Human Molecular Genetics, 1996, Vol. 5, No. 7 **899–912**

ARTICLE

Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution

Evan E. Eichler†, Fei Lu, Ying Shen, Rachele Antonacci, Vesna Jurecic, Norman A. Doggett[1], Robert K. Moyzis[1], Antonio Baldini, Richard A. Gibbs and David L. Nelson*

Department of Molecular and Human Genetics, Human Genome Center, Baylor College of Medicine, Houston, TX 77030, USA and [1]Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received February 22, 1996; Revised and Accepted April 26, 1996

We have identified a 26.5 kb gene-rich duplication shared by human Xq28 and 16p11.1. Complete comparative sequence analysis of cosmids from both loci has revealed identical Xq28 and 16p11.1 genomic structures for both the human creatine transporter gene (*SLC6A8*) and five exons of the CDM gene (*DXS1357E*). Overall nucleotide similarity within the duplication was found to be 94.6%, suggesting that this interchromosomal duplication occurred within recent evolutionary time (7–10 mya). Based on comparisons between genomic and cDNA sequence, both the Xq28 creatine transporter and *DXS1357E* genes are transcriptionally active. Predicted translation of exons and RT-PCR analysis reveal that chromosome 16 paralogs likely represent pseudogenes. Comparative fluorescent *in situ* hybridization (FISH) analyses of chromosomes from various primates indicate that this gene-rich segment has undergone several duplications. In gorilla and chimpanzee, multiple pericentromeric localizations on a variety of chromosomes were found using probes from the duplicated region. In other species, such as the orangutan and gibbon, FISH signals were only identified at the distal end of the X chromosome, suggesting that the Xq28 locus represents the ancestral copy. Sequencing of the 16p11.1/Xq28 duplication breakpoints has revealed the presence of repetitive immunoglobulin-like CAGGG pentamer sequences at or near the paralogy boundaries. The mobilization and dispersal of this gene-rich 27 kb element to the pericentromeric regions of primate chromosomes defines an unprecedented form of recent genome evolution and a novel mechanism for the generation of genetic diversity among closely related species.

INTRODUCTION

Duplications of genes involving the conservation of structure and gene order have been observed both intrachromosomally and interchromosomally in a variety of organisms (1–4). Intrachromosomal paralogies are generally restricted to discrete cytogenetic band locations. Recent investigations into the genomic organization of several human gene clusters such as the human *CEA* gene family (carcinoembryonic antigens) (5), human olfactory receptor cluster (6), the *ZNF* (zinc finger) gene family on human 19p12 (7), the pregnancy-specific glycoprotein gene family (5) and the *MAGE* (melanoma antigen gene) cluster on the short arm of the X chromosome (8) indicate that endoduplication of a single gene or an ancestral repertoire of genes has been the likely mechanism for the generation of the observed tandem array structures. Some cross-species comparisons reveal that such endoduplications have occurred quite recently in human evolution (7); while others, such as the duplications involved in the archetypal *Hox* cluster organization, are much more ancient, being closely conserved among all vertebrate species (4,9,10).

In contrast to intrachromosomal paralogies, duplications of genes between chromosomes are believed to be mechanistically and functionally distinct (11,12). Such duplications transfer cognate genes to new genomic environments, with the possibility to acquire novel combinations of long-range *cis*-acting regulatory elements and ultimately new functional roles in the organism

*To whom correspondence should be addressed
†Present address: Human Genome Center, Biology and Biotechnology Research Program, L-452, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

(13). The large paralogous segments between *S. cerevisae* chromosomes II and IV involving the functional *ARC* and *COR* gene clusters (1), the two copies of the insulin-growth factor genes on human chromosomes 11 and 12 (14) and the four copies of the vertebrate *HOX* cluster are a few examples of interchromosomal duplications which are presumed to have acquired functional diversity.

One theory postulated to account for the occurrence of interchromosomal paralogous genes has involved the duplication of chromosome number associated with major chordate speciation events (15–18). Tetraploidization followed by the re-establishment of the disomic state by macromutational processes such as translocations, inversions and Robertsonian fusions have been invoked to explain the large-scale paralogous gene organizations and the similarity in cytogenetic banding pattern observed between some human chromosomes, such as 11p and 12p (15,16,18–20). Three tetraploidization events are believed to have occurred during the evolution of mammals from protochordate ancestor, with the latest occurring approximately 250 mya (18). Interestingly, interspecific comparisons of gene organization and the estimation of the timing of duplication events between the α- and β-globins (21), the two insulin growth factor gene families (14,22) and the *HOX* gene clusters in vertebrates (4), all suggest that interchromosomal exchange are ancient events, occurring 300–500 million years ago. These results lend further support to the argument that interchromosomal paralogies in these families represent the remnants of early tetraploidization events (13,18).

We report the identification and complete sequence analysis of an interchromosomal paralogous region between human cytogenetic bands Xq28 and 16p11.1. At least two genes, the creatine transporter (*SLC6A8*) and the CDM (*DXS1357E*) gene, have been involved in this duplication event. Comparative sequence analysis of 26.5 kb of paralogous DNA and 12.5 kb of flanking sequence indicates an overall nucleotide similarity of 94.6%, suggesting that this duplication event has occurred relatively recently (7–10 mya) in the hominoid lineage of evolution. Comparative fluorescent *in situ* hybridization (FISH) analysis of orthologous 16p11.1 and Xq28 loci among the lesser and greater apes confirm that the duplications have occurred relatively recently during the pongid/hominid lineage of evolution and that the chromosome X paralogous gene-rich segment likely represents the ancestral template copy. Among gorillas and chimpanzees, duplications of the Xq28 sequence have been observed on multiple chromosomes which are non-syntenic to human chromosome 16. In each case, the targets of the duplication events have been within the pericentromeric regions of chromosomes, suggesting a region-specific directed mechanism for the generations of these interchromosomal duplications. Sequence determination of the breakpoints on human Xq28 and 16p11.1 has identified complicated interspersed CAGGG pentamer repeats located near or at the boundaries of interchromosomal paralogy. The similarity of these repeats to immunoglobulin switch-like recombination elements (CTGGG), the pericentromeric site-specificity of duplications, the burst of intrachromosomal paralogies involving these genes in a short period of evolution and the maintenance of the transcriptionally active X locus suggest that recombinogenic processes such as replicative transposition have been involved in the duplication, mobilization and integration of the Xq28 gene-rich segment to non-homologous regions of the primate genome. The documentation of such a large (26.5 kb) transposition which has occurred recently in the hominoid genome identifies a novel mechanism for interchromosomal paralogy which may have implications in generation of diversity of Na+Cl– dependent transporters and other multigene families of primates.

RESULTS

Identification of the Xq28-16p11.1 paralogy

Representative cosmids from various Xq28 contigs were mapped by Southern hybridization to X chromosome somatic-cell hybrid deletion panels and by FISH in order to determine the physical location of each set of contiguous cosmid clones. Cosmid clones from one contig (designated X1) consistently demonstrated strong hybridization signals corresponding to cytogenetic bands Xq28 and 16p11.1 by FISH (Fig. 1). Screening of arrayed cDNA libraries with these cosmids identified two different placental cDNA clones, p3A1 and p96D7. Sequence analysis of the 1.9 kb and 1.25 kb *Not*I inserts of p96D7 and p3A1 clones revealed that the cDNAs corresponded to the ORF of the recently described human creatine transporter gene(s) (*SLC6A8*) (GenBank accession L31409; GenBank accession S74309) (23,24) and the CDM (*DXS1357E*) gene (GenBank accession Z31696) (25). The first exon of the CDM gene had been previously mapped to a location 450 bp proximal to the 5′ UTR of the *ALD* (adrenoleukodystrophy) gene (25,26), confirming the Xq28 location of the X1 cosmid contig within the *BGN* (biglycan) and *L1CAM* (L1 cell-adhesion molecule) interval (Fig. 1) (27).

To eliminate the possibility that the cross-hybridization of the X1 contig cosmids to 16p11.1–16p11.2 and Xq28 was due to chimeric genomic clones, the p3A1 and p96D7 cDNA inserts were used as probes to screen a flow-sorted chromosome 16 arrayed cosmid library. A total of 15 cosmids were initially identified, the majority of which had been assembled by fingerprinting methods into a chromosome 16 contig, C177 (28). Cosmid Southern blots of *Pst*I and *Eco*R1 restriction digests of representative X1 and C177 cosmids probed with p3A1 and p96D7 cDNA inserts confirmed that paralogs of both CDM and the creatine transporter gene were found on 16 and X chromosome derived cosmids (data not shown). Cosmids within contig C177 were mapped by Southern analysis of somatic cell hybrid deletion panels of chromosome 16 to the *CY153*/*CY192B* interval, placing the contig between 16p11.1 and 16p11.2, as had been predicted (Fig. 1). Further screening of an STS derived from cosmid c318A11 within contig C177 established overlap with CEPH megaYACs My895G9 and My953C10 and refined the location by FISH and prior mapping data (28) of this contig to near the 16p11.1/16p11.2 Giemsa staining boundary. Finally, C177 cosmids used as probes in FISH of human chromosome metaphase spreads consistently hybridized to both 16p11.1–16p11.2 and Xq28 confirming the paralogy between these two chromosomal cytogenetic bands.

