Divergent Origins and Concerted Expansion of Two Segmental Duplications on Chromosome 16

E. E. Eichler, M. E. Johnson, C. Alkan, E. Tuzun, C. Sahinalp, D. Misceo, N. Archidiacono, and M. Rocchi

An unexpected finding of the human genome was the large fraction of the genome organized as blocks of interspersed duplicated sequence. We provide a comparative and phylogenetic analysis of a highly duplicated region of 16p12.2, which is composed of at least four different segmental duplications spanning in excess of 160 kb. We contrast the dispersal of two different segmental duplications (LCR16a and LCR16u). LCR16a, a 20 kb low-copy repeat sequence A from chromosome 16, was shown previously to contain a rapidly evolving novel hominoid gene family (*morpheus*) that had expanded within the last 10 million years of great ape/human evolution. We compare the dispersal of this genomic segment with a second adjacent duplication called LCR16u. The duplication contains a second putative gene family (KIAA0220/SMG1) that is represented approximately eight times within the human genome. A high degree of sequence identity (~98%) was observed among the various copies of LCR16u. Comparative analyses with Old World monkey species show that LCR16a and LCR16u originated from two distinct ancestral loci. Within the human genome, at least 70% of the LCR16u copies were duplicated in concert with the LCR16a duplication. In contrast, only 30% of the chimpanzee loci show an association between LCR16a and LCR16u duplications. The data suggest that the two copies of genomic sequence were brought together during the chimpanzee/human divergence and were subsequently duplicated as a larger cassette specifically within the human lineage. The evolutionary history of these two chromosome-specific duplications supports a model of rapid expansion and evolutionary turnover among the genomes of man and the great apes.

Recent studies in primate comparative primate genomics have indicated that gene order and chromosomal organization have remained largely invariant over relatively short periods of evolutionary time (Haig 1999; O'Brien and Stanyon 1999). This is especially evident among the chromosomes of humans and the great apes, where cross-species hybridization studies using human chromosome paints as probes reveal extensive homology (Jauch et al. 1992; Muller et al. 1999, 2000). Between chimpanzee and human genomes relatively few differences have been reported beyond those originally described initially by karyotype studies. A collection of pericentric inversions and the well-known fusion between chromosomes XII and XIII (chimpanzee) distinguish cytogenetically the genomes of man and the great apes (Dutrillaux et al. 1986; Garver 1980; Yunis and Prakash 1982; Yunis et al. 1980). With the exception of a few anomalous primate genomes, such as the hylobatids (Jauch et al. 1992; Koehler et al. 1995), most of the catarrhine primates appear to possess very large, conserved ancient linkage groups (Haig 1999).

Recent sequencing and analysis of the human genome has added another level of complexity to chromosomal evolution among hominoids. An unexpected property of the human genome is the relatively large number of small duplications that are spread throughout both pericentromeric and euchromatic regions of chromosomes (Eichler 1998, 2001; International Human Sequencing Consortium 2001; O'Keefe and Eichler 2000). These duplications range in size from 1 to 100 kb in length. They are organized into modules or segments corresponding to a portion of intron-exon structure of ancestral loci. The majority of these "duplicons" are organized in a nontandem fashion throughout the genome, indicating that their expansion has occurred by a mechanism other than unequal crossing over. Sequence comparisons of these segments reveal a high degree of sequence identity,

From the Department of Genetics and Center for Human Genetics, Case Western Reserve School of Medicine and University Hospitals of Cleveland, Cleveland, OH 44106 (Eichler and Johnson), Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106 (Alkan, Tuzun, and Sahinalp), and DAPEG, Sezione di Genetica, Via Amendola 165/A, 70126 Bari, Italy (Rocchi). We thank D. Zivkovic for technical assistance. This work was supported by grants NIH GM58815, NIH HG002385, and DOE ER62862-1013741-0005006 (to E.E.E.) and grants from PRIN, CE, MURST, and Telethon (to M.R.). Computational support was obtained by a grant from the Charles B. Wang Foundation. The financial support of the W. M. Keck Foundation and a Howard Hughes Medical Institute grant to Case Western Reserve University, School of Medicine are also gratefully acknowledged. Address correspondence to Evan Eichler, PhD, Department of Genetics, Case Western Reserve University, BRB720, 10900 Euclid Ave., Cleveland, OH 44106, or e-mail: eee@po.cwru.edu. This paper was delivered at a symposium entitled "Primate Evolutionary Genetics" sponsored by the American Genetic Association at Town and Country Resort and Convention Center, San Diego, CA, USA, May 19–20, 2001.

suggesting that they were duplicated and transposed recently during evolution (Bailey et al. 2001).

Two classes of duplications may be distinguished based on their distribution pattern in the human genome (Bailey et al. 2001; O'Keefe and Eichler 2000). Transchromosomal duplications refer to those interchromosomal duplications located in close proximity to heterochromatic sequence—pericentromeric and subtelomeric duplications (Eichler et al. 1997; Guy et al. 2000; Horvath et al. 2000; Jackson et al. 1999; Trask et al. 1998; van Geel et al. 1999, 2000). A second class of duplications, termed chromosome-specific duplications, is distributed in an interspersed fashion along a chromosome or chromosomal arm (Guy et al. 2000; Loftus et al. 1999; Shaikh et al. 2000). In both cases, these segmental duplications are organized into larger mosaics often consisting of multiple different duplications with diverse evolutionary origin (Eichler 1998, 2001). The mechanism by which these segments have spread from ancestral loci throughout the genome is unknown. The amount of genomic material (5%) and high degree of sequence identity (~96%) among duplicated segments suggest an extraordinary degree of dynamism within the last few million years of hominoid genome evolution (Bailey et al. 2001). This level of structural variation among primates appears to be restricted to very local regions of chromosomes and has not been detected at the resolution of whole chromosome painting experiments (Jauch et al. 1992; Muller et al. 1999, 2000; Stanyon et al. 1995; Weinberg and Stanyon 1995).

