Cytogenet Genome Res 108:73–82 (2005)
DOI: 10.1159/000080804

Cytogenetic and Genome Research

Molecular evolution of the human chromosome 15 pericentromeric region

D.P. Locke,[a] Z. Jiang,[a] L.M. Pertz,[a] D. Misceo,[b] N. Archidiacono[b] and E.E. Eichler[a]

[a] Department of Genetics, Center for Computational Genomics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, OH (USA);
[b] Dipartimento di Anatomia Patologica e di Genetica, Sezione di Genetica, University of Bari, Bari (Italy)

Abstract. We present a detailed molecular evolutionary analysis of 1.2 Mb from the pericentromeric region of human 15q11. Sequence analysis indicates the region has been subject to extensive interchromosomal and intrachromosomal duplications during primate evolution. Comparative FISH analyses among non-human primates show remarkable quantitative and qualitative differences in the organization and duplication history of this region – including lineage-specific deletions and duplication expansions. Phylogenetic and comparative analyses reveal that the region is composed of at least 24 distinct segmental duplications or duplicons that have populated the pericentromeric regions of the human genome over the last 40 million years of human evolution. The value of combining both cytogenetic and experimental data in understanding the complex forces which have shaped these regions is discussed.

Segmental duplications are duplicated blocks of genomic DNA, often containing high copy repetitive elements as well as intron-exon structure (IHGSC, 2001). Recent studies into the extent of segmental duplication in the human genome estimate approximately 5% of the entire genome consists of duplications (Bailey et al., 2001, 2002a). Pericentromeric regions, in particular, have been shown to be enriched in such sequences (Bailey et al., 2001). In fact, the pericentromeric region of more than half of all human chromosomes is comprised of blocks of segmental duplications extending between the centromeric satellite sequence and unique sequence (Bailey et al., 2001; IHGSC, 2001).

A two-step model has been proposed for the evolution of the complex structure of duplications within these regions, involving an initial seeding of material into a pericentromeric region, and subsequent swapping of that duplicated material, or larger composite blocks of duplications, between chromosomes (Horvath et al., 2000b). However, only five such pericentromeric regions have been thoroughly characterized at both the structural level within the human genome, and from the evolutionary perspective (Guy et al., 2000; Horvath et al., 2000a, b; Footz et al., 2001; Brun et al., 2003). Discerning the progression and pattern of duplication events that have generated the mosaic of segmental duplications in the pericentromeric region of many human chromosomes involves a comparison to the genomes of closely related primate species. Several studies have investigated the evolutionary history of individual duplicated segments (Arnold et al., 1995; Eichler et al., 1996, 1997; Regnier et al., 1997; Zimonjic et al., 1997; Orti et al., 1998; Horvath et al., 2000b, 2003; Golfier et al., 2003). Few studies to date, however, have investigated the evolutionary history of a large pericentromeric region encompassing several duplicated segments.

Several segmental duplications have been mapped to the pericentromeric region of chromosome 15, including immunoglobulin heavy chain (IgH) V and D segment, gamma-amino-

Supported in part by NIH grant GM58815 to E.E.E. In addition, the authors gratefully acknowledge Telethon, CEGBA (Centro di Eccellenza Geni in campo Biosanitario e Agroalimentare), MIUR (Ministero Italiano della Università e della Ricerca; Cluster C03, Prog. L.488/92) and the European Commission (INPRIMAT, QLRI-CT-2002-01325) for financial support.

Received 18 November 2003; manuscript accepted 9 December 2003.

Request reprints from: Evan Eichler, PhD, Associate Professor
University of Washington, Genome Sciences, HSB K336B
Box 357730, 1705 NE Pacific St., Seattle, WA 98195 (USA)
telephone: 206-543-9526; fax: 206-685-7301; e-mail: eee@gs.washington.edu

butyric acid receptor subunit alpha 5 (GABRA5), neurofibromatosis 1 (NF1), B-cell CLL/lymphoma 8 (BCL8) and KIAA0187 (BMS1L) derived duplications (Tomlinson et al., 1994; Kehrer-Sawatzki et al., 1997; Regnier et al., 1997; Barber et al., 1998; Ritchie et al., 1998; Crosier et al., 2002; Dyomin et al., 2002; Fantes et al., 2002). Several of these duplications have been characterized as a polymorphic "cassette" of approximately 1 Mb in size which varies in copy number in the human population (Barber et al., 1998; Ritchie et al., 1998; Fantes et al., 2002). The underlying nature of this polymorphic region is not well understood, however, due to the fact that a complete sequence map of 15q11 has not been resolved to date by the Human Genome Project. The assembly and analysis of highly duplicated pericentromeric regions has required the development of specialized strategies that use stringent standards of nucleotide identity to determine paralogous versus homologous overlap (Horvath et al., 2000a). In addition, the development of bioinformatic tools and methods has been essential to resolving overlaps.

In this study we have applied both a primate cytogenetic and phylogenetic approach, to explore the evolutionary history of the pericentromeric region of human chromosome 15. We first sought to construct a contig of BAC sequences from the pericentromeric region of 15q11 using a strict threshold of sequence identity as evidence of allelic overlap. Subsequent sequence analysis of the contig identified a tiling path of clones for comparative FISH in human, common chimpanzee, gorilla, orangutan, macaque and baboon. Comparison of the FISH signals obtained in human hybridizations with sequence analysis of the human genome assembly has facilitated the identification of regions potentially absent from the human genome sequence. In addition, we were able to identify a multitude of individual duplicons within the contig and perform phylogenetic comparisons to all paralogous sequences within the human genome assembly. The cytogenetic and phylogenetic evidence suggest a substantial portion of the mosaic structure observed within 15q11 emerged from a burst of primate segmental duplication which occurred shortly after the divergence of the African and Asian great ape lineages.