The extent of synteny between 16p11.1 and Xq28

In order to estimate the size of duplication between 16p11.1 and Xq28, a cosmid contig of approximately 150 kb was developed proximal to the creatine transporter locus extending distally to the 3′ UTR of the *ALD* gene of Xq28 (Fig. 1). In addition, various cDNA and genomic probes including pbgn900 (a 920 bp cDNA fragment of the biglycan locus), TA19 (a genomic clone including exons 6 and 7 of *ALD*), c6 (a 3.5 kb genomic subclone which includes the first exon of creatine transporter gene) and H8 (a

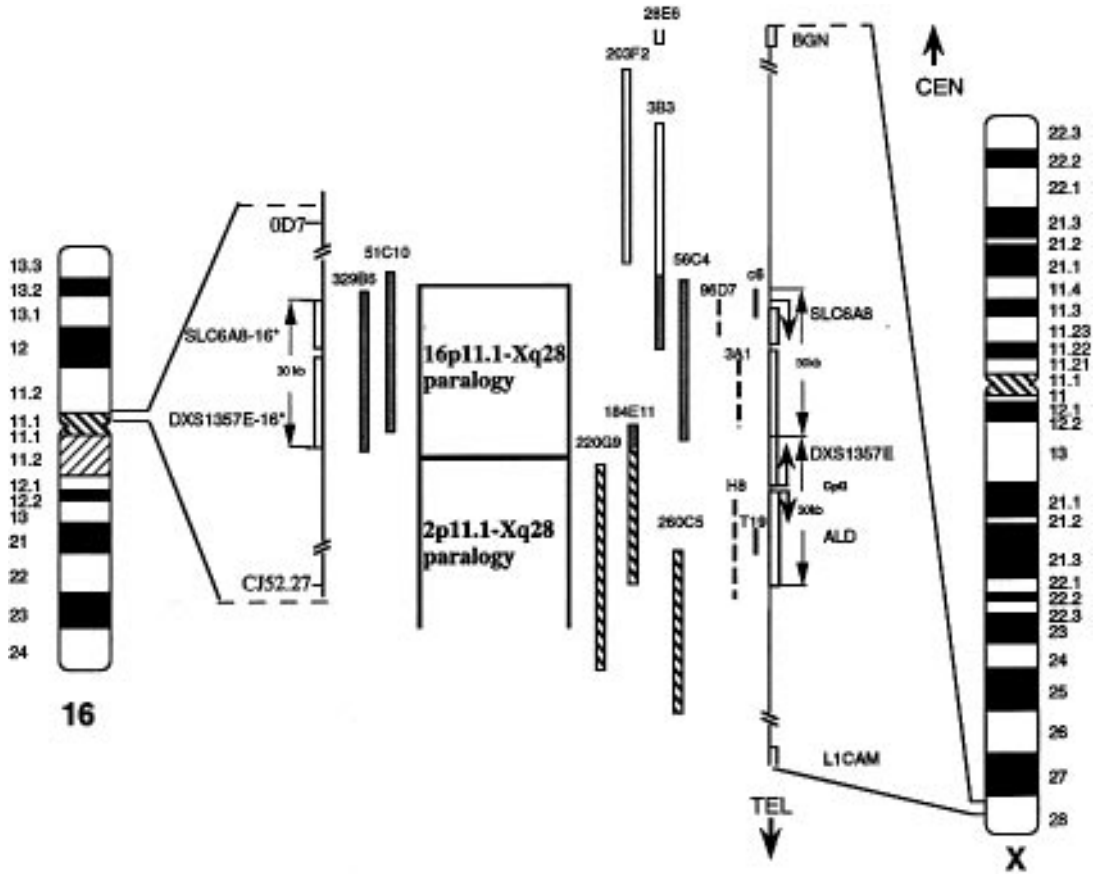


Figure 1. Extent of paralogy between Xq28 and 16p11.1. Depicted are ideograms of chromosomes 16 and X and the organization of cosmid, cDNA and genomic subclones in 16p11.1 and Xq28 regions. Cosmids are represented by vertical bars; cDNA clones are shown as dashed vertical lines, and genomic subclones are depicted as solid vertical lines. The extent of paralogy between 16p11.1 and Xq28 was initially estimated by FISH. Shaded vertical bars indicate 16p11.1-Xq28 crosshybridization, hatched bars indicate 2p11.1-Xq28 crosshybridization and open bars identify cosmids which hybridize only to cytogenetic band Xq28. The orientation of transcription of various genes in the Xq28 is shown by vertical fish hook arrows. Genomic distances are estimated based on*Pst*I fragment overlap between cosmids, hybridization patterns of various cDNA clones and previously reported distances for the ALD region (26). The centromere-telomere orientation of the 16p11.1 paralogous region is not known. *No transcripts corresponding to *SLC6A8* and *DXS1357E* have been identified.

2.5 kb cDNA fragment corresponding to exons 1–9 of*ALD*) were employed to facilitate in the identification and orientation of cosmids within and flanking this region (26). The position of these clones and a minimum tiling path of cosmids in the contig is summarized in Figure 1. Using representative cosmids of the region, FISH of human metaphase chromosome spreads was used to assay the extent of paralogy between X and 16 (Fig. 1). FISH analysis indicates that the duplicated region between X and 16 is relatively small, approximately 30 kb, and that it does not include the X chromosome *BGN* and *ALD* genes. Interestingly, a second domain of paralogy was identified distal to the X-16 duplicated region. Using intron/exon genomic clones (TA14) and *ALD* cDNA fragments as probes, FISH analysis indicates that a paralog of the adrenoleukodystrophy gene likely exists at cytogenetic band 2p11.1–2p11.2 (25,26) (Fig. 1).

The Xq28 CDM and creatine transporter loci are transcriptionally active

A shotgun M13 library was prepared from an Xq28 cosmid, u56c4, which cross-hybridized strongly to both 16p11.1 and Xq28 and which contained substantial portions of the putative *DXS1357E* and *SLC6A8* genes. M13 subclones were sequenced using dideoxy dye-primer cycle sequencing methods and assembled using XDAP software, reconstituting the 33 kb genomic insert of cosmid clone u56C4 (GenBank accession U36341). Comparative sequence analysis between u56C4 and previously published cDNA sequence of the *SLC6A8* (GenBank accession L31409) and the *DXS1357E* gene (GenBank Z31696) revealed greater than 99.8% nucleotide similarity, resulting in identical predicted amino-acid composition. These data indicate that the Xq28 genes likely represent the transcriptionally active loci for published *DXS1357E* and *SLC6A8* transcripts. The organization of the creatine transporter gene (4.3 kb transcript) (23) is compact. It consists of 13 exons spanning 8.5 kb of genomic DNA (Fig. 2 and Table 1). Alignment of the putative amino-acid sequence of the creatine transporter with the gene's physical structure reveals a modular exon organization, with one exon for every transmembrane domain over the first eight exons. In an area where compaction of intron/exon structure is greatest (exons 8–10; average intron size 50 bp), the modular organization breaks down with transmembrane protein domains spanning intron/exon boundaries (Fig. 2 and Table 1). The *DXS1357E* gene is located in close proximity (5 kb) to the *SLC6A8* gene and is

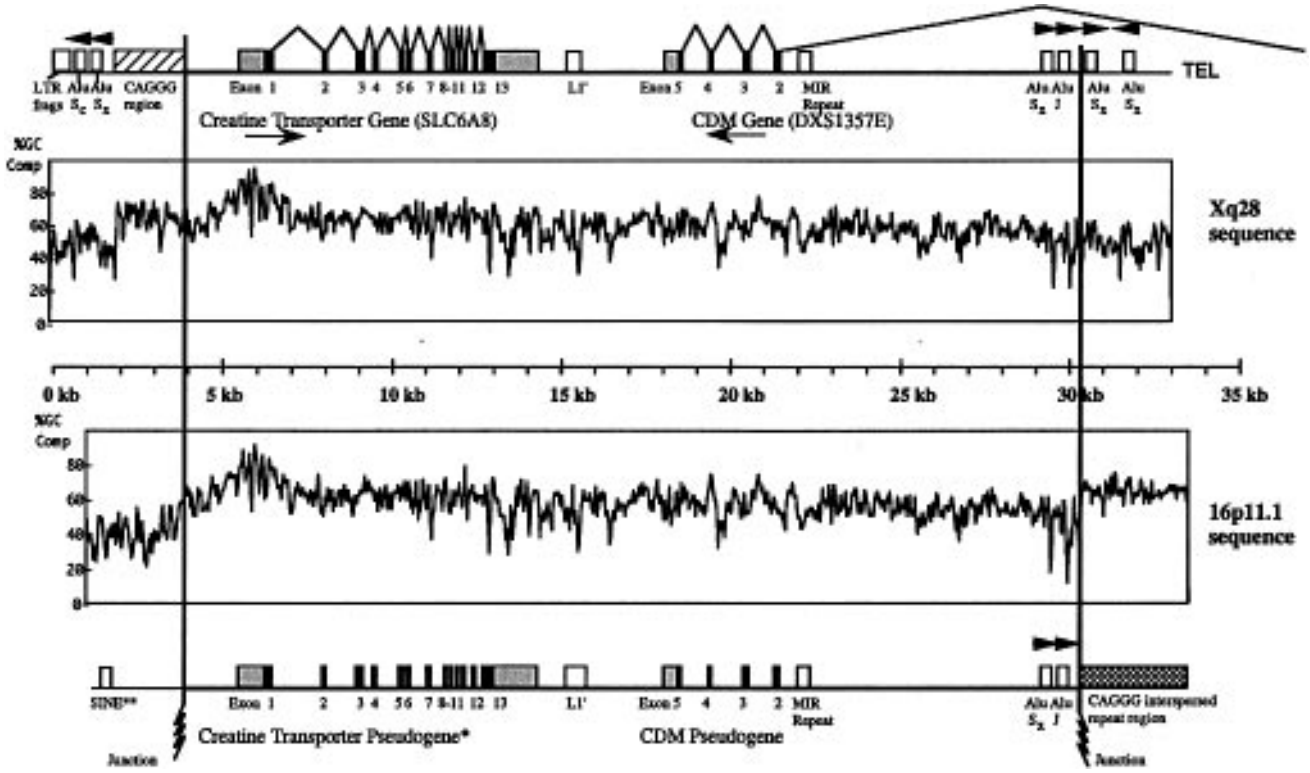


Figure 2. Sequence of cosmids c329B6 and u56C4. The complete sequence of cosmid u56C4 (33 023 bp insert; X chromosome) and c329B6 (32 505 bp insert; chromosome 16) is depicted. %GC content was calculated in 100 bp windows using GeneWorks software (v. 2.1.1). The physical structure of the creatine transporter gene (*SLC6A8*) and CDM (*DXS1357E*) genes are shown, and compared to cognate sequences on the 16p11.1 cosmid. Exons are depicted as filled boxes and UTR are symbolised as stippled boxes. The position and identity of repetitive elements was determined using the PYTHIA program (75) and are indicated as empty boxes in the sequence. Vertical bars define the boundaries of paralogy between 16p11.1 and Xq28. Note the duplication does not include the entire coding portion of the CDM gene. The position of CAGGG repeats at both the 5′ boundary of 56C4 and the 3′ boundary of 329b6 are depicted as hatched boxes.

Table 1. Intron/exon boundaries

Gene	Exon/Intron	Intron #	Intron length (bp)	Intron/Exon	Exon #
SLC6A8	CGGAG/gtgag	Intron 1	1539	ccag/GTGTG	2
	GAAAG/gtgag	Intron 2	799	cccag/GCCTG	3
	GGGAG/tgagc	Intron 3	422	cctag/GGACA	4
	GAAAG/gtaca	Intron 4	934	cccag/AGNTC	5
	CTCAG/gtgag	Intron 5	88	tctag/GTGTG	6
	ACAAG/taagc	Intron 6	96	ctcag/GGACG	7
	GTCAG/gtagg	Intron 7	319	cacag/GGCCG	8
	GCCAG/gtttg	Intron 8	139	cacag/TTTGT	9
	CTGAT/gttag	Intron 9	77	cccag/GGCGG	10
	TACGG/taggt	Intron 10	87	cgcag/GAGCT	11
	GCATG/gtaag	Intron 11	86	cgtag/GGCAT	12
	CTGAG/gtaag	Intron 12	185	tgcag/CGCTG	13
DXS1357E	GCAGC/ctgga	Intron 4*	1053	ctgac/CTGCA	4
	TAGTT/cttga	Intron 3	828	ctcac/TTTGC	3
	GCTCC/ctggg	Intron 2	904	ctcac/CTTCT	2
	GTCAC/caggc	Intron 1	>11 kb	?	1

Sequences at the intron/exon boundaries of the Xq28 *SLC6A8* and part of the *DXS1357E* gene are shown. The length of each intron is indicated, revealing a compact genomic organization of *SLC6A8*. *DXS1357E* exons are numbered as if the remaining 400 bp of the gene were contained in a single exon.

transcribed with opposite polarity. Five exons of the *DXS1357E* gene were sequenced corresponding to positions 474 to 1315 of the *DXS1357E* transcript (Fig. 2 and Table 1). Based on previous mapping and sequencing of the first exon of the *DXS1357E* gene, we estimate that the gene spans ~20 kb of sequence.