During the initial mapping and subsequent sequencing of chromosome 16, an unusual and complex class of segmental duplications was identified whose distribution was, for the most part, specific to the short arm of chromosome 16 (Johnson et al. 2001; Loftus et al. 1999; Stallings et al. 1992, 1993). We recently characterized the most prolific chromosome-specific duplication on chromosome 16, termed LCR16a—low copy repeat chromosome 16 sequence A. Mapping and sequencing of this 20 kb duplication indicated that it had expanded from 1 to approximately 15 copies during the emergence of human and great ape species (5–10 million years ago) (Johnson et al. 2001). The duplication had been distributed in a nontandem fashion to multiple locations along the entire length of chromosome 16, including cytogenetic band positions 16q22.2, 16q23, 16p11, 16p12.2, 16p12.3, 16p13.1, and 16p13.3. A gene family, termed *morpheus,* was identified within the LCR16a duplication. Remarkably, several exons of this gene family showed accelerated rates of amino acid replacement. Extreme positive selection (K_a/K_s = 5.0) was reported for exons 2 and 4, where the effective rate of amino acid replacement was found to exceed by approximately 30-fold that for typical genes under negative selection. The accelerated rate of amino acid change suggested that duplication and diversification of the LCR16a module had occurred as a result of meiotic drive, adaptive evolution, or sexual selection (Johnson et al. 2001).

With a few exceptions, the LCR16a duplications were not solitary but were located in close proximity to other duplicons (termed LCR16a-e, LCR16o-s) (Johnson et al. 2001; Loftus et al. 1999). In this study we present comparative and phylogenetic analysis of an adjacent duplicated segment termed LCR16u and contrast its evolutionary history to that of the LCR16a duplication. Our analysis indicates that among humans that expansion of the LCR16u segment occurred in conjunction with the LCR16a duplicon. In contrast, copies of the LCR16u show significantly less coordinate expansion in man's closest relative, the chimpanzee. Based on the degree of sequence divergence, comparative fluorescence in situ hybridization (FISH), and genomic hybridization results, the data suggest that the LCR16u and LCR16a modules originated from different regions of chromosome 16 in a common catarrhine ancestor. Subsequent rounds of lineage-specific expansion occurred in both the human and chimpanzee lineages. Subtle differences in chromosomal organization and gene order are therefore predicted within these specific genomic regions of humans and great apes.

Materials and Methods

Computational Analysis

A suite of genomic software tools were used to analyze and characterize the sequence content of the duplications. Initially two accessions, AF001549 and AC-003007, were merged into a single sequence contig of approximately 310 kb. The genomic segment was masked for common repeats (using RepeatMasker, default settings) and the program PARASIGHT (Bailey et al. 2001) was used to delineate the junction sequences and the extent of overlap for each duplicated segment with respect to the 16p12.2 reference sequence. We performed optimal global pairwise alignments between various duplicated segments using the program ALIGN (Myers and Miller 1988). Only pairwise sequence alignments greater than 1 kb with a minimum of 90% identity were considered in this analysis. Alignments with greater than 99.5% sequence identity and which did not show junction boundaries were deemed allelic. Estimates of genetic distance (pairwise deletion) were calculated using Kimura's two-parameter model (Kimura 1980). Standard error was calculated using the Boostrap method (MEGA2) (Kumar et al. 1993). Both total genomic and exon-containing DNA were considered in this analysis. Sim4 was used to optimally compare cDNA versus genomic DNA (Florea et al. 1998). Unique sequence differences within the predicted exons from genomic sequence compared to EST sequences were used to identify transcriptionally competent loci. A 12.3 kb portion of the LCR16u duplication was extracted from six different accessions. Using these sequences, we constructed a multiple sequence alignment (ClustalW) and neighbor-joining phylogenetic tree.

Fluorescence In Situ Hybridization

Chromosome metaphase spreads and interphase nuclei were prepared from lymphoblastoid cell lines representative of four hominoid species (*Homo sapiens, Pan troglodytes, Gorilla gorilla, Pongo pygmaeus*) and three Old World monkeys (*Macaca mulatta, Presbytis cristata, Cercopithecus aethiops*). In situ hybridizations were performed using previously described standard conditions (Lichter et al. 1990). A human chromosome BAC (bacterial artificial chromosome) clone (61E3) was used as probe in this study. The clone has been entirely sequenced (AC003007) and contains a 110 kb insert that spans both the LCR16u and LCR16v duplicons. To eliminate the effect of cross hybridization of common repeat sequences, probes were blocked by C_{ot} DNA prior to hybridization. In the determination of copy number and chromosomal band location we examined at least 20 independent metaphase and interphase nuclei. When necessary, hybridizations were performed in conjunction with human whole-chromosome painting probes to confirm chromosomal assignment.