Materials and methods

Validation of the 15q11 assembly

Due to the duplicated nature and overall complexity of the region, we utilized three independent methods of verifying the assembly of the 15q11 pericentromeric region. First, we utilized seed sequences which had been mapped to chromosome 15 by sequence comparison with monochromosomal hybrid sources as described previously (Horvath et al., 2000a, b). Clones RP11-1360M22 (Horvath et al., 2003) and RP11-509A17 (unpublished data) were validated as chromosome 15 clones in this manner. Second, we reassembled the region using stringent standards for considering allelic sequence overlap as opposed to paralogous overlap. Reiterative sequence similarity searches against the non-redundant nucleotide (NT) and high throughput genome sequence (HTGS) databases were performed. Overlaps required a dove-tail configuration with a minimum of >99.9% sequence identity of alignment and at least 10 kb in length. During this process, seed sequences were filtered for high copy repeats using the lower-case RepeatMasker option (Smit and Green, 1999) which allowed for extension through these regions using MEGABLAST (Zhang et al., 2000). Third, we selected a tiling path of BAC clones for FISH analysis (described below) to validate the mapping of these clones to chromosome 15.

STS analysis

At the distal end of the 860-kb contig an STS was developed and PCR amplified using a pair of oligonucleotides designed to GenBank accession no. AC023310: namely; AC023310.3 (5′-GAAATTTATGGTCAATCTCCCC-3′) and AC023310.4 (5′-TATTGCCCAATAGGATGTCG-3′). The PCR product was subsequently radiolabeled and hybridized to RP11 BAC library filters as described (Eichler et al., 1997). The inserts of the resulting positive BACs were end sequenced as described below and the end sequences used as queries in similarity searches against the 860-kb contig, as well as the NT and HTGS sequence databases. From this analysis, a single clone, RP11-1115P6, was identified which linked the 860-kb proximal contig to a 265-kb contig of clones RP11-32B5 and RP11-275E15.

BAC end sequence analysis

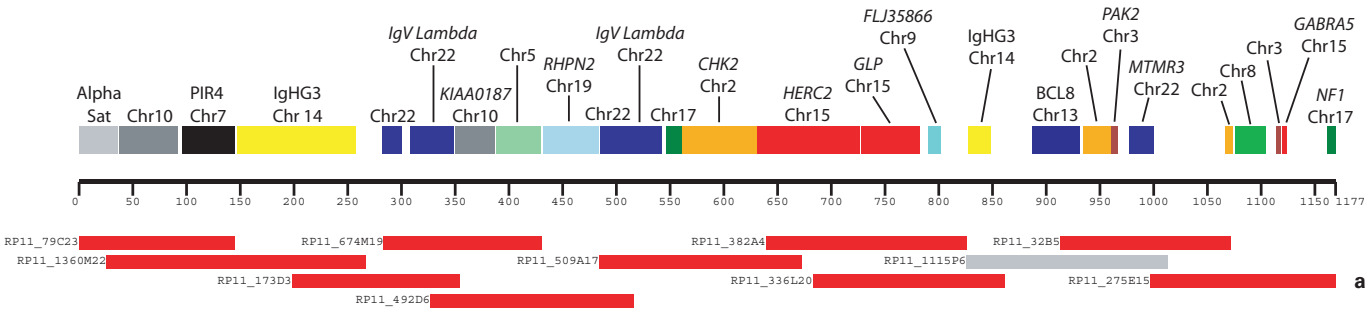
To verify the identity of all clones used in the FISH analyses, all clones were subjected to end sequencing analysis. BAC DNA was extracted from 250 ml LB, chloramphenicol bacterial cultures grown overnight by column purification (Nucleobond, Clontech) and resuspended in 100 µl H_2O. After determining the DNA concentration by spectrophotometry, 2 µg of BAC DNA was subjected to automated dideoxy-terminator cycle sequencing using the ABI Big Dye terminator sequencing chemistry (Applied Biosystems) using vector primers T7 and SP6. Sequencing reaction products were purified using G-50 Sephadex purification columns and analyzed on an ABI 3100 DNA Analyzer (Applied Biosystems). BLAST sequence similarity searches against GenBank confirmed the identity of the clones. The clones which comprise the verified contig include RP11-79C23 (AC138701), RP11-1360M22 (AC127381), RP11-173D3 (AC087386), RP11-674M19 (AC142539), RP11-492D6 (AC126603), RP11-509A17 (AC026495), RP11-382A4 (AC138748), RP11-336L20 (AC023310), RP11-1115P6 (submitted for sequencing based on this analysis in collaboration with the Whitehead Institute Center for Genomic Research), RP11-32B5 (AC068446), RP11-275E15 (AC060814).

Duplicon delineation

Underlying duplicons (ancestral segmental duplications) were delineated using two independent methods. The first strategy determined the minimal evolutionary shared segment (MESS) from a global analysis of all pairwise alignment as described previously (Bailey et al., 2002b). Briefly, sequence similarity searches using 15q11 sequence as a query are made against both the human genome assembly (build34, July 2003) and the NT and HTGS databases. All pairwise alignments are evaluated for percent identity, length, and chromosomal location. The alignments are manually curated using the graphical display program Parasight (Jeff Bailey, unpublished) which facilitates the determination of duplicon boundaries and distribution throughout the genome. Ancestral duplication fragments correspond to those which are either the shortest, most divergent pair and/or an ancestral gene structure can be identified where a complete intron-exon complement can be deduced. Not all duplicons may be delineated using this method. The second method entails the comparison of the human and mouse sequence to identify the "best match" within the mouse genome sequence. The mouse-human alignment data were produced by BLASTZ comparison of the mouse and human genomes (Schwartz et al., 2003), and the linkage of BLASTZ mouse-human alignments into what are termed chains and nets (W.J. Kent 2003). As the duplications found in 15q11 represent primate-specific events, the comparison of human duplicated segments with the mouse genome greatly simplifies the analysis. One caveat, however, is that regions duplicated in the mouse genome are uninformative in such a comparison, although the level of segmental duplication in the mouse genome is lower than that of the human genome (Cheung et al., 2003; Bailey et al., 2004). Once the mouse loci that corresponded to the 15q11 pericentromeric sequences were identified, the mouse sequences were searched against the genome assembly to identify putative ancestral human loci. Thus we utilized two methods of defining the position and length of duplicons that originated from multiple interchromosomal locations within the contiguous set of 15q11 pericentromeric clones.