Comparative sequence analysis of the 16p11.1-Xq28 duplication

A second cosmid clone from the chromosome 16 library, c329B6, was sequenced in its entirety (32.5 kb) (GenBank accession #U41302) and compared to the paralogous sequence of u56C4 (GenBank accession #U36341). Bestfit alignment of the two sequences reveals that the duplication between 16p11.1 and Xq28 involves 26.5 kb of genomic DNA, including the entire creatine transporter (*SLC6A8*) and five exons of the *DXS1357E* gene (Fig. 2). A total of 76 insertions/deletions (indels) and 1122 transitions/transversions were observed within the 26.5 kb region of overlap, resulting in an overall similarity of 94.6% between the two paralogous sequences (Table 2). A significantly greater (p <0.0001; Z = 7.2) per cent similarity was observed for the putative coding portions of the 16-X paralogous genes when compared to the overall similarity of the sequences (94.6% vs. 97.1%, respectively) (Table 2). Since the chromosome 16 truncated *DXS1357E* gene was likely non-functional at the time of the duplication event between these two chromosomes, we compared the divergence rate between putative coding segments (CDS) of the *DXS1357E* and *SLC6A8* genes between X and 16. The divergence rate for *DXS1357E* CDS exons is significantly greater (p <0.05; Z = 1.68) than the divergence rate for *SLC6A8* coding exons. In addition, no indels were observed within the CDS of the chromosome 16 paralogous *SLC6A8* gene (Fig. 2). Predicted amino-acid composition of the 16-creatine transporter paralog,

Table 2. Per cent nucleotide similarity between various portions of the 16p11.1-Xq28 duplication

Sequence Type	# of single bp changes	# of indels	# of basepairs compared	% Similarity
Non-genic	952	68	21919	94.4
CDS (SLC6A8)	52	0	1777	97.1
CDS (DXS1357E)	14	1	379	95.3
5′ UTR (SLC6A8)	48	3	604	90.1
3′ UTR (SLC6A8)	41	2	1362	96.4
3′ UTR (DXS1357E)	15	0	435	96.6
Total	1122	74	26476	94.6

The overall % similarity for 26.5 kb of paralogous sequence was calculated as 94.6% (see Materials and Methods). % nucleotide similarity as well as the number of indels and single bp changes among various genic and non-genic portions are compared. The percent similarity of the coding exons of the 19p11.1-Xq28 *SLC6A8* gene is significantly greater than that of the compared CDS of the *DXS1357E* gene ($p<0.05$). The 5′ UTR of the *SLC6A8* gene show greater nucleotide divergence than non–genic DNA.

however, does indicate the presence of a 'premature' stop codon in exon 4 (Trp_{TGG}>>$Amber_{TGA}$). Sequence analysis of PCR amplified and subcloned products from three unrelated individuals confirmed the presence of this premature stop codon in the chromosome 16 cosmid.

Breakpoints of 16p11.1-Xq28 paralogy contain CAGGG interspersed repeats

Figure 3 depicts the sequence immediately flanking the boundaries of the 16p11.1 and Xq28 paralogy domain. CAGGG or inverted CCCTG repeats were located near or within all four breakpoints (Fig. 3a,b). The 3′ flanking region (relative to the transcriptional orientation of the Xq28 *SLC6A8* gene) of the 16p11.1 domain, in particular, shows a complex arrangement of CAGGG pentamer repeats. A total of 141 direct CAGGG interspersed repeats were observed in 3.1 kb flanking the 16p11.1 duplicated portion, resulting in a density of one repeat every 22.5 bp (Fig. 3b). Several higher-order repeat structures were also observed within the repeat region, all involving the CAGGG pentamer. For example, 19 CAGGG doublets $(CAGGG)_2$ as well as several hexamer doublets including $(CAGGGC)_2$, $(ACAGGG)_2$, $(GCAGGG)_2$ and $(CCAGGG)_2$ were found in this region. The most complicated higher-order repeat structure identified was a 38-mer which, once again, included two pentamer CAGGG motifs located at either end of the repeat (see Fig. 3a,c for consensus). CAGGG repeats were also found distributed through the 5′ flanking region of the Xq28 paralogy domain, albeit much less frequently (1 CAGGG/CCCTG repeat every 124 bp). In contrast to the 16p11.1 flank, many of the repeats are inverted (CCCTG). Hundreds of imperfect repeats (i.e. those which differ by one bp from the CAGGG/CCCTG consensus) were observed on both the 5′and 3′ flanks of the Xq28 and 16p11.1, respectively. With the exception of a few imperfect (CAGGG) repeats located near or within the breakpoints (Fig. 3a), the reciprocal flanks (the 5′ region flanking the 16p11.1 breakpoint and the 3′ region flanking the Xq28 boundary) are virtually devoid of these repeat structures. It may be noteworthy that the 3′ breakpoint of Xq28 occurs within the cluster of direct *Alu* repeats (Fig. 2).

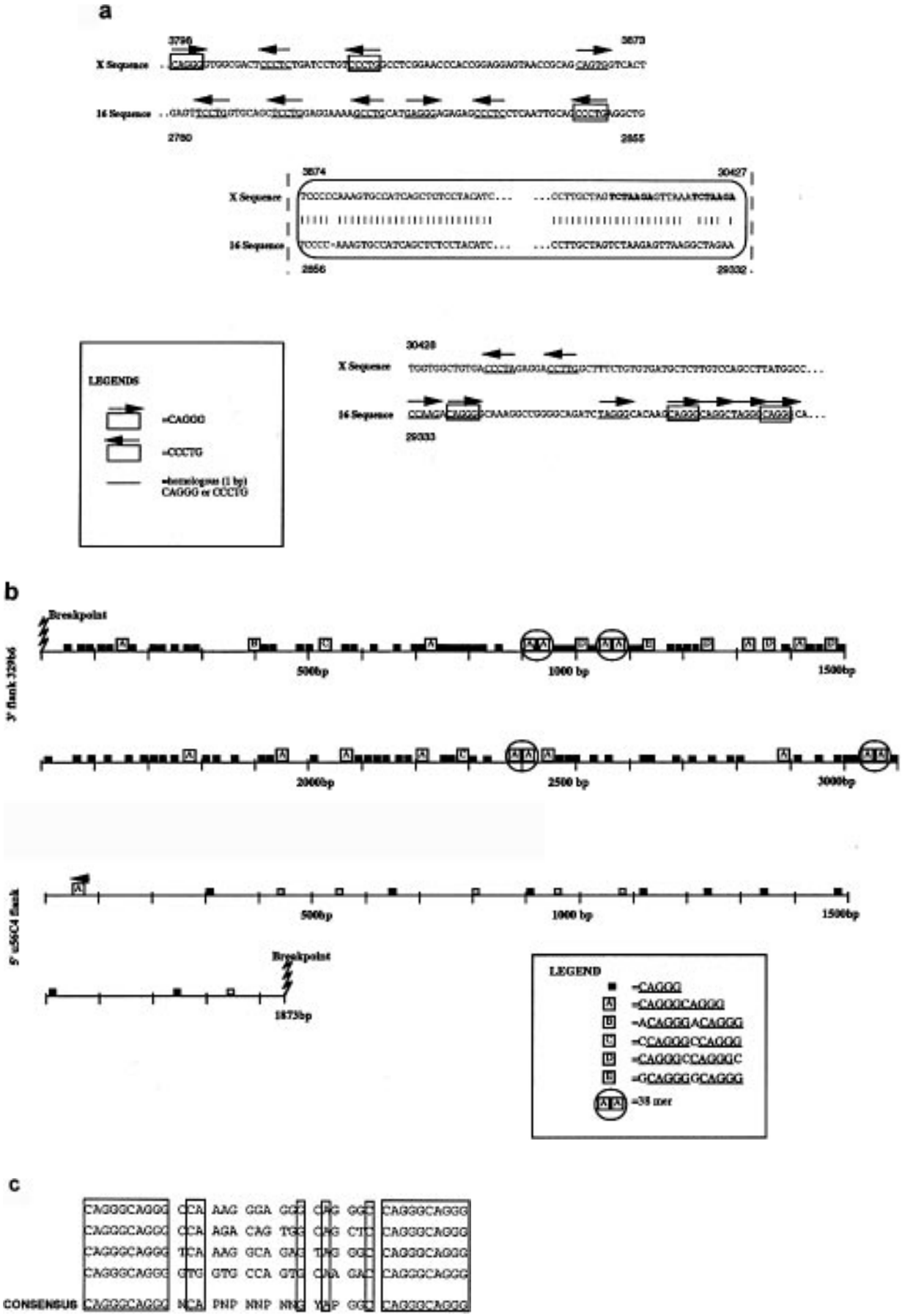
Table 3. Summary of comparative FISH analysis

Species	u56C4 (Xq28 cosmid)	329b6 (16p11.1 cosmid)
Homo sapiens	Xq28; 16p11.1(XVI)	Xq28; 16p11.1 (XVI)
Pan troglodytes	Xq telomeric; 19 pericentromeric (VII); 12 pericentromeric (IIp)	Xq telomeric; 19 pericentromeric (VII); 12 pericentromeric (IIp)
Gorilla gorilla	Xq telomeric; 6 pericentromeric (VII); 12 pericentromeric (IIp); 14 pericentromeric (XIII); 15 pericentromeric (XV); 16 pericentromeric (XVIII); 17 pericentromeric (XVI)	Xq telomeric; 6 pericentromeric (VII); 12 pericentromeric (IIp); 14 pericentromeric (XIII); 15 pericentromeric (XV); 16 pericentromeric (XVIII); 17 pericentromeric(XVI)
Pongo pygmaeus	Xq telomeric	Xq telomeric
Hylobates lar	Xq telomeric	Xq telomeric; 8 pericentromeric (XXIII); 17 pericentromeric (XIV); 15 pericentromeric (XV)

The FISH chromosomal assignments of the u56C4 and 329b6 probes for each species are summarized. Roman numberals indicate phylogenetic groups, which define syntenic chromosomal portions among these primate species. All non-Xq28 signals were located within the pericentromeric region of each chromosome.

