0	*0.0001*	*1*
Chr 14q†	12	8.5 ± 5.7	0.0604	1,454	5	18.9 ± 8.1	0.0001	1
Chr 15q†	7	7.9 ± 5.5	0.1413	4,526	73	17.8 ± 7.9	*0.0000*	*0*
Chr 16p†	2	3.2 ± 3.6	0.2087	3,798	92	18.1 ± 7.9	*0.0000*	*0*
Chr 16q†	7	4.0 ± 3.9	0.0595	1,021	3	18.8 ± 8.0	0.0000	0
Chr 17p†	6	1.7 ± 2.6	0.0061	73	51	18.3 ± 7.9	*0.0000*	*0*
Chr 17q*	18	5.2 ± 4.5	*0.0000*	*0*	30	18.9 ± 8.1	0.0047	76
Chr 18p	1	1.0 ± 2.0	0.3679	6,498	20	18.9 ± 8.1	0.0862	4,344
Chr 18q†	3	5.6 ± 4.7	0.1082	1,907	0	19.0 ± 8.1	0.0000	0
Chr 19p†	6	2.0 ± 2.7	0.0120	3,183	6	18.9 ± 8.0	0.0004	0
Chr 19q†	3	2.7 ± 3.2	0.2205	5,092	0	18.9 ± 8.1	0.0000	0
Chr 20p	4	2.2 ± 2.9	0.1082	1,769	14	18.9 ± 8.1	0.0522	1,360
Chr 20q†	6	3.0 ± 3.4	0.0504	779	4	15.1 ± 7.2	0.0006	5
Chr 21q	3	2.9 ± 3.4	0.2237	5,605	23	19.0 ± 8.0	0.0556	1,890
Chr 22q	6	3.0 ± 3.5	0.0504	824	29	18.1 ± 8.0	0.0045	82
Chr Xp†	4	5.3 ± 4.6	0.1641	3,882	3	19.0 ± 8.0	0.0000	0
Chr Xq†	7	9.1 ± 5.9	0.1145	3,094	7	18.9 ± 8.0	0.0011	11

The number of ancestral duplicons originating within each chromosome arm (donor loci) and the number of observed duplicate copies observed within each 5-Mb pericentromeric region (acceptor loci) were counted (Methods). The chromosomal arms in which pericentromeric regions have a significant transition block (>10 kb) of alpha-satellite or other centromeric satellite DNA are in bold font. We observed 271 ancestral regions duplicated to 741 pericentromeric locations in the human genome. The expected number (±2 s.d.) indicates the range for expected number of ancestral duplications or pericentromeric duplications by random distribution model. We tested for significant departures from a random genome distribution model by both Poisson sampling and by simulation. Significance values were corrected for multiple tests (Bon-Ferroni $P < 0.0012$). Regions significantly enriched in duplication are italicized, whereas regions 'protected' for duplication are underlined. The 95% confidence interval indicates the range of the expected number of ancestral or pericentromeric duplications based on a simulation of 10,000 replicates. The simulation reports the number of tests (from 10,000 replicates) that were equal to or exceeded the observed count for regions that were enriched, or the number of tests that were equal to or less than the observed count for regions that were protected. Most ancestral donor loci are randomly distributed but pericentromeric regions show a highly nonrandom distribution due largely to secondary duplications among specific cohorts of chromosomes.

*Significant departure from a random distribution model for donors.

†Significant departure from a random distribution model for acceptors.