Phylogenetic analysis

Non-coding sequences were extracted from each of the 24 duplicons and subjected to phylogenetic analysis (Table 2). A threshold of >5 kb was chosen as the likelihood of detecting smaller duplications by comparative FISH was unlikely. Multi-sequence alignments, generally >1 kb if possible, were gener-



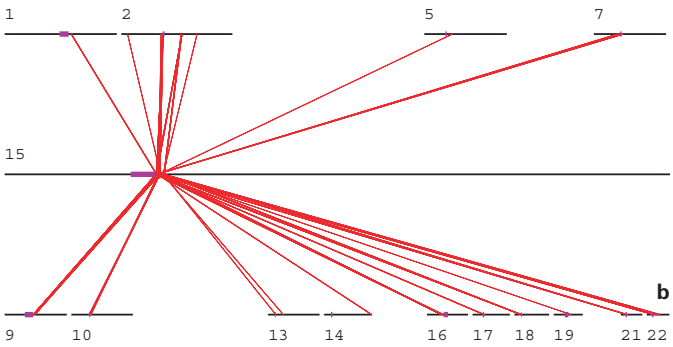


Fig. 1. (**a**) Organization of human 15q11. The tiling path of BAC clones is shown, terminating proximally in monomeric alpha satellite sequence and extending distally from left to right. Tick marks along the black line are placed at 50-kb intervals. Clones shown in red have been completely sequenced or are in working draft status. The single clone shown in grey (RP11-1115P6) is depicted according to STS hybridization and BAC end sequence placement. Above the black line the complex mosaic of duplicons is depicted with color coding according to the chromosome of the ancestral segment prior to duplication. The identity of pseudogenes associated with these duplicons, if known, is indicated. (**b**) Interchromosomal duplication pattern of 15q11. The diagram shows the complexity of interchromosomal duplications (chromosome connecting lines) which represent alignments >20 kb in length and >90 % identity (based on analysis of the July 2003 finished human genome assembly). Centromeres are indicated as purple boxes. Most of the duplications occur between pericentromeric regions or ancestral pericentromeric regions within the human genome.

ated using ClustalW version 1.82 (Thompson et al., 1994) and pairwise distance calculations and phylogram construction was performed using the MEGA software package version 2.1 (Kumar et al., 1993). The neighbor-joining method was used to generate phylograms and pairwise distance calculations were corrected for multiple substitutions by using the Kimura 2-parameter model of nucleotide substitution (Kimura, 1980).

To estimate the divergence time of the chromosome 15 duplicons and their associated ancestral loci , the formula r = k/2T was employed, where r is the rate of nucleotide changes per bp per year, k is the distance calculated between the ancestral and chromosome 15 sequence, and T is the time of divergence of the molecules. The rate of nucleotide change in the 15q11 pericentromeric contig was determined by alignment of two duplicons for which corresponding chimpanzee sequence was available. Specifically, alignments of RP11-79C23 with AC122174 (15.1 kb), and RP11-509A17 with AC124220 (8.9 kb) were used to calculate independent rates based on a divergence time of 6 Million years (My) between the chimpanzee and human lineages (Goodman, 1999). The independent rates were averaged, resulting in a rate of 1.9×10^{-9} nucleotide changes per bp per year. This estimate is in close agreement with previously published mutation rates for duplicated sequences (Horvath et al., 2003).

Comparative FISH

Metaphase chromosome preparations were prepared from lymphoblastoid cell lines derived from humans (*Homo sapiens* [HSA]) and five non-human primate species including three great ape species (*Pan troglodytes* [PTR], *Gorilla gorilla* [GGO], *Pongo pygmaeus* [PPY]), and two old-world monkey species (*Macaca mulatta* [MMU] and *Papio hamadras* [PHA]). Hybridizations were performed using standard conditions with BAC DNA probes labeled with either biotin-16-dUTP or digoxigenin-11-dUTP as previously described (Horvath et al., 2000b). At least 20 metaphases were examined for each hybridization. In situ hybridization experiments were repeated and only consistent signals were recorded in order to minimize potential extraneous signals from these multi-site clones. Chromosome identity was determined by DAPI staining and reported according to the guidelines of the International Standard for Cytogenetic Nomenclature (ISCN 1985).

Results and discussion

In silico analysis of the 15q11 pericentromeric region

The presence of highly duplicated sequences has been problematic for the sequencing and assembly of the human genome due to either an under-representation of paralogous sequences within genome databases or due to misassembly of duplicated sequence (Bailey et al., 2001, 2002a). Consequently, such regions require additional scrutiny. To avoid such potential pitfalls, we independently assembled the pericentromeric region of 15q11 (Fig. 1).

The pericentromeric sequence assembly in 15q11 is composed of two subcontigs: the most proximal contig consists of eight overlapping BACs and spans 865 kb, while the distal sequence contig consist of two BACs which span 265 kb (Fig. 1). Our analysis of 15q11 generally conforms to the finished human genome sequence assembly (build34, July 2003) with one important exception. STS hybridization experiments were performed to identify a clonal link, RP11-1115P6 which

bridges the existing sequence gap between these two contigs (see Methods). Given the average insert size of the BAC library as 196 kb (RP11-Segment 5; www.chori.org/bacpac), and the extent of the BAC-end sequence overlap of the traversing clone, RP11-1115P6, we estimate the sequence gap to be approximately 60 kb. Combined, the 15q11 pericentromeric region is approximately 1.2 Mb in length, representing one of the longest contiguous clone assemblies within human pericentromeric DNA.