Duplications are localized to the pericentromeric regions of primate chromosomes

In order to assess the origin of the duplication event, comparative FISH analysis using both u56C4 (chromosome X cosmid) and c329B6 (chromosome 16 cosmid) as probes was performed on metaphase chromosome spreads from various primate species. With the exception of gibbon, signals from both probes co-localized within each species (Table 3; Fig. 4). While the Xq28 signals were generally punctate, those generated on the other chromosomes were much more intense and broader in form among all species. In the chimpanzee, c329B6 and u56C4 cohybridized to the syntenic region of Xq28. In addition, signals were observed in the pericentromeric region of chimpanzee chromosomes 19 and 12, which are syntenic to human chromosomes 7 and 2p (Table 3). In the gorilla, the duplication event appears to have occurred multiple times, resulting in signals syntenic to human chromosomes X, 7, 2p, 13, 15, 16 and 18. With the exception of the Xq28 locus, all signals were observed within the first cytogenetic band from the centromere (Fig. 4). Southern analysis using p96D7 (*SLC6A8* cDNA fragment) and p3A1 (*DXS1357E* cDNA fragment) on genomic zooblots from various primates confirmed that the duplication in the gorilla and chimpanzee had included the creatine transporter and CDM genes, resulting in multiple cross-hybridizing bands and increased intensity of hybridization on some restriction fragments (data not shown). In contrast to human, gorilla and chimpanzee, FISH using c329B6 and u56C4 on orangutan metaphase spreads showed signals only on the distal end of X chromosome, suggesting that the Xq28 represents the ancestral template copy. Gibbon was the only species which demonstrated differential patterns of chromosome hybridization with the two probes. The c329B6 probe hybridized uniquely to gibbon pericentromeric regions chromosomes 8, 15 and 17 (syntenic to human chromosomes 22, 15 and 14). Both the u56C4 and the c329B6 probe, once again, hybridized to the distal end of the X chromosome.


a
X Sequence
16 Sequence
LEGENDS
=CAGGG
=CCCTG
=homologous (1 bp) CAGGG or CCCTG
b
Breakpoint
3' flank 329bd
5' α56C4 flank
500bp
1000 bp
1500bp
2000bp
2500 bp
3000bp
1875bp
LEGEND
=CAGGG
=CAGGGCAGGG
=ACAGGGACAGGG
=CCAGGGCCAGGG
=CAGGGCCAGGGC
=GCAGGGGCAGGG
=38 mer
c
CONSENSUS

We have surveyed by FISH distantly related groups of the human population (Mbuti Pygmy, Israeli Caucasian, Brazilian Karitiana, Melanesian Nasoi and Auca Indian) in order to assess potential polymorphic variability in the location of the duplication. All five individuals showed strong hybridization of the u56C4 probe to both 16p11.1 and Xq28. The c329B6 probe hybridized intensely to 16p11.1 and more weakly to Xq28 (data not shown). These results indicate that the 16p11.1 and Xq28 duplication has been fixed in the human population for at least 200 000 years. Interestingly, hybridization of c329B6 to 2p11.1 was observed in a few individuals (Caucasian, Karitiana and Pygmy).

Xq28 locus represents the ancestral template

The consistent cohybridization of 56c4 and c329B6 probes to the distal end of the X chromosome may be taken as strong evidence that the Xq28 region represents the ancestral locus of the CDM and creatine transporter genes. To determine if this region is syntenic in a more distantly related organism, a panel of murine YACs extending from the *Bgn* to *G6pd* locus (~1062 kb) (29) was hybridized with human cDNA probes p96D7 and p3A1. Both the creatine transporter and CDM probes hybridized to YACs *D19H6* and *C176B11*, placing both genes in the same 185 kb interval between *Bgn* and *L1cam*, at a location at least 100 kb distal to the mouse *Bgn* locus. The position of the genes, thus, is highly conserved between man and mouse on the X chromosome. Only bands cognate to the Xq28 fragments were observed upon hybridization of the creatine transporter cDNA (p96D7) to mouse genomic Southerns, indicating that other duplications of this region in the mouse are unlikely.

DISCUSSION

Origin of the 16p11.1-Xq28 paralogy

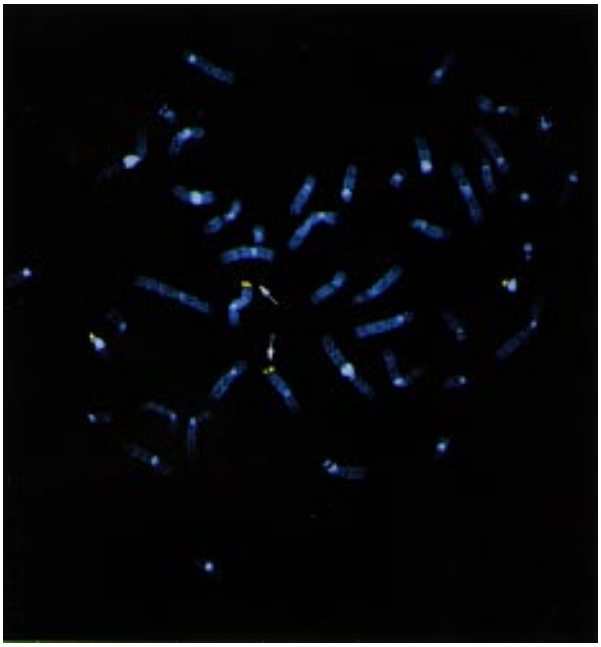
We have identified and characterized an evolutionarily recent and proliferative duplication event involving a gene-rich region of Xq28 and the pericentromeric regions of primate chromosomes. The overall similarity between the human 16p11.1 and Xq28 regions over 26.5 kb of paralogous sequence is 94.6% (Table 1). Based on estimates of mutation rate for silent site substitutions (30,31) and for intronic sequences (32) (5×10^{-9}–7×10^{-9} mutations per site per year), we calculate that the 16p11.1 and Xq28 may have diverged as recently as 7–10 mya. Most estimates of nucleotide divergence between ancestrally related sequences, however, have been determined by comparisons of orthologous sequences and the time of species separation predicted from paleontological lines of evidence (30–33). Among orthologous gene sequences (including introns and some intergenic segments) (33), selection pressure may be operative, limiting the rate of mutation. The recent transfer of genetic material to new chromosomal positions and the different (or lack of) selection pressure which it may experience is clearly distinct from orthologous sequence comparisons. Perhaps more appropriate measures of the rate of divergence within the 16p11.1-Xq28 paralogy domain can be made from comparisons of the β-globin gene and its paralogous pseudogenes. These estimates of divergence which range as high as 13×10^{-9} would predict that the 16p11.1 and Xq28 sequences may have separated as recently as 4 mya. Comparisons of paralogous gene segments located in tandem (globin gene cluster), once again, may not be entirely equivalent to interchromosomal parologies, due to the potential for sequence homogenization by unequal crossing over or gene conversion (11,12). The sequence characterization of a recent duplication of genes between Xq28 and 16p11.1, thus, should provide researchers with a unique opportunity to study the rates of divergence of DNA between non-homologous chromosomes within the same species.

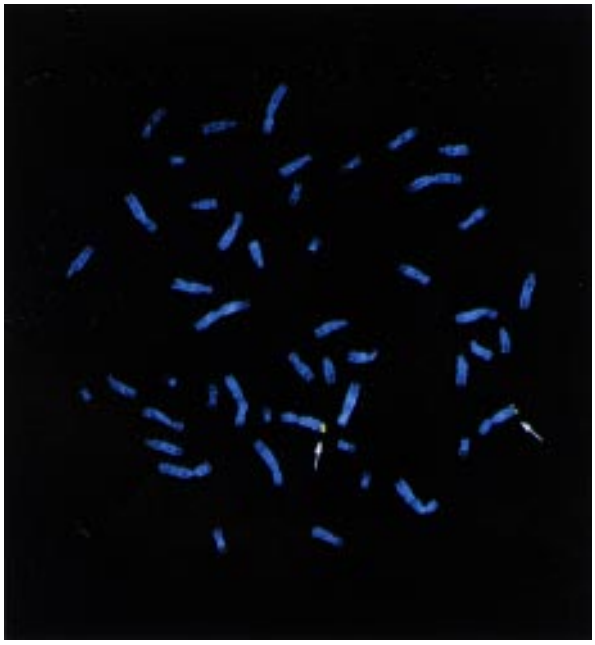
Functionality of the chromosome 16 paralogous gene segments

Interestingly, the rate of divergence is not uniform throughout the entire 16p11.1-Xq28 duplicated region (Table 2). The coding segments of 16p11.1 *SLC6A8* exons are more highly conserved than the intronic or intergenic sequence (97.1 % vs. 94.4%; Table 2). The predicted amino acid sequence of the 16p11.1 creatine transporter paralog, however, identifies a premature stop codon within exon 4 (compare position 8633 of GenBank accession #U41302 with position 1224 of paralogous Xq28 cDNA, GenBank accession L31409). This suggests that the 16p11.1 creatine transporter is non-functional or that it encodes a truncated four trans-membrane domain protein. RT–PCR assays specific for the chromosome 16 *SLC6A8* transcript have failed to identify expression from various tissue sources (brain, hippocampus, retina, liver, intestine, lymphoblastoid, fibroblast and melanoma; data not shown). Interestingly the normal stop codon of the Xq28 *SLC6A8* transcript is absent in the paralogous 16p11.1 sequence (TGA>>GGA) (compare position 12824 GenBank accession #U36341 with position 11753 GenBank accession #U41302).

In contrast to the *SLC6A8* gene, the 16p11.1 *DXS1357E* 'coding' sequences have diverged much more rapidly (95.1% similarity) and have incurred mutations which alter the putative translational reading frame of the protein (when compared to the Xq28 sequence). Since the paralogy domain breaks within an intron of the chromosome 16 *DXS1357E* sequence (Figs 2,3), effectively truncating the gene, it is likely that the 16p11.1 *DXS1357E* paralog was a non-functional pseudogene at the time of the duplication event. This may explain why the divergence of *DXS1357E* coding exons (4.9%) is greater than the coding exons of the *SLC6A8* gene (2.9%). *In toto*, our investigations suggest that although the *DXS1357E* gene has been a non-processed (promoter-less) pseudogene since the time of its divergence, the high conservation of the coding exons of the 16p11.1 *SLC6A8* suggest that it may have had a functional role at some time during hominoid evolution.