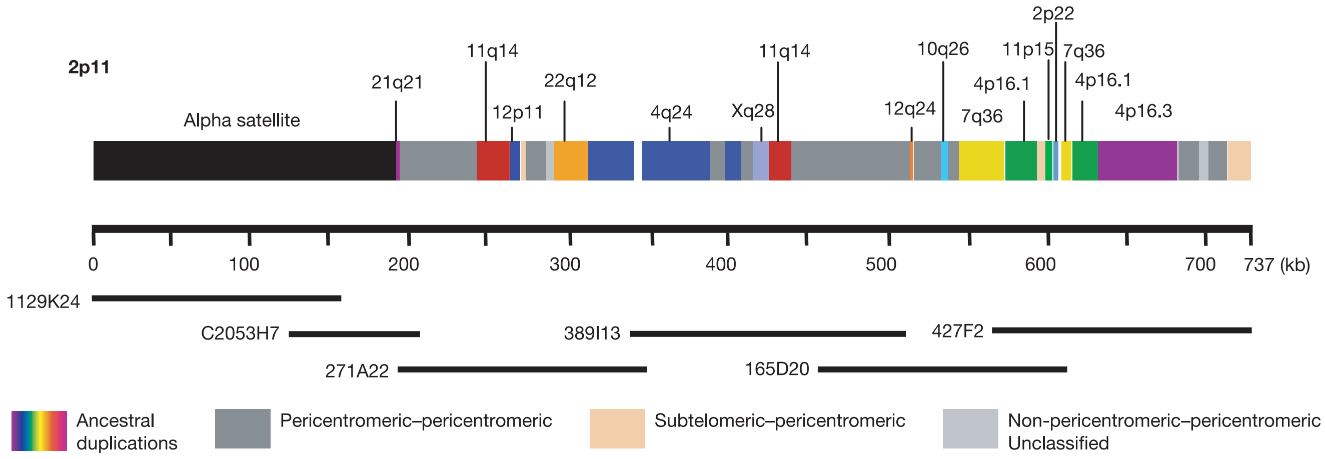


Figure 5 Ancestral duplicons within 2p11. The modular organization of a pericentromeric region, 2p11, is depicted based on the classification of the underlying pairwise alignments. Duplicated segments that originate outside of the pericentromeric region, termed ancestral duplicons, are shown in colour (ancestral cytogenetic band locations are delineated). Unshaded regions correspond to regions where no underlying duplication could be detected. The minimal tiling path of large-insert BAC clones is drawn to scale below each line. A 737-kb validated sequence contig that provides the first autosomal transition into higher-order alpha-satellite repeat DNA is shown here. Approximately 98% of this region is composed of duplicated material of which 57% can be traced back to non-pericentromeric regions of the genome. These correspond to 13 ancestral duplicons of which 9 were experimentally confirmed by non-human primate analyses.

Methods

Characterization and validation of pericentromeric DNA

The definition of pericentromeric DNA and its boundary is arbitrary. Noticeable changes in exon density, satellite repeat and interchromosomal duplication content were observed at 2.6 and 5.2 Mb (Fig. 3). In this study, therefore, we limited our analysis of pericentromeric DNA to both 2 and 5 Mb. To facilitate the assembly and to assess the quality of these regions, we performed STS content analysis, FISH and detailed characterization of alpha-satellite repeat content (Supplementary Methods). The origin of ancestral duplications for the 2p11 region was determined based on detailed comparative and phylogenetic analysis between humans and non-human primate species (Supplementary Methods). All genome-wide analyses were performed based on the July 2003 finished genome assembly (build34) and are available at http://humanparalogy.gene.cwru.edu.

Duplication analyses

We used a BLAST-based detection scheme[6] to initially identify all pairwise similarities representing duplicated regions ($\geq$1 kb and $\geq$90% identity) within the finished sequence of the human genome (July 2003). A total of 25,239 welded pairwise alignments were generated, of which 8,343 mapped within 5 Mb of human pericentromeric DNA. To eliminate potential artefactual duplications due to misassembly, we considered only those alignments that could be confirmed using a second assembly-free method of detection (whole-genome shotgun sequence detection[24]). To test the significance of the observed enrichment of duplicated sequences within pericentromeres (5 Mb around the centromere), a conservative model for the duplication process by randomly reassigning contiguous blocks of duplicated sequence to new locations was used. In order to detect more divergent duplications, a second all-by-all genome BLASTZ discontinuous search was performed within the finished genome to recover more divergent ($>$75%) and shorter ($>$250 bp) alignments (W.M., unpublished data) (Supplementary Methods).