Based on this finished high-quality sequence, we analyzed the global segmental duplication content of 15q11 using previously described methods and graphical viewing software (see Methods). A total of 228 pairwise alignments were detected with significant sequence similarity (>1 kb >90%) to this portion of 15q11. The average similarity of these alignments was 94.7% with an average length of 7.5 kb and a range of 1.0 kb to 77.3 kb. Figure 1b simplistically depicts the interchromosomal duplication content (> 90% sequence identity; >20 kb) for this 1.2 Mb portion of chromosome 15q11. A more detailed analysis of the underlying alignments is also presented (Supplemental Fig. 1, www.karger.com/doi/10.1159/000080804). Similar to previous analyses of human pericentromeric regions (Jackson et al., 1999; Guy et al., 2000; Horvath et al., 2000a, 2000b; Footz et al., 2001), in silico analysis confirms that the pericentromeric region of 15q11 is composed of a complex mosaic of small and large duplications. Most regions share homology to three or more distinct regions of the genome. Also, most interchromosomal sites with similarity to the 15q11 pericentromeric region occur within other pericentromeric regions of the human genome. Notable exceptions to this trend include sequence homology to the subtelomeric region of chromosome 14, a large block of homology to the Prader-Willi/Angelman breakpoint associated HERC2 and LCR15 duplications and the euchromatic site within chromosome 2q21 which corresponds to an ancestral pericentromeric region. Specifically, the alignments generated by comparison of the 15q11 contig with the ancestral pericentromeric region of chromosome 2q21 resulted in 10 alignments of 8.8 kb average length and 91.9% average percent identity, with alignments ranging from 2.4 kb to 20.5 kb.

We also identified a block of 28 kb of alpha satellite sequence within the most proximal portion of this sequence contig (RP11-79C23). The alpha-satellite sequence shows no evidence of higher-order repeats (unpublished DOTTER analysis) nor significant sequence similarity (> 90%) to previously characterized higher order-repeat sequences for chromosome 15. Based on previous published alpha satellite/non-alpha satellite transition regions (Horvath et al., 2000a, b; Schueler et al., 2001; Guy et al., 2003), it is likely that this monomeric block of alpha satellite demarcates the boundary between higher-order and monomeric alpha satellite repeat sequences typical for such alpha satellite/non-alpha satellite transitions.

Gene content within duplicons was identified through comparison with the UCSC Genome Browser (www.genome.ucsc.edu) resulting in the following partial gene structures: immunoglobulin heavy chain gamma 3 (IGHG3), immunoglobulin lambda locus (IGL), KIAA0187 (BMS1L), rhophilin like protein (RHPN2), CHK2 checkpoint homolog (CHEK2), hect domain and rcc1 domain protein 2 (HERC2), golgi autoantigen, golgin subfamily a, member 6 (GOLGA6), FLJ35866 (C9orf79), neurobeachin (NBEA) (also known as B-cell lymphoma CLL/lymphoma 8 [BCL8]), p21 activated kinase 2 (PAK2), myotubularin related protein 3 (MTMR3), gamma-aminobutyric acid receptor subunit alpha 5 (GABRA5) and neurofibromatosis 1 (NF1) gene segments. Not all duplications analyzed in this study, however, contained gene content, including the pericentromeric interspersed repeat, PIR4.

FISH analysis of human 15q11

FISH has served as a powerful tool to assess the quality of sequence assembly within human pericentromeric regions (Cheung et al., 2001; Horvath et al., 2001; Bailey et al., 2002b). Since most duplicated segments are large (> 20 kb) and share considerable sequence identity (> 95%), the consistent presence or absence of multi-site FISH signals has been used to suggest potential errors or gaps within the genome assembly. We, therefore, selected a tiling path of 11 clones, confirmed their identity by BAC end sequence analysis and assessed their multi-chromosomal distribution by FISH on human metaphase chromosomes (Table 1). From the cytogenetic perspective, these 11 clones produced a total 71 distinct metaphase signals for an average of ~6 chromosomal signals per probe. Interestingly, 12 of these signals appeared to have no underlying support from the most recent human genome assembly, for a discordance rate of 16.9% (Table 1). Although this rate appears relatively high, it should be pointed out that four of the discordant signals map to chromosome 1 and correspond to 4 contiguous BAC clones spanning ~400 kb of sequence. Chromosome 1 has the highest density of sequence gaps with slightly less than a quarter of all remaining gaps mapping to this chromosome. This suggests that these four clones likely identify a single sequence gap (400 kb in size) within the pericentromeric region of this chromosome. Similarly, the absence of in silico signals for 14q11 (RP11-32B5 and RP11-257E15) likely corresponds to a pericentromeric gap on this chromosome. If we consider these as single large gaps within the finished human genome, the discordance rate drops to 8.5%. In other words, 8.5% of human pericentromeric regions with sequence similarity to 15q11 are still not faithfully represented within the "finished" human genome and therefore likely represent gaps that remain to be sequenced and assembled.

From the in silico perspective of the genome assembly, the correlation is also good, but not absolute. In total, 80.3% of the sites considered as potentially duplicated by computational analyses were confirmed by FISH signals on human metaphase chromosomes (Table 1). Interestingly, a consistent lack of signal for chromosome 10 duplications was observed for six of the clones used in this study, with two potential explanations. First, the sequence analysis may be detecting sequences that are duplicated below the threshold of FISH to produce clearly visible signals. Focusing on clone RP11-173D3, the sequence relationship between this clone and chromosome 10 extends approximately 39 kb with an average identity of 96.4%, which is typically sufficient to produce a FISH signal. Alternatively, the chromosome 10 region is highly polymorphic in the human population and the material used to assess the distribution of