Library Hybridization

Large insert BAC libraries were examined for three primate species: human (RPC-11, segment 1), chimpanzee (RPCI-43; *P. trog-*

lodytes), and the olive baboon (RPCI-41; *Papio hamadryas*). Two probes (16.19 and 61E3.14) were generated by polymerase chain reaction (PCR) amplification from human DNA. Probe 16.19 is a 4.497 kb PCR fragment corresponding to LCR16a. Probe 61E3.14 is a 1.242 kb PCR fragment corresponding to genomic positions 92548 to 93790 within AC003007. The sequence of the oligonucleotides used to generate the probes is

CHLAR16.1

5-AGGGATGTGGTTACCTTTTGGAGG

CHLAR16.9

5-ACCAATGCCAGTACTAGCAACTCC

61E3.1

5′-TACTGCCCCCTAACTTTGGATGTC

61E3.4

5′-GGTGGTCACACAAGTGAAGACATG

PCR amplification was performed using human DNA as template and previously described conditions (Johnson et al. 2001). Different filter sets for each library were hybridized independently using previously described conditions (Horvath et al. 2000). The depth of coverage for each library was 5.9-, 3.5-, and 5.1-fold for RPCI-11, RPCI-43, and RPCI-41, respectively. The estimated copy number was computed by dividing the number of positives by the depth of coverage for each library.

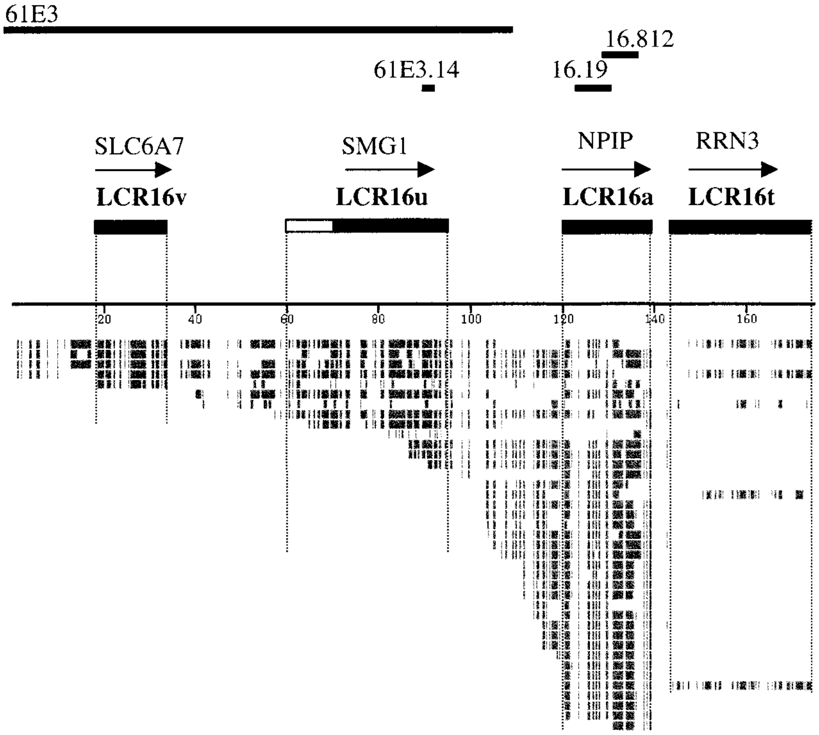

Figure 1. Duplication structure. The schematic displays the organization and extent of duplication for a 170 kb region of 16p12.2. The sequence was extracted from complete finished sequence (AC003007, AF001549). The genomic sequence was masked for common repeat sequences (*Alus*, LINEs) and used as a reference to identify segmental duplications from genomic sequence. The gray bars beneath the scale bar indicate the extent of duplication based on the analysis of 47 genomic accessions (the gaps or discontinuities represent the position of common repeats). Minimal shared segments with gene structure are showed above the horizontal line as black bars. Four duplicons—LCRa, LCR16t, LCR16u, and LCR16t—are defined based on a previous nomenclature scheme (Loftus et al. 1999). The presence of duplicated genes (with its putative transcription orientation based on the ancestral expressed loci) is indicated by the arrows.

Results and Discussion

Duplication Structure

We selected a 300 kb region of 16p12.2 due to the presence of a single copy of the LCR16a duplication and its proximity to several other uncharacterized chromosome 16-specific duplications (Johnson et al. 2001; Loftus et al. 1999). In order to study the context of additional duplications in the vicinity of LCR16a, we performed a detailed computational analysis. Initially two completely sequenced genomic segments (AC003007 and AF001549) were assembled into a single sequence contig of 310 kb. The assembled sequence was compared against all available public human genome project data. Paralogous segments were identified using the program PARASIGHT, as previously described (Bailey et al. 2000; International Human Sequencing Consortium 2001). In this analysis we only considered duplicated segments where the individual alignments showed greater than 90% sequence identity and where the aligned segments exceeded 1 kb in length. The pattern of duplicated segments for an approximately 170 kb portion of 16p12.2 is shown in Figure 1.