Ancestral origin of pericentromeric duplications

Using mouse synteny data, we could classify 741 of the 8,343 pericentromeric duplication alignments as ancestral and thereby define the directionality of the duplication event (see Supplementary Methods for details). Regions were clustered if duplicon subgroups mapped $<$100 kb of one another and the sequence divergence of pericentromeric alignments showed $<$2% difference. These ancestral duplications correspond to 8.1 Mb or 30% of all pericentromeric duplicated base pairs (8.1 out of 26 Mb). The remaining 18 Mb corresponded to pericentromeric duplications that did not have a euchromatic origin and/or regions where insufficient mouse–human synteny data existed to claim directionality of the duplication event. Ancestral loci were examined for the presence of intron–exon structure based on known genes (http://genome.ucsc.edu). A gene segment was considered if at least a single exon and intron could be identified within the duplicated segment. The Y chromosome is not included in this analysis due to the lack of mouse synteny information.

Gene and transcript analysis

We examined transcriptional potential by considering both the number of genes and exons in 500-kb windows (including gaps) sliding by 100-kb increments from the centromere for each chromosome. Four sets of data were considered: Refseq annotated genes (28,452), known annotated genes (38,482), mRNAs (130, 762) and human ESTs with intron/exon structure (2.3 million). Both best-placement and tied genes/transcripts were distinguished based on BLAT score criteria (http://www.genome.ucsc.edu). Exon density was defined as the number of non-overlapping exons identified within a genomic window. We separately examined mRNAs and genes that mapped to duplicated regions near the centromere (5 Mb), considering only those where the sequence identity between genome and cDNA exceeded 99%. A total of 25 known and Refseq genes and 31 mRNA clusters, with intron/exon structure, were identified within the duplicated regions. The calculation for novel transcripts or genes was estimated based on sequence divergence of genomic sequence. Marmoset DNA shows $\sim$11–12% (ref. 34) divergence from human and they are estimated to have diverged 35–40 million years ago. On the basis of sequence divergence, the segmental duplications observed here probably do not exist in marmoset and therefore the 28 novel genes/mRNAs arose in the last 35 million years of evolution (28 genes per 35 million years = $\sim$1 gene per million years).

Received 21 February; accepted 2 July 2004; doi:10.1038/nature02806.