Table 1. FISH localization of human chromosome 15 pericentromeric BAC clones

Clone	HSA[a]	PTR	GGO	PPY	MMU	PHA
RP11-79C23	1, 2, 7, 9, 13, ***14***, 15, 16, 17, 18, 21, ***22***	IIp, IIq, XII, IX, XIII, XIV, XV, XVI, XVII, XVIII, XXI, XXII	XIV, XV, XVI, XXI[b,c,d]	All Chromosomes[c,d]	NS[c]	NS[c]
Build34 Support[e]	1, 2, 7, 9, ***10***, 15, 16, 17, 18, 21					
RP11-1360M22	1, 2, 7, 9, 13, 14, 15, 16, 18	ND	IX, XIII, XIV, XV, XVI	IIp, XVII, X, XIV, XXII[f]	XIV	XIV
Build34 Support	1, 2, 7, 9, ***10***, 13, 14, 15, 16, ***17***, 18					
RP11-173D3	***1***, 2, 14, 15, 16, 22	XIII, XIV, XV, XVI	IIp, XIII, XIV, XV[c]	XXI, XXII, Y	NS[c]	NS[c]
Build34 Support	2, ***9***, ***10***, 14, 15, 16, 22					
RP11-674M19	***1***, 2, 5, 6, 15, 16, 22	IIp, IX, X, XIV, XV, XVI, XXI, XXII	IIp, IIq, VII, XI, XIII, XV, XVI, XXII	V, X, XXI, XXII	V, X, XVII	V, XVII
Build34 Support	2, 5, ***9***, ***10***, 15, 16, 22					
RP11-492D6	***1***, 2, ***8***, 9, ***13***, ***14***, 16, 22	IIp, V, IX, X, XIII, XIV, XV, XVI, XIX, XXII	IIp, V, VII, IX, XIII, XIV, XV, XVI, XVII	X, XXII, X	XX	XX
Build34 Support	2, ***5***, 9, ***10***, ***15***, 16, ***19***, 22					
RP11-509A17	***1***, 2, 9, 15, 16, 22	IIp, X, XIV, XV, XVI, XXI, XXII	XIII, XV, XVI	XV	XV	XV
Build34 Support	2, 9, ***10***, 15, 16, 22					
RP11-382A4	15, 16	XV, XVI	XV	XV	XV	XV
Build34 Support	15, 16					
RP11-336L20	15	XV	XV	XV	XV, XIV	XV
Build34 Support	15					
RP11-1115P6	2, 13, 14, 15, 18, 21	NS[3]	XV	XV, XXI	XX	XX
Build34 Support	2, 13, 14, 15, 18, 21, ***22***					
RP11-32B5	2, ***11***, 13, ***14***, 15, 18, ***20***, 21	IIq, XIII, XV, XVIII, XXI	XIII, XV	XV	XX	XX
Build34 Support	2, 13, 15, 18, 21, ***22***					
RP11-275E15	2, 13, ***14***, 15, 18, 22	IIp, IIq, XIII, XIV, XV, XVIII, XXI	XIII, XV	IIp, IIq, XIII, XIV, XV, XVIII, XXI	XV	XV
Build34 Support	2, 13, 15, 18, ***21***, 22					

[a] ***Bold italics*** indicates a discordance between the cytogenetic and computational data. ND = Not done. NS = No signal.
[b] Other signals present, but not definitively assigned.
[c] High background, due to low stringency hybridization.
[d] Alpha satellite sequence in RP11-79C23 produced extensive centromeric cross hybridization.
[e] Determined by examination of sequence similarity to the genome assembly (build34), predicted signals are based upon >90% sequence similarity >10 kb in length.
[f] Note the primate X chromosome is designated by "X" and primate chromosome 10 by "X".

duplicated sequences by FISH is lacking this sequence (i.e. a homozygous deletion). The presence of chromosome 10 signals (RP11-674M19, RP11-492D6 and RP11-509A17 – shown in Fig. 2, discussed below) within the chimpanzee genome indicates that this is likely a limitation of the analysis and that the prediction of a chromosome 10 sequence relationship by in silico analysis is indeed correct. These results highlight some of the general limitations to investigating highly duplicated regions of the genome using probes consisting of potentially highly duplicated pericentromeric sequences.

Comparative FISH of non-human primate chromosomes

In order to provide some insight into the evolutionary dynamics of this region during human evolution, we compared the distribution of metaphase FISH signals between human and non-human primates (chimpanzee, gorilla, orangutan, macaque and baboon) for the same underlying human BAC clones (see above). A wide variety of multi-site patterns was observed among non-human primate chromosomes, demonstrating the extremely complex evolutionary history of the human 15q11 pericentromeric region. In general, both qualitative and quantitative differences in the distribution of FISH signals were noted. Also, as more distantly related primate species were examined, a general reduction in the number of signals was observed. While some of these differences may represent loss of signal due to sequence divergence, this observation is consistent with the phylogenetic analyses (see below) which clearly indicate an expansion during great ape evolution. Overall, only five of the 11 BAC clones used in this study yielded results for all primate species examined (Table 1). Generally, two sets of signals can be distinguished based on a compilation of the comparative FISH results: sites shared between closely related species, and putative lineage specific duplication events. It should be noted that the presence of a FISH signal in one species, and not its closest relatives in the panel could also be the result of lineage specific loss in the related species as opposed to a lineage specific duplication. For example, hybridization with BAC RP11-509A17 produced a FISH signal in the orthologous human chromosome 15 region in all species examined (Fig. 2). Hybridization of this clone to PHA, MMU and PPY produced a single signal. In contrast, multiple signals were observed among all African ape lineages examined as GGO exhibited signals on chromosomes XIII, XV and XVI, while in PTR signals were observed on chromosomes IIp, X, XIV, XV, XVI, XXII. Thus, the lack of a chromosome XIII signal in PTR is indicative of a lineage specific duplication in GGO, or alternatively the sequence was lost after the divergence of the gorilla and chimpanzee lineages. The hybridization of RP11-509A17