Figure 3. CAGGG repeats at the Xq28–16p11.1 breakpoints. (**a**) Approximately 70 bp of sequence flanking either side of the identified boundaries of the Xq28–16p11.1 duplication is shown. CAGGG and inverted CCCTG sequence motifs are indicated in boxes with arrows defining the orientation of the sequence relative to Xq28 CRT transcription. Pentamer sequences differing by a single bp are underlined. Vertical bars define the breakpoints. (**b**) The distribution of perfect CAGGG and CCCTG sequences flanking the 5′ boundary of Xq28 u56C4 (1.8 kb) and the 3′ boundary of 16p11.1 c329b6 (3.1 kb) is depicted. CAGGG pentamers occur at a density of once every 124 bp at the 5′ boundary of Xq28 and at a density of once every 22.5 bp at the 3′ boundary of 16p11.1. The occurrences of CAGGG doublets and other hexamer doublet derivatives are also indicated. (**c**) Four 48-mers were identified within the 16p11.1 3′ flank. The sequences were aligned using PILEUP (GCG software package) and a concensus was generated.





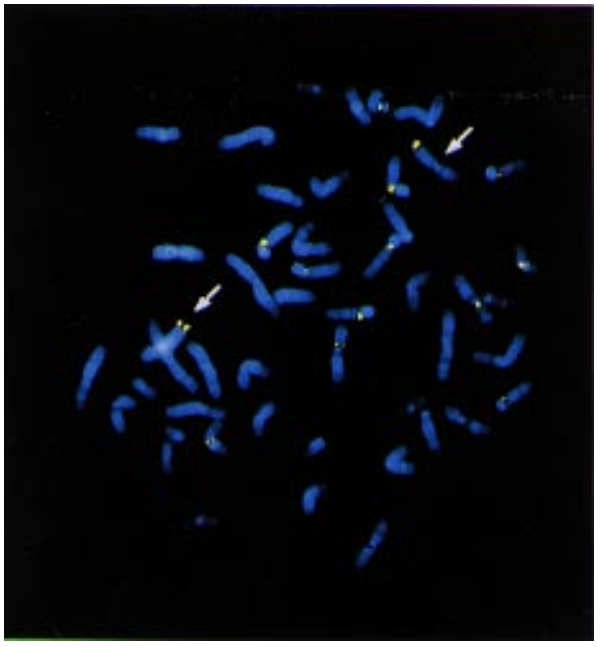


Figure 4. Comparative FISH analysis of primate chromosomes. Fluorescent *in situ* hybridization of cosmid u56C4 is shown for metaphase chromosomal preparations of (**a**) human, (**b**) gorilla (*G. gorilla*) and (**c**) orangutan (*P. pygmaeus*). Hybridization signal to the distal end of the X chromosome is indicated in each case by an arrow. The other chromosomal assignments are summarized in Table 3. Identical FISH localizations were obtained from the chromosome 16 cosmid probe, c329B6 (data not shown).

Pericentromeric-directed mechanism for the duplication of the Xq28 region

Comparative FISH analysis of primate chromosomes confirm that the duplication event is likely to have occurred after the separation of the orangutan from the human hominoid line of evolution (10–19 mya) but prior to the trichotomy of the human-gorilla-chimpanzee clade (4–5 mya). For example, in gibbon and orangutan, both 16p11.1 and Xq28 probes colocalize only to the distal end of the X chromosome (Table 3). In conjunction with mapping data in the mouse which places the murine creatine transporter locus in the *Bgn* and *L1cam* interval of the X chromosome, these data suggest that the Xq28 locus represents the ancestral template copy of the duplications. FISH of gorilla, human and chimpanzee chromosomes indicate that the duplication event occurred multiple times. In each species, signals are consistently observed on the distal end of the X chromosome and at the pericentromeric regions of various chromosomes (Table 3; Fig. 4). Surprisingly, none of the pericentromeric locations in the chimpanzee (phylogenetic groups IIp and VII) (34) are syntenic to human 16p11.1 (Table 3). Gorilla chromosome metaphase spreads, in contrast, show pericentromeric signals including synteny to human 16p11.1 (phylogenetic group XVI), chimpanzee syntenic groups, IIp and VII, as well as gorilla-specific pericentromeric duplications (Table 3). It is possible that the arrangement of duplications in Gorilla represents the ancestral state of the pongid/hominid evolutionary line. This would suggest that a burst of duplications occurred over a short period of time, transposing copies of the Xq28 CDM-creatine transporter sequence to various pericentromeric positions. Evolutionary lineages, such as those leading to humans and chimpanzees subsequently may have lost copies of the duplicated genes on various chromosomes leading to a restriction of paralogy to 16p11.1 in humans (XVI synteny group) and IIp and VII in chimpanzee (Table 3). Alternatively, duplication and mobilization of the Xq28 sequence may involve on-going replicative and non-replicative transposition events,

resulting in various pericentromeric locations (syntenic and non-syntenic) among the different species. Our survey of human diversity in which all different ethnic groups showed signals at 16p11.1 and Xq28 indicates that duplication and mobilization events of the Xq28 sequence have remained quiescent for the last 200 000 years of human evolution (35,36).

A second domain of paralogy was identified in our FISH analysis of the *BGN/ALD* interval of Xq28. Cosmids encompassing the 5′ portion of the *DXS1357E* gene as well as substantial portions of the *ALD* gene were shown to cross-hybridize to cytogenetic bands Xq28 and 2p11.1 (Fig. 1). Using probes derived exclusively from *ALD* cDNA, we demonstrated by FISH that it is likely that paralog(s) of the *ALD* gene exist near the pericentromeric region of the short arm of chromosome 2. These results confirm molecular analysis which predicted that an additional copy of the adrenoleukodystrophy gene existed in the human genome (26). The gene-cluster spanning from the creatine transporter gene to the adrenoleukodystrophy gene of Xq28, thus, appears to have been actively involved in pericentromeric duplications. It is intriguing that the 2p11.1 cytogenetic band location of the *ALD* paralogy domain coincides with the syntenic pericentromeric location of *SLC6A8/DXS1357E* cross-hybridization observed in chimpanzee and gorilla (phylogenetic group IIp). One possible explanation is that the entire Xq28 creatine transporter/adrenoleukodystrophy region was duplicated by mechanisms such as replicative transposition to the short arm of chromosome 2 in the early ancestor of pongids and hominids. In the human lineage, a subsequent round of non-replicative transposition may have mobilized and directed the creatine transporter-CDM portion to 16p11.1, leaving the *ALD* copy at 2p11.1. Further characterization of the 2p11.1 paralogy domain of humans and primates will be necessary to critically evaluate this model.

16p11.1–16p11.2: a hotspot for interchromosomal duplications

Two other interchromosomal duplications have previously been described involving the 16p11.1–16p11.2 cytogenetic interval; namely, a duplication of the V_H segments of the immunoglobulin heavy chain locus from cytogenetic band 14q32.3 to 16p11.2 (37) and the identification of a >15 kb paralogy domain between 6p25-pter and 16p11.1 involving the minisatellite locus, λMS29 (38). There are striking similarities between these duplications and the Xq28–16p11.1 paralogy domain. All three examples involve the mobilization of a substantial portion of genomic sequence (15–30 kb) relatively recently (1–10 mya) to the 16p11.1–16p11.2 cytogenetic interval. In each case, the transposition has involved the exchange of genetic material between subtelomeric and pericentromeric loci. In addition, the subtelomeric paralogs (V_H segments on 14q32.3, *DNF21S1* on 6p25-pter; *SLC6A8/DXS1357E* region on Xq28) have been deemed antecedent (37,38). Interestingly, when such events have involved the transposition of genes, the 16p11.1–p11.2 paralogs have generally been found to be transcriptionally silent (37). These data, as well as the general intrachromosomal instability of this region (39,40), suggest that the 16p11.1–11.2 region may be particularly prone to undergo macromutational events resulting in local chromosomal duplications.

Junctions of the 16p11.1-Xq28 paralogy domain suggest that duplications may be mediated by recombination

We have identified and characterized the sequence at the junctions of the 16p11.1 and Xq28 paralogy domain. The 3′ junction (centromere to telomere orientation) of the Xq28 region is located within a cluster of three head-to-tail *Alu* repeats (Figs 2 and 3a). A second cluster of *Alu* repeats was observed 2 kb proximal to the 5′ junction of the Xq28 region arranged in opposite orientation to the *Alu* breakpoint cluster (Fig. 2). No other *Alu* repeats were observed in nearly 28 kb of genomic distance between these two clusters, suggesting a non-random distribution of *Alu* repeats in the creatine transporter-CDM region (41). Rearrangement of *Alu* clusters, resulting in the formation of local duplications and deletions, has been implicated in the etiology of various human genetic diseases (42–45). Although the mechanism by which these events occurs is largely unknown (43), it has been proposed that the sequence homology between *Alu* repeats facilitates illegitimate recombination events, resulting in disruption of gene structure and concomitant gene dysfunction (46). The inverted clusters of *Alu* repeats at either end of the Xq28 breakpoints may, similarly, provide the sequence homology required to bring the Xq28 breakpoints in close proximity (Fig. 5) for intrachromosomal recombination to occur.

Within 15 bp of all four breakpoints of the 16p11.1-Xq28 paralogy domain, we observed the presence of both perfect and imperfect CAGGG (or inverted CCCTG) sequence motifs (Fig. 3a). These motifs are particularly enriched in the region proximal to the 5′ junction of the Xq28 sequence. CAGGG or CCCTG motifs occur with a density of once every 124 bp (Fig. 3b) throughout the 2 kb region between the 5′ *Alu* cluster and the Xq28 5′ paralogy boundary (Fig. 3). A much more dense (1 CAGGG every 22.5 bp) and complicated organization of CAGGG sequences (Fig. 3) is observed at the 3′ boundary of the 16p11.1 paralogy region. Many higher order structures including pentamer and hexamer doublets of the type $(CAGGG)_2$ $(CAGGGC)_2$, $(ACAGGG)_2$, etc. were observed in this region. In contrast to the 5′ boundary region of Xq28, no inverted CCCTG motifs were found distal to the 16p11.1 3′ boundary (Fig. 3a,b). The CAGGG sequence motifs present at or near the 6p11.1-Xq28 paralogy junctions bear a striking resemblance to interspersed and direct CTGGG repeats found in association with many of the immunoglobulin heavy-chain switch recombination regions (47–50). Switch recombination regions are responsible for the deletion of C_H genes of immunoglobulins vis-à-vis deletion circle intermediates (51,52) resulting in the placement of heavy chain VDJ segments in proximity to ‘new’ constant regions (48). These recombination-mediated events are responsible for increasing the diversity of immunoglobulin genes in B-lymphocytes by altering the antigenic clearance properties of the antibody. In addition to their similarity to recombination switch pentamers, there are several other lines of evidence which suggest that CAGGG motifs at the 16p11.1-Xq28 region may be hotspots for recombination: CAGGG motifs are similar to the prokaryotic chi element (GCTGGTGG) and minisatellite core consensus sequences which strongly promote recombination and gene conversion (53,54); CAGGAGG sequences are frequently found at the translocation breakpoints of *bcl2* and *c-myc* genes (55); $(CAGGG)_n$ and $(CAGG)_n$ repeats at mouse loci *Ms6-hm* and *Hm-2*, respectively are extremely variable exhibiting high