Our analysis indicates a complex pattern of duplications. More than 95% of the duplicated segments map to other positions on chromosome 16, including cytogenetic map positions at 16q24.3, 16q22.2, 16p11.1, 16p12.1, 16p12.2, 16p12.3, and 16p13.1. These positions are separated by megabases of intervening sequence, suggesting transposition and duplication of sequence over considerable genomic distances. The degree of sequence identity among the duplications ranges from 95.4% to 99.8%. Based on the available data within GenBank, both the extent of duplication and copy number vary considerably along the length of this portion of 16p12.2. The most prolific element, LCR16a, is duplicated 15 times along chromosome 16 (Johnson et al. 2001). Although the pattern of duplications suggests a gradient, we characterized different duplicons (segmental duplications) based on minimal shared overlap, sequence divergence, and the presence of gene structure from putative ancestral loci. Within this 170 kb region of 16p12.2, four distinct duplications could be identified. They were assigned nomenclature in accordance with previous descriptions of low-copy repeat sequences for chromosome 16 (Johnson et al. 2001; Loftus et al. 1999).

LCR16a corresponds to a minimal shared segment extending from 119,549 to 139,970 (Figure 1) as defined by sequence comparison between genomic clones U95742 and AF001549. The segment encodes a novel hominoid gene family termed *morpheus.* One member of this gene family, NPIP (nuclear pore interacting protein; accession AF132984), has been localized to the periphery of the nuclear membrane, where it is predicted to associate with the nuclear pore complex. Exons 2 and 4 of this gene family show extreme positive selection. Most copies of LCR16a were situated in close proximity to other duplicated segments on chromosome 16 (Johnson et al. 2001; Loftus et al. 1999).

LCR16t was identified approximately 2 kb distal to LCR16a. It represents a partial

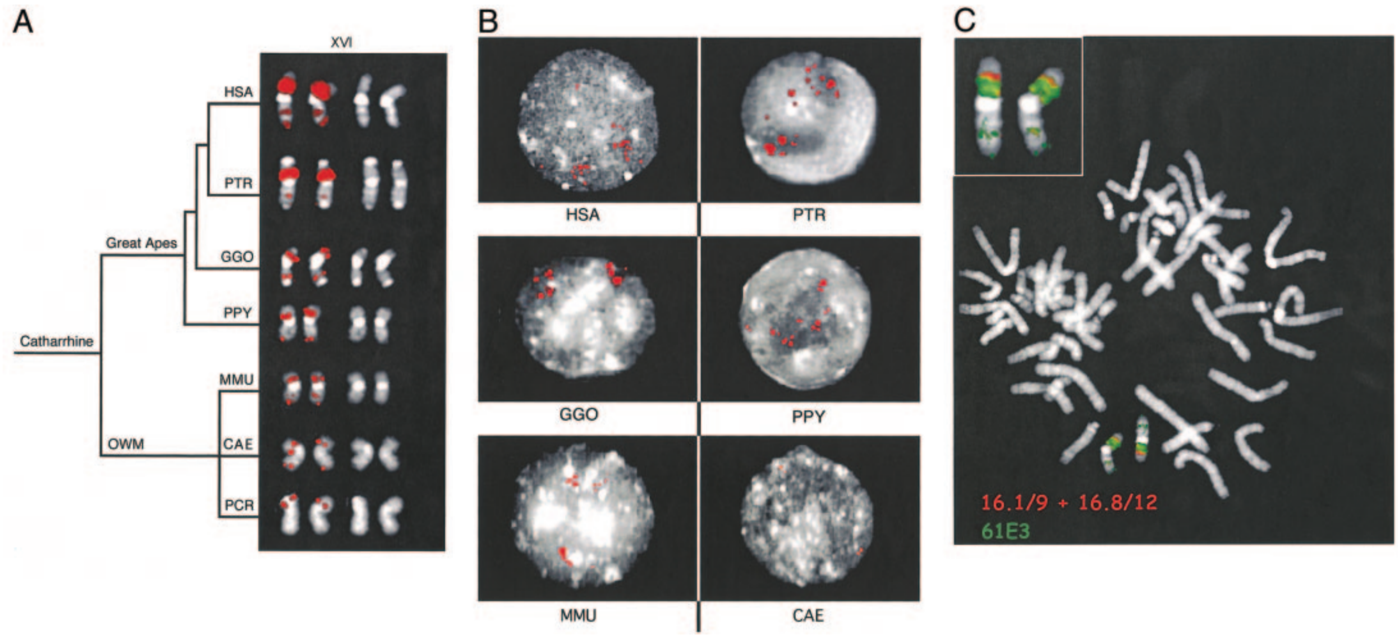

Figure 2. Comparative FISH of 61E3. A human chromosome 16 BAC clone (61E3) was used as a FISH probe against a panel of primate chromosomes. **(a)** Extracted chromosome 16 metaphase hybridized with 61E3 are shown for three Old World (OW) species (MFU = *Macaca fascicularis,* PCR = *Presbytis cristata,* CMO = *Callicebus mollochus*) and four hominoids (HSA = *Homo sapiens,* PTR = *Pan troglodytes,* PPA = *Pan paniscus,* GGO = *Gorilla gorilla,* PPY = *Pongo pygmaeus*). The hybridization results are depicted in the context of a generally accepted phylogeny of the species (Goodman 1999). The roman numeral above metaphase chromosomes is in accordance with standard cytogenetic nomenclature (ISCN 1985). No signals were observed for any other chromosome other than chromosome XVI. The DAPI image of that chromosome is shown to the right for comparison. **(b)** Hybridization of interphase nuclei from a subset of Old World and hominoid species. **(c)** Two-color hybridization of the LCR16a probe (16.18, 161/9; red signal) and the LCR16u-containing probe (61E3; green signal) against a complete human metaphase chromosome. Overlapping signals appear yellow. Note the specificity of hybridization to chromosome XVI.

genomic duplication (13 exons, ~35 kb, greater than 98% sequence identity) of the RNA polymerase I transcription factor gene (*RRN3*). The functional copy of this gene has been assigned to 16p13.11.