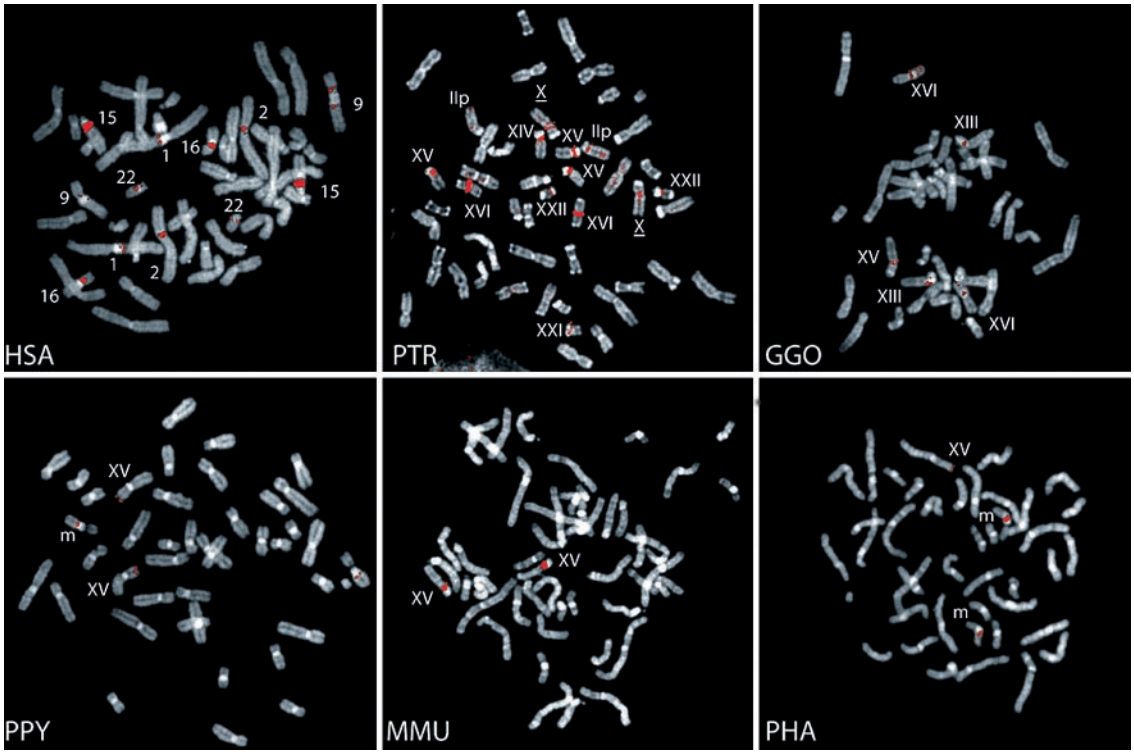


Fig. 2. Non-human primate comparative FISH. An example of the comparative FISH results is shown for 15q11 pericentromeric human BAC clone RP11-509A17. HSA = *Homo sapiens*, PTR = *Pan troglodytes*, GGO = *Gorilla gorilla*, PPY = *Pongo pygmaeus*, MMU = *Macaca mulatta*, PHA = *Papio hamadras*. Note the progressive expansion of the number of interchromosomal FISH signals in GGO, PTR and HSA. Interchromosomal sites of hybridization exclusive to one species, such as the chromosome XIII signal observed in GGO, are indicative of lineage specific rearrangements. Signals labeled with an "m" indicate non-chromosome 15 marker probes used for hybridization controls.

to human metaphase chromosomes, however, produced signals on chromosomes 1, 2, 9, 15, 16 and 22. We believe these results suggest extensive variability in the pericentromeric regions of primate chromosomes.

Phylogenetic analysis

Large-scale genomic sequence analyses of primate DNA have shown that human and lemur non-coding DNA diverge ~15% (Liu et al., 2003; Thomas et al., 2003). Assuming that the majority of the DNA within duplicated regions evolves in a neutral fashion, it is likely that the duplications we analyzed emerged specifically within the primate lineage. Experimental determination of the ancestral origin of the duplicated segments is a tedious process requiring comparative and phylogenetic analysis of each underlying duplicon (Crosier et al., 2002; Guy et al., 2003; Horvath et al., 2003). As a first step in understanding the evolutionary history, however, it is important to delineate the origin of the initial duplication events. Our analysis of such regions over the last seven years has revealed some important trends (Eichler et al., 1996, 1997; Horvath et al., 2000a, 2000b, 2003; Bailey et al., 2002b). First, most of the ancestral loci originate from euchromatic regions and are not associated with pericentromeric DNA. Second, while the ancestral duplications may be contained within larger duplicated pairwise alignments (termed blocks), ancestral loci are usually more divergent and demarcate a smaller segment termed a minimal evolutionary shared segment (Bailey et al., 2002b).

We then applied a mouse-human sequence alignment methodology (see Methods) to predict the ancestral donor locus for 24 duplicons >5 kb within the contig. We were able to identify 19/24 putative donor loci employing this approach, which was consistent with a pilot study of 12 experimentally determined ancestral regions within 2p11 which showed generally good correspondence – 9/12 regions were correctly identified. An example of this analysis is shown for BAC clone RP11-509A17 (Fig. 3), for which the comparative FISH analyses are also shown (Fig. 2). The top panel of Fig. 3 illustrates the result of the ancestral locus prediction using the mouse-human alignment strategy. The prediction appears to be quite effective for low copy duplications, such as the chromosomes 2 and 17 duplicons, which by analysis of the human genome assembly, shown in the bottom panel of Fig. 3, appear to be duplicated only once. Even for intrachromosomal duplications such as the chromosome 15 duplicon derived from the HERC2 locus in 15q13, the mouse-human alignment data correctly pinpoint the ancestral segment. Determining the ancestry of moderately duplicated segments, such as the chromosome 22 duplication noted in Fig. 4 are also approachable with this technique.