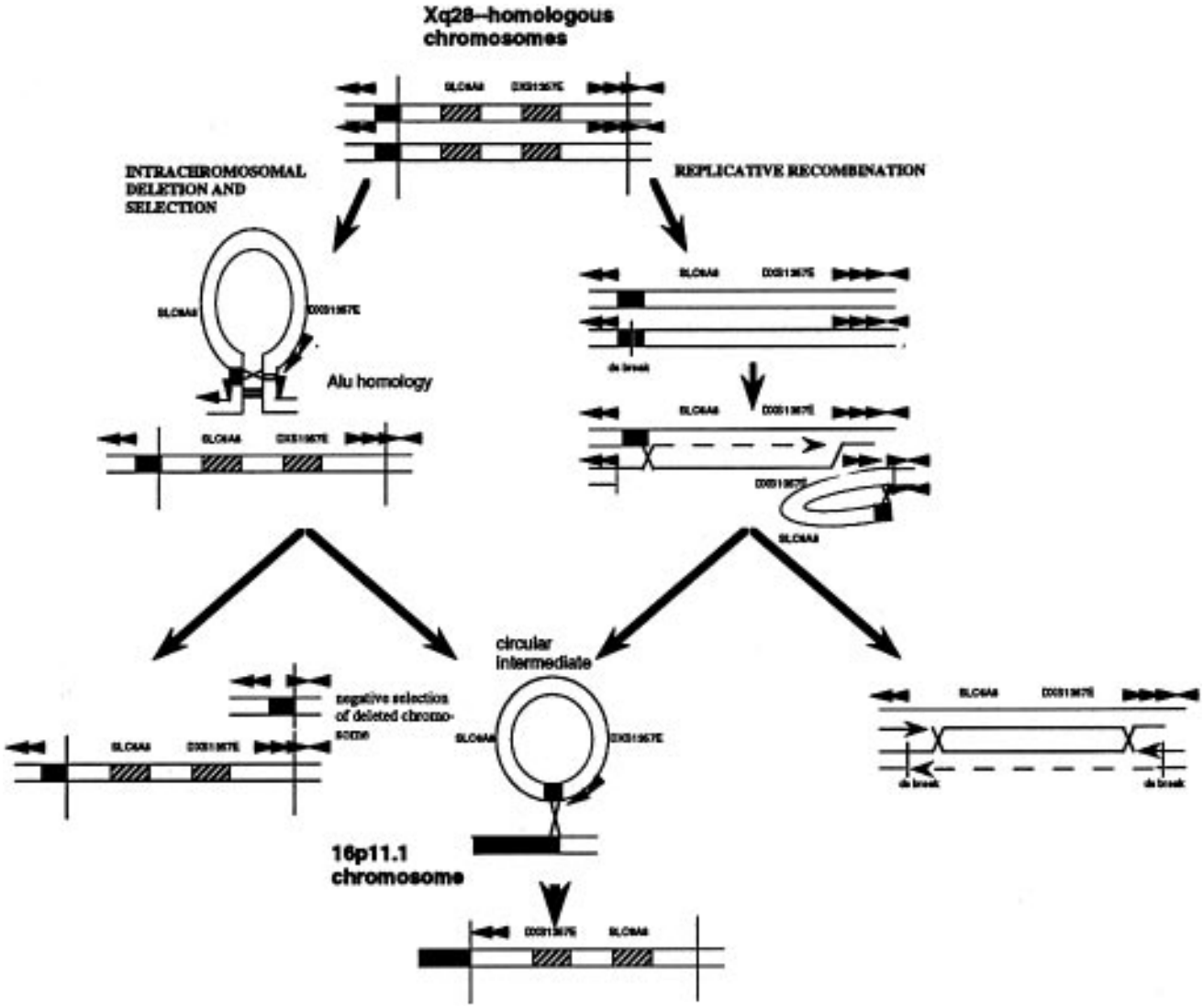


Figure 5. Possible mechanisms for the duplication of the Xq28 CDM-creatine transporter sequence to the pericentromeric regions of chromosomes. Two possible models for the Xq28-pericentromeric duplication are proposed. Both mechanisms assume that CAGGG pentamer repeats are hyper-recombinogenic, that the Xq28 CTR-CDM region represents the ancestral copy and that duplication involves the formation of a circular episomal intermediates (see Discussion). The position and orientation of *Alu* repeats in the Xq28 region are indicated by horizontal arrows, the CDM (*DXS1357E*) and creatine transporter genes (*SLC6A8*) are symbolized as crosshatched boxes and the CAGGG pentamer repeat regions are depicted as filled boxes. The first model invokes the formation of a large intrachromosomal hairpin loop, facilitated by the alignment of inverted *Alu* repeats, and deletion of the Xq28 CDM-creatine transporter region. The episomal intermediate integrates vis-à-vis CAGGG repeats present at 16p11.1 and other autosomes. Segregation of the deleted chromosome and selection removes the chromosome from subsequent generations (i.e. only the non-deleted X chromosome survives germline transmission). The second model is similar with the exception that the Xq28 region is not lost, but is duplicated by gene conversion during replication. The displaced strand instead of being degraded forms a circular intermediate by recombination at the CAGGG repeats. In this model, breakage within the *Alu* cluster is fortuitous, possibly due to weak homology of CAGGG repeats (see Fig. 3).

rates of somatic and germline mutations (56,57); CAGGGCAGGG doublets are reminiscent of G4 DNA which form unique interstrand and intra-strand DNA structures which are believed to promote meiotic pairing and exchange (58–60). Overall, these findings and the observation that CAGGG motifs are located near or at all four breakpoints, strongly argue that the recombinogenic properties of these sequences may have been responsible for the duplication and/or the integration of sequences near the pericentromeric regions of chromosomes. It is noteworthy that comparative FISH analysis of gibbon has shown that sequences other than the *SLC6A8*/*DXS1357E* paralogy are responsible for hybridization of the pericentromeric regions of chromosomes 4 and 7 in this species (Table 3). The CAGGG repetitive structure (of the chromosome 16 probe) may be responsible for this hybridization pattern, suggesting that these structures may exist at the pericentromeres prior to the integration of *SLC6A8*/*DXS1357E* paralogous sequence.

Two hypothetical models are proposed to account for the transposition of creatine transporter-CDM Xq28 sequence to 16p11.1 pericentromeric region (Fig. 5). Both models assume that CAGGG motifs are highly recombinogenic and that a small (~27 kb) episomal intermediate acts as the carrier of the sequence. The formation of such circular structures is attractive for two reasons. First of all the CAGGG motifs bear a striking resemblance to

CTGGG pentamers found in association with switch recombination. Switch recombination elements have been shown to promote the formation of circular deletion products with many of the recombinant breakpoints occurring near the CTGGG motif (48,51,52). Secondly, the integration of episomal intermediate (possibly through CAGGG homology) would account for the inverted organization of the 16p11.1 paralog relative to the CAGGG superstructure on chromosome 16. An alternate model, not presented, may involve trans-chromosomal exchanges between 16p11.1 and Xq28, possibly mediated by non-homologous pairing during meiosis. Such events, however, would be expected to preferentially result in translocations. Although no Xq28–16p11.1 translocations have been reported (possibly due to non-viability), translocations between 16p11.1–2 and 2p11.1–2 (the second domain of paralogy identified in this study) have been observed in cases of spontaneous childhood acute lymphoblastic leukemia (61).

Summary

We have described a novel interchromosomal paralogy involving 26.5 kb of gene-rich sequence between Xq28 and 16p11.1. Our data indicate that this paralogy has arisen quite recently in human evolution (7–10 mya) and have likely involved duplications of the Xq28 'master' sequence to the pericentromeric regions of chromosomes, possibly through recombinogenesis of CAGGG sequence motifs. These data suggest that processes other than tetraploidization (followed by reestablishment of the disomic state) may be involved in the expansion of gene families in the primate genome. The rapid proliferation of other mammalian 'cluster' multigene families which are dispersed near the telomeres and centromeres of chromosomes such as the ZNF (zinc finger) gene family (7) and the olfactory receptor cluster (62) may be mediated by similar bursts of transposition of short stretches of genomic DNA. Such genome plasticity would facilitate gene diversity, dramatically increasing the adaptive potential of an organism. The proliferation of the mammalian Na+ Cl– dependent transporter gene family, of which *SLC6A8* is a member, may be regarded in this light. The conservation of gene structure and their diverse function of this 12 membrane spanning domain gene family in amino-acid, neurotoxin, neurotransmitter, osmolyte and ion transport, may have involved duplication and dispersal from an ancestral template copy. The pericentromeric-directed transposition of the creatine transporters may provide a clue to the mechanism mediating such evolutionary diversity in vertebrate genomes.

MATERIALS AND METHODS

Library hybridization

An arrayed X-chromosome library, (LLOXNCO1 'U'), was obtained from Lawrence Livermore National Laboratory consisting of 25 000 clones. Clones were grown in a 4 × 4 array format on 16 nylon membrane filters at a density of 1536 clones per filter. Filters were prehybridized for 1 h at 65°C with 0.25 M $NaPO_4$, 0.25 M NaCl, 5% SDS, 10% PEG and 1 mM EDTA; and blocked with 20 µg/ml herring sperm DNA. Random hexamer-generated and end-labelled probes were prepared using manufacturer's specifications (Pharmacia; NEB) and hybridizations were performed as described. End-labelled oligonucleotide and random hexamer-generated probes were purified through G-25 and G-50 Sephadex columns, respectively. All hybridizations were performed overnight at 65°C in a rotisserie oven. Filters were washed three times for 30 min each at 65°C with 0.05 M $NaPO_4$, 0.5% SDS and 1 mM EDTA solution and exposed to autoradiographic film. A second flow-sorted chromosome 16-derived cosmid library (LA16NC02) was obtained from the Los Alamos National Laboratory (63). Three filters consisting of 4032 arrayed clones which had been pre-selected during the course of developing a physical map of chromosome 16 were screened with both cDNA and genomic clones. Total cosmid DNA was used as probe to screen an array of PCR-amplified clonal inserts from human placental and heart cDNA libraries in order to isolate corresponding cDNA clones (64).