LCR16u was identified by sequence comparison between AC020716 and 16p12.2. It contains a partial duplication (17 exons, 96.9% sequence identity) of a novel phosphatidylinositol kinase-related kinase (*SMG-1*) gene (Denning et al. 2001). Based on the closest genomic sequence match, we mapped *SMG-1* (a 54 exon gene spanning 62.5 kb of genomic sequence) to 16p12.3 (AC020716). We estimated the LCR16u duplication at approximately 35 kb based, in part, on this comparison. A partial cDNA, KIAA0220 (D86974), corresponds to 16p12.2, indicating that this locus is transcriptionally competent. At least six copies of LCR16u can be identified mapping to 16q22.2, 16p13.1, 16p12.2, and 16p12.3.

LCR16v is a 13.7 kb duplication (95.6% sequence identity). The extent of duplication is based on sequence comparison to genomic accession AC007442. It contains the first exon and a portion of the first intron of the gene, SLC7A5 (solute carrier 7, family member 5), which has been mapped to 16q24.3 (Maglott et al. 1994).

Comparative FISH

Comparative FISH analysis was conducted to obtain a first approximation of the chromosomal distribution of these duplicated loci among various primate species. Previously probes corresponding to the LCR16a duplication were *in situ* hybridized against metaphase chromosomal spreads and interphase nuclei from various Old World monkey and hominoid species. A major proliferation in copy number of LCR16a (3 to 15 copies) was observed among humans and the African apes. All Old World monkeys showed a single copy corresponding to the putative ancestral locus at 16p13.1 (Johnson et al. 2001). In this study we performed a complementary set of experiments using the probe 61E3. Probe 61E3 is a human chromosome 16 BAC clone that has been completely sequenced as part of the Human Genome Project (113 kb insert; AC003007) (Loftus et al. 1999). It contains both LCR16u and LCR16v duplications, but excludes the LCR16a duplication. We examined by FISH the chromosomal distribution of the 61E3 probe (Figure 2). Similar to results from the LCR16a probe, the proximal portion of the short arm of chromosome 16 (16p11–16p13.1) hybridized most intensely among humans and the great apes. Examination of interphase nuclei revealed the presence of multiple signals along the length of chromosome 16. In contrast to the hominoid species, Old World monkey species show far fewer hybridization signals. These data suggest that both LCR16a and LCR16u have undergone recent hominoid expansion and have been transposed to multiple sites on chromosome 16.

Several important differences are noted between the LCR16a and 61E3 hybridizations. First, in orangutan and common chimpanzee metaphase, LCR16a hybridized to chromosomes other than chromosome XVI (Johnson et al. 2001) (chromosomes XIII and chromosomes VII, respectively). All hybridization signals for 61E3 were specific to chromosome XVI (Figure 2c). Second, two distinct signals for probe 61E3 are present on the long arm of chromosome XVI for most of the primates examined. These likely correspond to regions syntenic to human loci 16q24.3 and 16q22.3, which contain duplicated portions of the 61E3 probe based on the human reference (see above). In contrast, LCR16a hybridizes only to 16q22.3 (Figure 2c). Of interest, among human chromosomes the LCR16a hybridization signals extend further telomerically along the short arm of chromosome 16p. Soli-

Table 1. Comparative hybridization of genomic BAC libraries

	Positive clones (estimated copy number)		
Probe	Human	Chim-panzee	Baboon
A	119 (20)	132 (38)	9 (1–2)
U	45 (8)	36 (10)	6 (1)
A and U	32 (5)	12 (3)	0
A without U	87 (15)	120 (35)	9 (1–2)
U without A	13 (3)	24 (7)	6 (1)

Probe A = 16.19.
Probe U = 61E3.14.

tary copies of the LCR16a duplication (i.e., not associated with other duplications) have been mapped to the distal portion of 16p13 (U95742). Finally, although there is a reduction in the number of signals among Old World monkeys when compared to hominoids, at least two signals can always be discerned for 61E3. This bipartite pattern is likely due to the complex nature of the 61E3 probe, which contains at least two different duplications of diverse evolutionary origin (LCR16u and LCR16v, see above).

Comparative Genomic Hybridization

In order to refine the molecular organization of the LCR16u and LCR16a duplications more precisely among primate species, we designed probes specific for each of the segmental duplications. Probe 16.19 represents a portion of the LCR16a duplication, while probe 61E3.14 specifically targets the LCR16u duplication. We independently hybridized each of these probes against large-insert BAC libraries from three primate species: human, common chimpanzee, and olive baboon. Based on the depth of coverage, we estimated the copy number of these probes within each respective genome (see Materials and Methods). The results are summarized in Table 1. Both "duplications" are present as a single copy in the baboon genome, confirming the reduced signal intensity/interphase signal copy number observed by FISH. Of interest, none of the BACs positive by 61E13.14 cross-hybridize with probe 16.19. Assuming that the Old World monkey is indicative of the ancestral condition, these data imply that the duplications arose after the divergence of the Old World and hominoid lineages (less than 25 million years ago) (Goodman 1999) and that the duplications emerged from two distinct loci on chromosome 16.