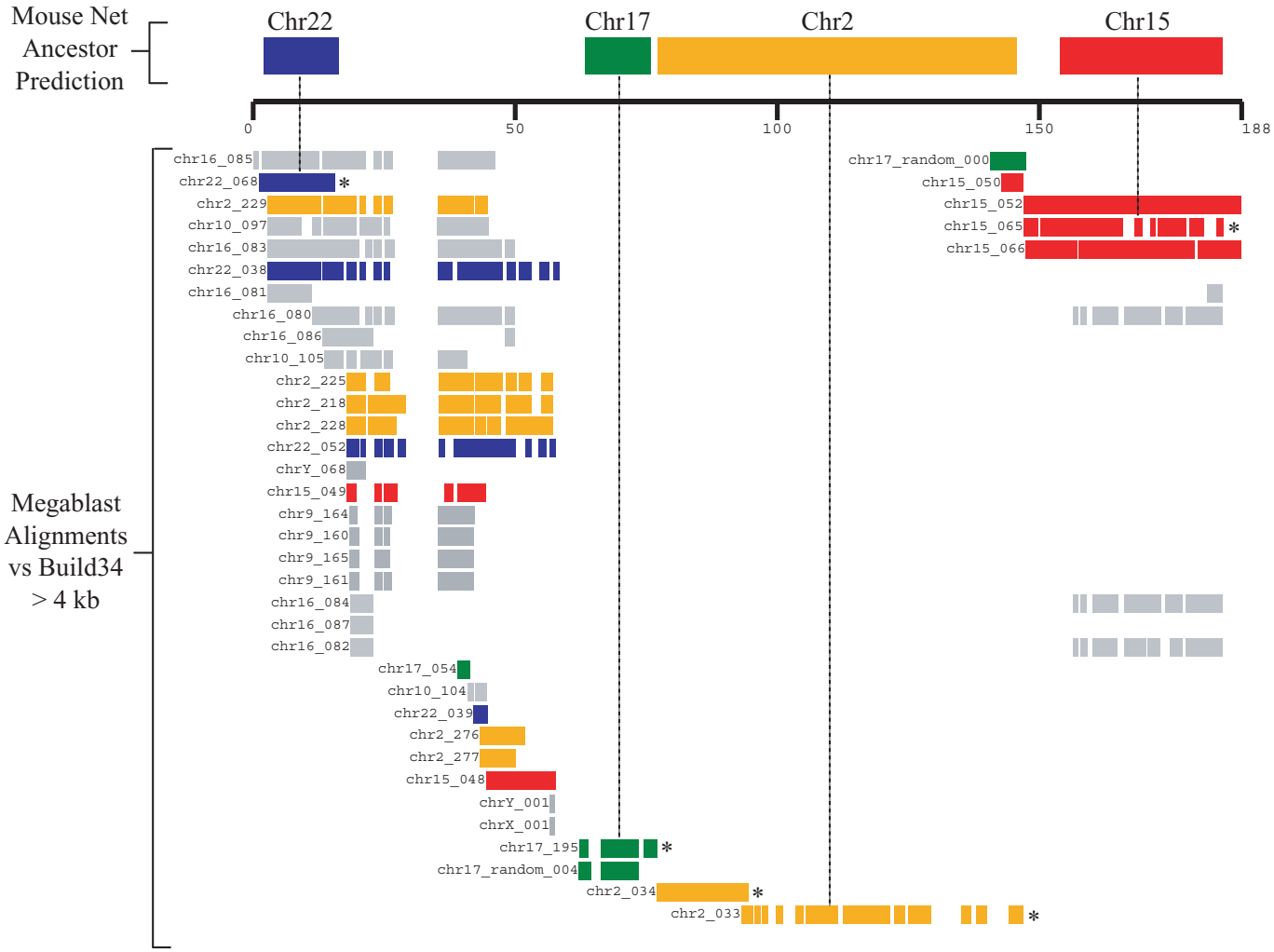


Fig. 3. Duplicon delineation. The diagram depicts the identification of ancestral duplicated segment (duplicon) using two different methods for 15q11 accession AC026495 (BAC RP11-509A17). The results of the Mouse Net determination of the ancestral segment are depicted above the line. Putative ancestral loci are color coded according to the optimal placement within the mouse-human synteny map (Methods). Below the black line, the alignments produced by a sequence similarity search against the most recent human genome assembly (build34, July 2003) are shown. Each alignment is represented by a horizontal box, corresponding to the coordinates of the alignment within the assembly. Global segmental duplication analyses are conducted by fractioning the genome into 400-kb segments, labeled consecutively from p-telomere to q-telomere. Thus the alignments are labeled on the left according to the 400-kb segment of the genome, the alignments to the respective chromosomal random bins are labeled with "random". For the purposes of illustration, only alignments >4 kb are depicted. The alignments with chromosomes other than those determined to be ancestral to a segment within AC026495 are shown in grey. Dotted lines connect the intervals in which ancestry was determined by Mouse Net analysis with the alignments produced by global sequence similarity searches. An asterisk indicates the alignments which agree with the ancestry prediction.

Using a molecular clock calibrated from human-chimp sequence alignments (see Methods) we then estimated the timing of the initial duplication event. Overall, the majority of duplication events appear older than would be suggested simply by the comparative FISH hybridization data. We attribute this to the limited effectiveness of FISH to detect substantially divergent paralogs among species as well as the complex organization of the underlying duplicons contained within each BAC probe. The most proximal half of the 15q11 region consists of relatively younger duplication events which emerged, with one exception, during the separation of the great ape species, 8–15 million years ago (up to and including the 2p24.3 duplicon; Table 2). In contrast, most of the duplications located distally appear to be significantly older, with all but two occurring >15 million years ago. These data generally support previous suggestions of a gradient model with respect to the centromere (Guy et al., 2000; Horvath et al., 2001) – younger evolutionary duplications occur near the centromere while more ancient ones accrue distally. It should be emphasized that these data, however, do not necessarily indicate precisely when the duplicons emerged on chromosome 15 but rather when the initial duplication occurred from an ancestral euchromatic to a pericentromeric region. Several studies have shown that pericentromeric duplications occur in a step-wise fashion, with subse-