Physical mapping of cosmid clones

Individual cosmid clones were labelled and hybridized to Southern blots of chromosome 16 and X chromosome somatic cell hybrid deletion panels as described previously (28,65,66). In order to further refine the map location of chromosome 16 cosmid, Southern blots of restriction digest panels of YAC clones which had been STS mapped to the 16p11.1–16p11.2 interval were probed. Due to the reported proximity of the Xq28 CDM gene to the *ALD* (adrenoleukodystrophy) gene (26), corresponding *ALD* cDNA (H8, 2.6 kb *Eco*R1 insert; exons 1–9) and genomic (TA19; 3.8 kb *Taq*I subclone of cosmid Qc11H12; exons 6 to 7) clones, as well as cDNA clones p96D7 and p3A1, were employed to confirm the *ALD-BGN* (biglycan) map interval location and to extend the cosmid physical map in this region (Fig. 1). A Southern blot of *Eco*R1 restriction digests of five murine YAC clones (D19H6, C176B11, D741C3, B7S6, and H864F2) (29) was probed with p96D7 and p3A1 to confirm the comparative map location of the *Cdm* and creatine transporter genes to the *Bgn*/*L1cam* interval of mouse.

Library construction

Shotgun libraries from two cosmids u56C4 (a 33.023 kb X chromosome insert) and c329B6 (32.505 kb chromosome 16 insert) were prepared in the M13delta vector (67) using previously described protocols (68). Libraries were constructed using an adaptor-based approach that increases the number of random clones by driving the blunt-ended ligation step with the addition of excess adaptor (67). The average insert size of M13 clones was approximately 1.5 kb. M13 clones were picked to inoculate 5 ml cultures and allowed to grow for no more than 7 h with vigorous shaking and aeration. Single stranded DNA was prepared using a PEG precipitation followed by GFC filter capture (69). The quantity and quality of DNA isolated was confirmed by gel electrophoresis on 1% agarose with ethidium staining.

Sequence assembly

Dye primer sequencing was performed using a Biomek 1000 workstation (Beckman Instruments Inc.), PE 9600 thermocycler (Perkin Elmer Cetus) and reagents provided by Applied Biosystems (70). A modified asymmetric PCR protocol (71,72) was used in the directed reverse sequencing phase of the project. Sequencing reaction products were analysed on a ABI 373A DNA Sequencer (Applied Biosystems). The collected data was trimmed using Seqprep software (MBCR, Department of Cell

Biology, BCM) and transferred to a UNIX platform for assembly and editing using XDAP software (73). Sequence map gaps (SMGs) and the known exon structure from 56C4 (Table 1) assisted in confirming assembly of the sequence reads from cosmid c329B6.

Sequence analysis

A GeneWorks software package (v.2.1, Intelligenetics) was used to analyse the GC composition and to generate a detailed restriction map of the XDAP consensus assembly of both cosmids. The locations of putative exons in the sequence was determined with the GRAIL 2 gene-recognition tool (74). A comparison of genomic (GenBank accession U36341 and U41302) and cDNA sequences (GenBank accession L31409 and GenBank Z31696) identified the position and size of *DXS1357E* and *SLC6A8* exons (23,25). The location and identity of *Alu* repetitive elements was determined using the PYTHIA search algorithm (75). Internal repetitive sequence motifs were identified and characterized using the PRINTREPEATS program. BestFit alignment (GCG Software package) was used to compare the 26.5 kb of paralogous sequence between the two cosmids.

In situ hybridization

In order to assess potential human diversity of the 16-X duplication, bicolor FISH was performed on metaphase chromosomal preparations from five lymphoblastoid cell lines from five diverse human populations: GM10493 (African Mbuti pygmy), GM11523 (Israeli Jew from Galilee), GM10965 (Brazilian Karitiana), GM10540 (Nasoi-speaking Melanesian) and GM11776 (Auca Indian tribe Waorani). In addition, chromosome metaphase spreads were prepared from human peripheral blood lymphocytes of a female donor or from lymphoblastoid cell lines derived from various species [chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*) and gibbon (*Hylobates lar*)]. Cosmid probes were labelled by nick-translation and hybridized to chromosomal preparations as previously described (76). High copy repeats were suppressed by reannealing the probe with 20-fold excess of human Cot 1 DNA (Life Technologies) prior to hybridization. Bicolor FISH was performed in which the X chromosome cosmid, u56C4, was biotin-labelled and detected using fluorescein isothiocyanate (FITC)-conjugated avidin (5 µg/ml) (Vector laboratories); and the chromosome 16 cosmid, c329B6, was digoxigenin-labelled and detected using a rhodamine-conjugated anti-digoxigenin antibody (Boehringer Mannheim). 4'6-diamidino-2-phenyllindole (DAPI) staining was simultaneously performed to confirm chromosomal cytogenetic band location.

ACKNOWLEDGEMENTS

We would like to thank C.-O. Sarde, J.-L. Mandel, C. C. Lee, A. Chaterjee and G. Herman for providing many of the Xq28 reagents from the *ALD-BGN* interval. We are grateful to N. Archidiacono, E. Lindsay and D. Muzny for excellent technical assistance. The chromosome specific gene libraries LLOXNC01 and LA16NC02 used in this work were constructed at the Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore CA 94550 and the Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545; respectively, under the auspices of the National Laboratory Gene Library Project sponsored by the US Department of Energy. This research was supported by US Dept. of Energy grant (94ER61830) to DLN, and by grants from NIH/NCHGR (HG00210 to DLN and HG00823 to RAG).

REFERENCES

1. Melnick, L. and Sherman, F. (1993) The gene clusters ARC and COR on chromosomes 5 and 10, respectively, of *Saccharomyces cerevisiae* share a common ancestry. *J. Mol. Biol.* **233**, 327–388.
2. Hankeln, T. and Schmidt, E. (1993) Divergent evolution of an 'orphan' histone gene cluster in *Chironomus*. *J. Mol. Biol.* **234**, 1301–7.
3. Arnold, N., Wienberg, J., Emert, K. and Zachau, H. (1995) Comparative mapping of DNA probes from the Vk immunoglobulin gene regions on human and great ape chromosomes by fluorescence in situ hybridization. *Genomics* **26**, 147–56.
4. Garcia-Fernandez, J. and Holland, P. (1994) Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370**, 563–6.
5. Teglund, S., Olsen, A., Khan, W., Frangsmyr, L. and Hammarstrom, S. (1994) The pregnancy-specific glycoprotein (PSG) gene cluster on human chromosome 19: fine structure of the 11 PSG genes and identification of 6 new genes forming a third subgroup within the carcinoembryonic antigen (CEA) family. *Genomics* **23**, 669–84.
6. Ben-Arie, N., Lancet, D., Taylor, C., Khen, M., Walker, N., Ledbetter, D., Carrozzo, R., Patel, K., Sheer, D., Lehrach, H. and North, M. (1993) Olfactory receptor gene cluster on human chromosome 17: Possible duplication of an ancestral repertoire. *Hum. Mol. Genet.* **3**, 229–35.
7. Bellefroid, E., Marine, J.-C., Ried, T., Lecocq, P., Riviere, M., Amemiya, C., Poncelet, D., Coulie, P., deJong, P., Szpirer, C., Ward, D. and Martial, J. (1993) Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymophoid cells. *EMBO J.* **12**, 1363–74.
8. Muscatelli, F., Walker, A., De Plaen, E., Stafford, A. and Monaco, A. (1995) Isolation and characterization of a MAGE gene family in the Xp21.3 region. *Proc. Natl Acad. Sci. USA* **92**, 4987–91.
9. McGinnis, W. and Krumlauf, R. (1992) Homeobox genes and axial patterning. *Cell* **68**, 283–302.
10. Kappen, C., Schughart, K. and Ruddle, F. (1989) Organization and expression of homeobox genes in mouse and man. *Ann. NY Acad. Sci.* **567**, 243–52.
11. Dover, G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–17.
12. Dover, G. (1988) DNA turnover and the molecular clock. *J. Mol. Evol.* **26**, 47–58.
13. Bird, A. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.* **11**, 14–21.
14. Chan, S., Cao, Q. and Steiner, D. (1990) Evolution of the insulin superfamily: cloning of a hybrid insulin/insulin-like growth factor cDNA from amphioxus. *Proc. Natl Acad. Sci.* **87**, 9319–823.
15. Ohno, S., Wolf, U. and Atkin, N. (1968) Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169–87.
16. Ohno, S. (1970) *Evolution by Gene Duplication*. Springer Verlag: Berlin,Heidelberg,New York.
17. Sidow, A. (1992) Diversification of the wnt gene family on the ancestral lineage of vertebrates. *Proc. Natl Acad. Sci. USA* **89**, 5098–102.
18. Lundin, L. (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16**, 1–19.
19. Bailey, G., Poulter, R. and Stockwell, P. (1978) Gene duplication in tetraploid fish: Model for gene silencing at unlinked duplicated loci. *Proc. Natl Acad. Sci. USA* **75**, 5575–9.
20. Ferris, S. and Whitt, G. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.* **12**, 267–317.
21. Jeffreys, A., Wilson, V., Blanchetot, A., Weller, P., Geurts van Kessel, A., Spurr, N., Solomon, E. and Goodfellow, P. (1984) The human myoglobin gene: a third dispersed globin locus in the human genome. *Nucleic Acids Res.* **12**, 3235–43.
22. Grzeschik, K. and Kazazian, H. (1985) Report of the committee on the genetic constitution of chromosomes 10, 11 and 12. *Cytogenet. Cell Genet.* **40**, 179–205.
23. Nash, S., Giros, B., Kingsmore, S., Rochelle, J., Suter, S., Gregor, P., Seldin, M. and Caron, M. (1994) Cloning, pharmacological characterization and genomic localization of the human creatine transporter. *Receptors and Channels* **2**, 164–174.