In human and chimpanzee, a dramatic increase in copy number of the LCR16a and LCR16u duplications is observed. A comparable copy number estimate for the LCR16u (8–10) duplication is predicted for both species, while the LCR16a locus has nearly doubled from approximately 20 copies in humans to approximately 38 copies in chimpanzee (Johnson et al. 2001). Analysis of Human Genome Project data confirms estimates in humans (7 copies of LCR16u and ~17 copies of LCR16a). Based on the 16p12.2 sequence (Figure 1), the two probes are separated by 38 kb. Our results (Table 1) reveal that 71% (32/45) of human BAC clones positive with the LCR16u duplication were also positive by hybridization for the LCR16a probe. These hybridization results are confirmed by sequence analysis of the human genome. Five of seven LCR16u duplications lie within close proximity to the LCR16a duplication. The organization of these regions is nearly identical to the sequence organization of the 16p12.2 region (Figure 1). These data suggest that in humans the LCR16u locus has been most often duplicated as part of a larger cassette that included the LCR16a segment. LCR16a duplications, however, have been more prolific and have been spread independent from the LCR16u duplication.

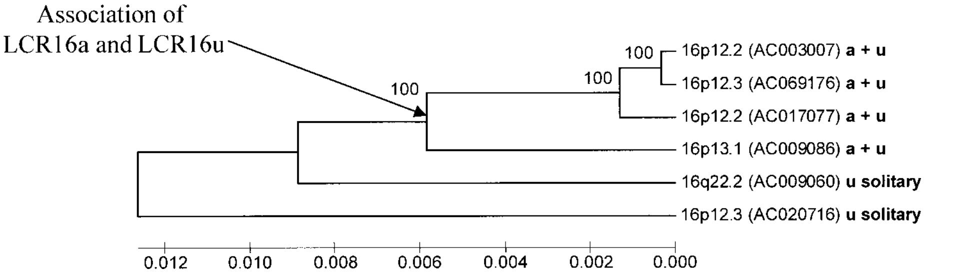

Figure 3. Phylogeny of human LCR16u duplication. An unrooted neighbor-joining phylogenetic [Kimura two-parameter model (Li 1997)] tree for the LCR16u duplication. The phylogenetic analysis is based on the alignment of 12.3 kb of genomic sequence extracted from six genomic accessions for which sufficient contiguous sequence could be obtained. The map position of these accessions is based on assignment within the most current assembly (www.genome.ucsc.edu, April 2001). Sequences that contain or are located immediately adjacent to an LCR16a duplication are denoted "a + u," while LCR16u sequences that do not show an association with LCR16a are indicated as solitary. The midpoint of all trees was set to one-half the distance between AC0020716 (the presumptive ancestral locus) and all other sequences. Only bootstrap values greater than 95% are shown (*n* = 1000 replicates).

One possible scenario may be that the LCR16a sequence represented a duplication focus or an initiation site of gene conversion from which larger segments (including the LCR16u segment) were occasionally generated. In contrast to human genome organization, the comparative genomic hybridization results for chimpanzee provide only limited evidence for concerted duplication of the LCR16u and LCR16a modules. Only 33% (12/36) of chimpanzee BACs were positive for both probes. These data suggest that the duplications have occurred in a lineage-specific fashion (i.e., after the separation of human and chimpanzee) from different source templates. Alternatively, coordinate LCR16u and LCR16a duplications occurred before the separation of human and chimpanzee lineages (less than 6 million years ago), but that the organization of many of the larger duplications (LCR16a + LCR16u) was subsequently scrambled in the chimpanzee lineage.

Phylogenetic Analysis

We investigated the phylogenetic relationship and structure of the LCR16u copies in an attempt to recapitulate its origin and evolutionary dispersal. A genomic segment (12.3 kb) from six LCR16u duplications was extracted and aligned using ClustalW software (Thompson et al. 1994). We constructed an unrooted phylogenetic tree based on this multiple sequence alignment (Figure 3). Copies were selected for which human transcripts had been detected (data not shown). We computed the pairwise genetic distance among all copies for both genomic sequence and transcribed exonic portions of the putative (SMG-1/KIAA0220) gene family. Unlike the LCR16a duplication, no dramatic differences in genetic distance were observed for exonic regions when compared to noncoding genomic sequences (Table 2). This suggests that the duplicated loci have been under largely neutral selection. The analysis unequivocally identified the 16p12.3 copy (represented by AC020716) as the most divergent. This corresponds to the functional SMG-1 locus with its complete complement of 51 exons. All other duplicates contain only 14 exons. We suggest that the other chromosome 16 copies are derivatives of this original ancestral locus.

In humans, the two most divergent LCR16u copies (AC009060 and AC020716)

Table 2. Genetic distance of transcribed versus genomic DNA portions of LCR16u

	1	2	3	4	5	6
1	—	0 (0)	0.002 (0.001)	0.013 (0.002)	0.024 (0.003)	0.022 (0.003)
2	0.001 (0)	—	0.002 (0.001)	0.014 (0.003)	0.024 (.003)	0.023 (0.003)
3	0.003 (0.001)	0.003 (0.001)	—	0.013 (0.002)	0.024 (0.003)	0.022 (0.003)
4	0.012 (0.001)	0.012 (0.001)	0.012 (0.001)	—	0.023 (0.003)	0.021 (0.003)
5	0.017 (0.001)	0.017 (0.001)	0.017 (0.001)	0.019 (0.001)	—	0.018 (0.003)
6	0.025 (0.001)	0.025 (0.001)	0.025 (0.002)	0.025 (0.001)	0.025 (0.001)	—

Kimura two-step parameter estimate of genetic distance (standard error).