1. Nagaki, K. *et al.* Sequencing of a rice centromere uncovers active genes. *Nature Genet.* **36,** 138–145 (2004).
2. Horvath, J. *et al.* Molecular structure and evolution of an alpha/non-alpha satellite junction at 16p11. *Hum. Mol. Genet.* **9,** 113–123 (2000).
3. Jackson, M. Duplicate, decouple, disperse: the evolutionary transience of human centromeric regions. *Curr. Opin. Genet. Dev.* **13,** 629–635 (2003).
4. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
5. Eichler, E. E. Masquerading repeats: Paralogous pitfalls of the Human Genome. *Genome Res.* **8,** 758–762 (1998).
6. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11,** 1005–1017 (2001).
7. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA* **101,** 1916–1921 (2004).
8. Bailey, J. A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70,** 83–100 (2002).
9. Bailey, J. A., Giu, L. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73,** 823–834 (2003).
10. Horvath, J. E., Bailey, J. A., Locke, D. P. & Eichler, E. E. Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum. Mol. Genet.* **10,** 2215–2223 (2001).
11. Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. Genomic and genetic definition of a functional human centromere. *Science* **294,** 109–115 (2001).
12. Horvath, J. E. *et al.* Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications. *Mol. Biol. Evol.* **20,** 1463–1479 (2003).
13. Ventura, M. *et al.* Neocentromeres in 15q24–26 map to duplicons which flanked an ancestral centromere in 15q25. *Genome Res.* **13,** 2059–2068 (2003).
14. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215,** 1525–1530 (1982).
15. Baldini, A. *et al.* An alphoid DNA sequence conserved in all human and great ape chromosomes: evidence for ancient centromeric sequences at human chromosomal regions 2q21 and 9q13. *Hum. Genet.* **90,** 577–583 (1993).
16. Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13,** 358–368 (2003).
17. Ventura, M., Archidiacono, N. & Rocchi, M. Centromere emergence in evolution. *Genome Res.* **11,** 595–599 (2001).
18. Eder, V. *et al.* Chromosome 6 phylogeny in primates and centromere repositioning. *Mol. Biol. Evol.* **20,** 1506–1512 (2003).
19. Eichler, E. E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17,** 661–669 (2001).
20. Locke, D. P. *et al.* Molecular evolution of the human chromosome 15 pericentromeric region. *Cytogenet. Genome Res.* (in the press).
21. Golfier, G. *et al.* The 200-kb segmental duplication on human chromosome 21 originates from a pericentromeric dissemination involving human chromosomes 2, 18 and 13. *Gene* **312,** 51–59 (2003).
22. Courseaux, A. *et al.* Segmental duplications in euchromatic regions of human chromosome 5: a source of evolutionary instability and transcriptional innovation. *Genome Res.* **13,** 369–381 (2003).
23. Horvath, J., Schwartz, S. & Eichler, E. The mosaic structure of a 2p11 pericentromeric segment: A strategy for characterizing complex regions of the human genome. *Genome Res.* **10,** 839–852 (2000).
24. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
25. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. *Genome Res.* **13,** 159–172 (2003).
26. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424,** 157–164 (2003).
27. Crosier, M. *et al.* Human paralogs of KIAA0187 were created through independent pericentromeric-directed and chromosome-specific duplication mechanisms. *Genome Res.* **12,** 67–80 (2002).
28. Jackson, M. S. *et al.* 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 (1999).
29. Locke, D. P. *et al.* Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **13,** 347–357 (2003).
30. Thomas, J. W. *et al.* Pericentromeric duplications in the laboratory mouse. *Genome Res.* **13,** 55–63 (2003).
31. Tuzun, E., Bailey, J. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* **14,** 493–506 (2004).
32. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5,** R23 (2004).
33. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286,** 2468–2474 (1999).
34. Schneider, H. *et al.* Molecular phylogeny of the New World monkeys (Platyrrhini, primates) based on two unlinked nuclear genes: IRBP intron 1 and epsilon-globin sequences. *Am. J. Phys. Anthropol.* **100,** 153–179 (1996).
35. Willard, H. F. & Waye, J. S. 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 (1987).
36. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. Alpha-satellite DNA of primates: old and new families. *Chromosoma* **110,** 253–266 (2001).
37. Lee, C., Wevrick, R., Fisher, R. B., Ferguson-Smith, M. A. & Lin, C. C. Human centromeric DNAs. *Hum. Genet.* **100,** 291–304 (1997).
38. An International System for Human Cytogenetic Nomenclature, High resolution-banding. *Cytogenet. Cell Genet.* **31,** 1–23 (1981).
39. Leem, S. H. *et al.* Closing the gaps on human chromosome 19 revealed genes with a high density of repetitive tandemly arrayed elements. *Genome Res.* **14,** 239–246 (2004).
40. Rudd, M. K. & Willard, H. F. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* (in the press).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We are grateful to the large-scale sequencing centres (Baylor College of Medicine, Cold Spring Harbor Laboratory, Genome Therapeutics Corporation, Harvard Partners Genome Center, Joint Genome Institute, The NIH Intramural Sequencing Center, The UK-MRC Sequencing Consortium, The University of Oklahoma Advanced Center for Genome Technology, The University of Texas Southwest, The Whitehead Institute for Biomedical Research, The Washington University Genome Sequencing Center and the Wellcome Trust Sanger Institute) for access to all large-scale finished sequence, genome assembly and trace sequence data from the human genome before publication. This work was supported by grants from NIH and DOE to E.E.E. and grants from P.R.I.N.C.E., MURST and Telethon to M.R.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.E.E. (eee@gs.washington.edu).