24. Sora, I., Richman, J., Santoro, G., Wei, H., Wang, Y., Vanderah, T., Horvath, R., Nguyen, M., Waite, S., Roeske, W. and Yamamura, H. (1994) The cloning and expression of a human creatine transporter. *Biochem. Biophys. Res. Comm.* **204**, 419–27.
25. Mosser, J., Sarde, C.-O., Vicaire, S., Yates, J. and Mandel, J.-L. (1994) A new human gene (DXS1357E) with ubiquitous expression, located in Xq28 adjacent to the adrenoleukodystrophy gene. *Genomics* **22**, 469–71.
26. Sarde, C.-O., Mosser, J., Koschis, P., Kretz, C., Vicaire, S., Aubourg, P., Poustka, A. and Mandel, J.-L. (1994) Genomic organization of the adrenoleukodystrophy gene. *Genomics* **22**, 13–20.
27. Nelson, D., Ballabio, A., Cremers, F., Monaco, A. and Schlessinger, D. (1995) In Nelson, D., Ballabio, A., Cremers, F., Monaco, A. and Schlessinger, D., Banff, A. B. (eds), *Report of the sixth international workshop on X chromosome mapping 1995.*
28. Stallings, R., Doggett, N., Callen, D., Apostolou, S., Chen, L., Nancarrow, J., Whitmore, S., Harris, P., Michison, H., Brauning, M., Sarich, J., Fickett, J., Cintosky, M., Sorenson, D., Torney, D., Hildebrand, C. and Moyzis, R. (1992) Evaluation of a cosmid contig physical map of human chromosome 16. *Genomics* **13**, 1031–9.
29. Chatterjee, A., Faust, C., Molinari-Storey, L., Kiochis, P., Poustka, A. and Herman, G. (1994) A 2.3-Mb yeast artificial chromosome contig spanning from Gabra3 to G5pd on the mouse X chromosome. *Genomics* **21**, 49–57.
30. Hayashida, H. and Miyata, T. (1983) Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatability complex. *Proc. Natl Acad. Sci. USA* **80,** 2671–2675.
31. Efstratiadis, A., Posakony, J., Maniatis, T., Lawn, R., O'Connell, C., Spritz, R., DeRiel, J., Forget, B., Weissman, S., Slightom, J., Blechl, A., Smithies, O., Bralle, F., Shoulders, C. and Proudfoot, N. (1980) The structure and evolution of the Human beta-globin gene family. *Cell* **21**, 653–68.
32. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) The evolution of genes: the chicken preproinsulin gene. *Cell* **20**, 555–65.
33. Margot, J. B., Demers, G. W. and Hardison, R. C. (1989) Complete nucleotide sequence of the rabbit beta-like globin gene cluster. *J. Mol. Biol.* **205**, 15–40.
34. ISCN (1985) Report of the standing committee on human cytogenetic nomenclature. *Birth Defects* **21**, 1–117.
35. Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. and Takahata, N. (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl Acad. Sci. USA* **92**, 532–6.
36. Bowcock, A. M., Ruiz-Linares, A., Minch, E., Kidd, J. R. and Cavalli-Sforza, L. L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–7.
37. Tomlinson, I., Cook, G., Carter, N., Elaswarapu, R., Smith, S., Walter, G., Buluwela, L., Rabbitts, T. and Winter, G. (1994) Human immunglobulin VH and D segments on chromosomes 15q11.2 and 16p11.2. *Hum. Mol. Genet.* **3**, 853–60.
38. Wong, Z., Royle, N. and Jeffreys, A. (1990) A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics* **7**, 222–234.
39. Dauwerse, J., Jumelet, E., Wessels, J., Saris, J., Hagemeijer, A., Beverstock, G., van Ommen, G. and Breuning, M. (1992) Extensive cross-homology between the long and the short arm of chromosome16 may explain leukemic inversions and translocations. *Blood* **79**, 1299–304.
40. Stallings, R., Doggett, N., Okumura, K. and Ward, D. (1992) Chromosome 16-specific repetitive DNA sequences that map to chromosomal regions known to undergo breakage/rearrangement in leukemia cells. *Genomics* **7**, 332–8.
41. Jurka, J. and Batzer, M. (1995) In Meyers, R. (ed.) *Human repetitive elements* (in press).
42. Marcus, S., Hellgren, D., Lambert, B., Fallstrom, S. and Wahlstrom, J. (1993) Duplication in the hypoxanthine phosphoribosyl-transferase gene caused by Alu-Alu recombination in a patient with Lesch Nyhan syndrome. *Hum. Genet.* **90**, 477–82.
43. Heikinnen, J., Hautala, T., Kivirikko, K. and Myllyla, R. (1994) Structure and expression of the human lysyl hydroxylase gene (PLOD): Introns 9 and 16 contain Alu sequences at the sites of recombination in Ehlers-Danlos syndrome Type VI patients. *Genomics* **24**, 464–71.
44. Markert, M., Hutton, J., Wiginton, D., States, J. and Kaufman, R. (1988) Adenosine deaminase (ADA) deficiency due to deletion of the ADA gene promoter and first exon by homologous recombination between two Alu elements. *J. Clin. Invest.* **81**, 1323–7.
45. Nicholls, R., Fischel-Ghodsian, N. and Higgs, D. (1987) Recombination at the human alpha-globin gene cluster: sequence features and toplogical constraints. *Cell* **49**, 369–78.
46. Morris, T. and Thacker, J. (1993) Formation of large deletions by illegitimate recombination in the *HPRT* gene of primary human fibroblasts. *Proc. Natl Acad. Sci. USA* **90**, 1392–6.
47. Davis, M., Kim, S. and Hood, L. (1980) DNA sequences mediating class switching in alpha-immunoglobulins. *Science* **209**, 1360–5.
48. Hengstschläger, M., Maizels, M. and Leung, H. (1995) Targeting and regulation of immunoglobulin gene somatic hypermutation and isotype switch recombination. *Prog. Nucleic Acids Res.* **50**, 67–99.
49. Vogt, P. (1990) Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved 'chromatin folding code'. *Hum. Genet.* **84**, 301–36.
50. Arakawa, H., Iwasato, T., Hayashida, H., Shimizu, A., Honjo, T. and Yamagishi, H. (1993) The complete murine immunglobulin class switch region of the alpha heavy chain gene-hierarchic repetitive structure and recombination breakpoints. *J. Biol. Chem.* **268**, 4651–5.
51. Iwasato, T., Shimizu, A., Honjo, T. and Yamagishi, T. (1990) Circular DNA is excised by immunoglobulin class switch recombination. *Cell* **62**, 143–9.
52. Matsuoka, M., Yoshida, K., Maeda, T., Usuda, S. and Sakano, H. (1990) Switch circular DNA formed in cytokine-treated mouse splenocytes: Evidence for intramolecular DNA deletion in immunoglobulin class switching. *Cell* **62**, 135–42.
53. Wahls, W., Wallace, L. and Moore, P. (1989) Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. *Cell*, 95–103.
54. Jeffreys, A., Neumann, R. and Wilson, V. (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation. *Cell* **60**, 473–85.
55. Krowczynska, A., Rudders, R. and Krontiris, G. (1990) The human minisatellite consensu at breakpoints of oncogene translocations. *Nucleic Acids Res.* **18**, 1121–7.
56. Gibbs, M., Collick, A., Kelly, R. G. and Jeffreys, A. J. (1993) A tetranucleotide repeat mouse minisatellite dsplaying substantial somatic instability during early preimplantation development. *Genomics* **17**, 121–8.
57. Kelly, R., Bulfield, G., Collick, A., Gibbs, M. and Jeffreys, A. (1989) Characterization of a highly unstable mouse minisatellite locus: Evidence for somatic mutation during early development. *Genomics* **5**, 844–56.
58. Sundquist, W. and Klug, A. (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature* **342**, 825–9.
59. Sen, D. and Gilbert, W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications in meiosis. *Nature* **334**, 364–6.
60. Sen, D. and Gilbert, W. (1990) A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature* **344**, 410–14.
61. Lowe, L., Heerema, N., Cheerva, A. and Palmer, C. (1992) A new nonrandom chromosomal abnormality, t(2:16)(p11.2;p11.2), possibly associated with poor outcome in childhood acute lymphoblastic leukemia. *Cancer Genet. Cytogenet.* **54**, 60–4.
62. Trask, B., Massa, H., Evans, J., Scherer, S., Friedman, C., Youngblom, J., Rouquier, S., Giorgi, D., Martin-Gallardo, A., Wong, D., Iadonato, S., Yokota, H., van den Engh, G., Hearst, J. and Sachs, R. (1995) In Bentley, D., Green, E. and Warterston, R. (eds), *Applications of FISH in genome analysis.* Cold Spring Harbor Press: Cold Spring Harbour, NY, p. 14.
63. Longmire, J., Brown, N., Meincke, L., Campbell, M., Albright, K., Fawcett, J., Campbell, E., Moyzis, R., Hildebrand, C., Evans, G. and Deaven, L. (1993) Construction and characterization of partial digest DNA libraries made from flow-sorted human chromosome 16. *GATA* **10**, 69–76.
64. Lee, C., Yazdani, A., Wehnert, M., Zhao, Z., Lindsay, E., Bailey, J., Coolbaugh, M., Couch, L., Xiong, M., Chinault, A. *et al.* (1995) Isolation of chromosome-specific genes by reciprocal probing of arrayed cDNA and cosmid libraries. *Hum. Mol. Genet.* **4**, 1373–80.
65. Parrish, J. E., Oostra, B. A., Verkerk, A. J. M. H., Richards, C. S., Reynolds, J., Spikes, A. S., Shaffer, L. G. and Nelson, D. L. (1994) Isolation of a GCC repeat showing expansion in FRAXF, a fragile site distal to FRAXA and FRAXE. *Nature Genet.* **8**, 229–35.
66. Stallings, R., Torney, D., Hildebrand, C., Longmire, J., Deaven, L., Jett, J., Doggett, N. and Moyzis, R. (1990) Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl Acad. Sci.* **87**, 6218–22.
67. Povinelli, C. M. and Gibbs, R. A. (1993) Large-scale sequencing library production: an adaptor-based strategy. *Analyt. Biochem.* **210**, 16–26.
68. Bankier, A. T., T., W. K. and Barrel, B. G. (1987) Random cloning and sequencing by the M13 dideoxynucleotide termination melthod. *Meth. Enzym.* **155**, 51–93.

69. Kristensen, T., Voss, H. and Ansorge, W. (1987) A simple and rapid preparation of M13 templates for manual and automated dideoxy sequencing. *Nucleic Acids Res.* **15**, 5507–16.
70. Civitello, A. B., Richards, S. and Gibbs, R. A. (1992) A simple protocol for the automation of DNA cycle sequencing reactions and polymerase chain reactions. *J. DNA Seq. Map.* **3**, 17–23.
71. Munzy, D., Richards, S., Shen, Y. and Gibbs, R. (1993) In Venter, C. (ed.) *PCR based strategies for gap closure in large scale sequencing projects.* Harcourt, Brace Jovanovich: London.
72. Wilson, R., Chen, C. and Hood, L. (1990) Optimization of asymmetric polymerase chain reaction for rapid fluorescent DNA sequencing. *BioTechniques* **8**, 184–9.
73. Dear, S. and Staten, R. (1991) A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.* **19**, 3907–11.
74. Uberbacher, E. and Mural, R. (1991) Locating protein-coding regions in human DNA sequences by multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA* **88**, 11261–5.
75. Milosavljevic, A. (1993) In Hunter, L., Shavlik, J. and Searls, D. (eds) *Discovering sequence similarity by the Algorithmic significance method.* AAAI Press.
76. Baldini, A., Miller, D., Shridhar, V., Rocchi, M., Miller, O. and Ward, D. (1991) Comparative mapping of a gorilla-derived alpha satellite DNA on great ape and human chromosomes. *Chromosoma* **101**, 109–14.