Distance based on exonic regions (2047 sites) = upper right of matrix.

Distance based on genomic sequence (12303 sites) = lower left of matrix.

1 = AC003007, 2 = AC069176, 3 = AC017077, 4 = AC009060, 5 = AC009086, 6 = AC020716.

are not associated with LCR16a. In contrast, all of the less divergent copies harbor the larger LCR16a-LCR16u duplication. The maximum degree of sequence divergence among those duplications that contain LCR16a and LCR16u is approximately 1.2% (Figure 3 and Table 2). A recent survey of intronic sequence divergence over a large number of loci between chimpanzee and human was estimated as approximately 1.24% (Chen and Li 2001). In toto, these data suggest that the first association of LCR16a and LCR16u duplications occurred at a time around the separation of the human and chimpanzee lineages (~6 million years ago) (Goodman 1999). Subsequent duplications and transpositions of the larger LCR16a and LCR16u cassette occurred predominantly within the human lineage. This is supported by much lower levels of divergence within the LCR16a + LCR16u clade (Figure 3). We propose that the expansion of many of the LCR16u segments within the chimpanzee genome occurred independently and that the predominant association of LCR16u and LC16a is idiosyncratic to human chromosome 16. Subtle differences in chromosomal organization and gene order are therefore predicted within these specific genomic regions of the humans and great apes. The functional significance of these subtle genomic rearrangements awaits validation. It is intriguing, however, that most of chromosome 16-specific duplications such as LCR16a, LCR16u, LCR16v, and LCR16t all harbor exon-intron structure. The duplication, transposition, and rearrangement of blocks of genomic sequence that we have described are similar to exon-shuffling events. Such events were largely believed to have been restricted to vertebrate antiquity (Gilbert 1985; Gilbert et al. 1997). This analysis, as well as other recent work, indicates that such events were not uncommon during primate evolution (Courseaux and Nahon 2001; Inoue et al. 2001; Johnson et al. 2001). Large-scale sequencing and comparative gene studies are warranted to further clarify the impact of such regions on primate evolution and as a potential force in speciation.

References

Bailey JA, Carrel L, Chakravarti A, and Eichler EE, 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. Proc Natl Acad Sci USA 97:6634–6639.

Bailey JA, Yavor AM, Massa HF, Trask BJ, and Eichler EE, 2001. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11:1005–1017.

Chen FC and Li WH, 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456.

Courseaux A and Nahon JL, 2001. Birth of two chimeric genes in the Hominidae lineage. Science 291:1293–1297.

Denning G, Jamieson L, Maquat LE, Thompson EA, and Fields AP, 2001. Cloning of a novel phosphatidylinositol kinase-related kinase: characterization of the human SMG-1 RNA surveillance protein. J Biol Chem 276: 22709–22714.

Dutrillaux B, Couturier J, Sabatier L, Muleris M, and Prieur M, 1986. Inversions in evolution of man and closely related species. Ann Genet 29:195–202.

Eichler EE, 1998. Masquerading repeats: paralogous pitfalls of the human genome. Genome Res 8:758–762.

Eichler EE, 2001. Segmental duplications: what's missing, misassigned, and misassembled—and should we care? Genome Res 11:653–656.

Eichler EE, Budarf ML, Rocchi M, Deaven LL, Doggett NA, Baldini A, Nelson DL, and Mohrenweiser HW, 1997. Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. Hum Mol Genet 6:991–1002.

Florea L, Hartzell G, Zhang Z, Rubin GM, and Miller W, 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8: 967–974.

Garver JJ, 1980. Genome organization in primates genetics (PhD dissertation). Leiden, The Netherlands: Leiden University.

Gilbert W, 1985. Genes-in-pieces revisited. Science 228: 823–824.

Gilbert W, de Souza S, and Long M, 1997. Origin of genes. Proc Natl Acad Sci USA 94:7698–7703.

Goodman M, 1999. The genomic record of humankind's evolutionary roots. Am J Hum Genet 64:31–39.

Guy J, Spalluto C, McMurray A, Hearn T, Crosier M, Viggiano L, Miolla V, Archidiacono N, Rocchi N, Scott C, Lee PA, Sulston J, Rogers J, Bentley D, and Jackson MS, 2000. Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q [in process citation]. Hum Mol Genet 9:2029–2042.

Haig D, 1999. A brief history of human autosomes. Philos Trans R Soc Lond B Biol Sci 354:1447–1470.

Horvath J, Schwartz S, and Eichler E, 2000. The mosaic structure of a 2p11 pericentromeric segment: a strategy for characterizing complex regions of the human genome. Genome Res 10:839–852.

Horvath J, Viggiano L, Loftus B, Adams M, Rocchi M, and Eichler E, 2000. Molecular structure and evolution of an alpha/non-alpha satellite junction at 16p11. Hum Mol Genet 9:113–123.

Inoue K, Dewar K, Katsanis N, Reiter LT, Lander ES, Devon KL, Wyman DW, Lupski JR, and Birren B, 2001. The 1.4-Mb CMT1A duplication/HNPP deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes. Genome Res 11:1018–1033.

International Human Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409:860–920.

ISCN, 1985. Report of the standing committee on human cytogenetic nomenclature. Birth Defects 21:1–117.

Jackson M, Rocchi M, Hearn T, Crosier M, Guy J, Viggiano L, Piccininni S, Ricco A, Marzella R, Archidiacono N, McMurray A, Sulston J, Rogers J, Bentley D, and Spalluto C, 1999. Characterisation of the heterochromatin/euchromatin boundary at 10q11 and identification of novel transcripts by repeat induced instability. Am J Hum Genet 65(suppl):A56.

Jauch A, Wienberg J, Stanyon R, Arnold N, Tofanelli S, Ishida T, and Cremer T, 1992. Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. Proc Natl Acad Sci USA 89:8611–8615.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, and Eichler EE, 2001. Positive selection of a novel gene family during the emergence of humans and the great apes. Nature 413:514–519.

Kimura M, 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111–120.

Koehler U, Arnold N, Wienberg J, Tofanelli S, and Stanyon R, 1995. Genomic reorganization and disrupted chromosomal synteny in the siamang (*Hylobates syndactylus*) revealed by fluorescence in situ hybridization. Am J Phys Anthropol 97:37–47.

Kumar S, Tamura K, and Nei M, 1993. MEGA: molecular evolutionary genetic analysis, version 1.0. University Park: Pennsylvania State University.

Li W, 1997. Molecular evolution. Sunderland, MA: Sinauer.

Lichter P, Tang CJ, Call K, Hermanson G, Evans GA, Housman D, and Ward DC, 1990. High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. Science 247:64–69.

Loftus B, Kim U, Sneddon V, Kalush F, Brandon R, Fuhrmann J, Mason T, Crosby M, Barnstead M, Cronin L, Mays A, Cao Y, Xu R, Kang H, Mitchell S, Eichler E, Harris P, Venter J, and Adams M, 1999. Genome duplications and other features in 12 Mbp of DNA sequence from human chromosome 16p and 16q. Genomics 60: 295–308.

Maglott DR, Durkin AS, Lane SA, Callen DF, Feldblyum TV, and Nierman WC, 1994. The gene for membrane protein E16 (D16S469E) maps to human chromosome 16q24.3 and is expressed in human brain, thymus, and retina. Genomics 23:303–304.

Muller S, Stanyon R, Finelli P, Archidiacono N, and Wienberg J, 2000. Molecular cytogenetic dissection of human chromosomes 3 and 21 evolution. Proc Natl Acad Sci USA 97:206–211.

Muller S, Stanyon R, O'Brien PC, Ferguson-Smith MA, Plesker R, and Wienberg J, 1999. Defining the ancestral karyotype of all primates by multidirectional chromosome painting between tree shrews, lemurs and humans. Chromosoma 108:393–400.

Myers EW and Miller W, 1988. Optimal alignments in linear space. Comput Appl Biosci 4:11–17.

O'Brien SJ and Stanyon R, 1999. Phylogenomics. Ancestral primate viewed. Nature 402:365–366.

O'Keefe C and Eichler E, 2000. The pathological consequences and evolutionary implications of recent human genomic duplications. In: Comparative genomics: empirical and analytical approaches to gene order dynamics, map alignment and the evolution of gene families. Volume 1: Computational biology (Sankoff D and Nadeau J, eds). Dordrecht, The Netherlands: Kluwer Academic; 29–46.

Shaikh TH, Kurahashi H, Saitta SC, O'Hare AM, Hu P, Roe BA, Driscoll DA, McDonald-Ginn DM, Zackai EH, Budarf ML, and Emanuel BS, 2000. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. Hum Mol Genet 9:489–501.

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 7:332–338.

Stallings R, Whitmore S, Doggett N, and Callen D, 1993. Refined physical mapping of chromosome 16-specific low-abundance repetitive DNA sequences. Cytogenet Cell Genet 63:97–101.

Stanyon R, Arnold N, Koehler U, Bigoni F, and Wienberg J, 1995. Chromosomal painting shows that "marked chromosomes" in lesser apes and Old World monkeys are not homologous and evolved by convergence. Cytogenet Cell Genet 68:74–78.

Thompson JD, Higgins DG, and Gibson TJ, 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

Trask BJ, Massa H, Brand-Arpon V, Chan K, Friedman C, Nguyen OT, Eichler EE, van den Engh G, Rouquier S, Shizuya H, and Giorgi D, 1998. Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. Hum Mol Genet 7:2007–2020.

van Geel M, Heather LJ, Lyle R, Hewitt JE, Frants RR, and de Jong PJ, 1999. The FSHD region on human chromosome 4q35 contains potential coding regions among pseudogenes and a high density of repeat elements. Genomics 61:55–65.

van Geel M, van Deutekom JC, van Staalduinen A, Lemmers RJ, Dickson MC, Hofker MH, Padberg GW, Hewitt JE, de Jong PJ, and Frants RR, 2000. Identification of a novel beta-tubulin subfamily with one member (TUBB4Q) located near the telomere of chromosome region 4q35. Cytogenet Cell Genet 88:316–321.

Weinberg J and Stanyon R, 1995. Chromosome painting in mammals as an approach to comparative genomics. Curr Opin Genet Dev 5:724–733.

Yunis JJ and Prakash O, 1982. The origin of man: a chromosomal pictorial legacy. Science 215:1525–1530.

Yunis JJ, Sawyer JR, and Dunham K, 1980. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. Science 208:1145–1148.

Corresponding Editor: Oliver A. Ryder