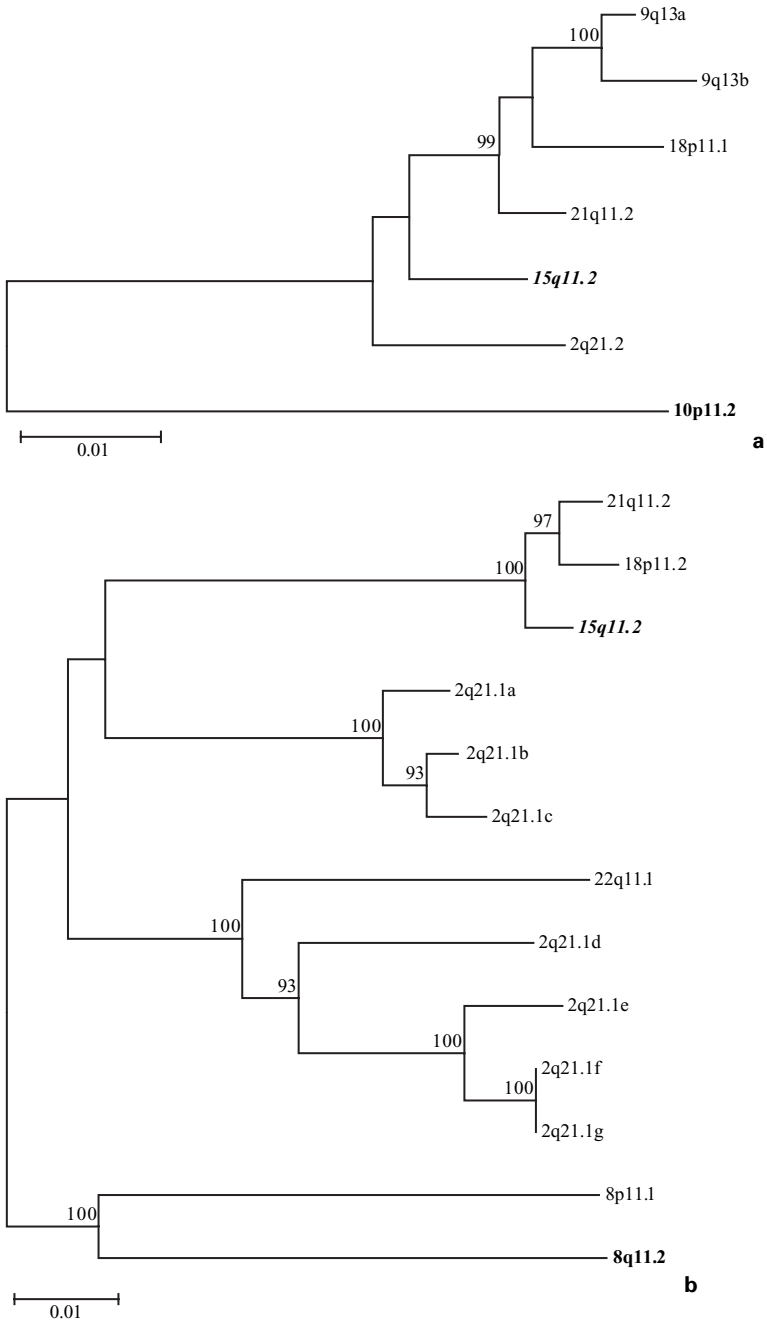


Fig. 4. Phylogenetic analyses. Two examples of neighbor-joining phylogenetic trees of duplicated sequences within the 15q11 pericentromeric contig are presented. The band position of each duplicated segment is indicated at the branch termini. Bootstrap values are placed as near as possible to branch points. (**a**) Analysis of the chromosome 10 duplicon adjacent to the monomeric alpha satellite sequence of the 15q11 pericentromeric contig indicates a progressive swapping of this segment to additional interchromosomal sites. The approximate time of the duplication to chromosome 15 is estimated at 21.7 Mya (Table 2) (**b**) The phylogenetic analysis of the single chromosome 8 duplicon from the 15q11 contig indicates a longer evolutionary history of this segment swapping among multiple sites. The approximate timing of the duplication to chromosome 15 is estimated to be 32.5 Mya. For both phylograms, the putative ancestral locus is indicated in bold, and the sequence derived from the chromosome 15 contig is indicated in bold italics. Bootstrap values >90 are indicated.

quent duplications of larger blocks spreading duplications among pericentromeric regions of the primate genome.

The molecular evolutionary history of the underlying duplicons – especially the sequence relationships between the ancestral duplicon donor sequence and the chromosome 15 duplicon – becomes more apparent when phylogenetic trees are independently considered for each duplicon (Fig. 4). As depicted for the chromosome 10p11.2 duplicon in Fig. 4a the topology of the phylogram indicates there has been a dispersal of this segment in recent evolutionary time to multiple human chromosomes. The data indicate that either 15q11 or the ancestral pericentromeric region of 2q21.2 were the targets of duplicative transposition of the segment from 10p11.2 approximately 20 million years ago. Subsequent duplications of a larger segment (Supplemental Fig. 1, www.karger.com/doi/10.1159/000080804) were responsible for the distribution of this segment within 21q11, 18p11 and 9q13. Similarly, the topology of the 8q11.2 duplicon suggests an equally complex model in which there were more ancient duplications (~ 30 million years ago) of the chromosome 8 segment to chromosomes 2 and 22, and a subsequent swapping of this segment to chromosome 15, 18 and 21 (Fig. 4b). Interestingly, the two segments depicted in Fig. 4,

Table 2. Estimated divergence time of 15q11 pericentromeric duplicons

Ancestral site[a]	Band[b]	Gene content[c]	Mouse net support[d]	Length[e]	k[f]	k [Std. Err.]	T (My)[g]	T [Std. Err.] (My)
chr10:37036374-37037623	10p11.21	ND	Yes	873	0.0833	[0.0085]	21.7	[2.2]
chr14:104360699-104362272	14q32.33	IGHG3	Yes	1573	0.0393	[0.0048]	10.3	[1.3]
chr22:29790232-29788858	22q12.2	ND	Yes	1326	0.0332	[0.005]	8.7	[1.3]
chr22:15399760-15397794	22q11.2	IGL	No	1927	0.0308	[0.0043]	8	[1.1]
chr10:42591455-42593351	10q11.21	BMS1L	Yes	1895	0.0364	[0.0041]	9.5	[1.1]
chr5:61563805-61568337	5q12.1	ND	Yes	4516	0.0230	[0.0022]	6	[0.6]
chr19:38173059-38174945	19q13.11	RHPN2	Yes	1564	0.0314	[0.0041]	8.2	[1.1]
chr22:27410306-27408671	22q12.2	CHEK2	Yes	1531	0.0340	[0.0045]	8.9	[1.2]
chr22:15399760-15397794	22q11.2	IGL	No	1927	0.0307	[0.0048]	8	[1.3]
chr17:78386246-78385225	17q25.3	ND	Yes	1019	0.0616	[0.0079]	16.1	[2.1]
chr2:13549006-13547180	2p24.3	ND	Yes	1813	0.0521	[0.0059]	13.6	[1.5]
chr15:26073457-26074977	15q13.1	HERC2	Yes	2370	0.0145	[0.0026]	3.8	[0.7]
chr15:60276902-60277780	15q22.2	GOLGAG6	Yes	795	0.1469	[0.0145]	38.3	[3.8]
chr9:85962454-85960996	9q22.1	FLJ35866 (C9orf79)	No	1447	0.1014	[0.0079]	26.5	[2.1]
chr14:104700190-104702407	14q32.33	IGHG3	No	2201	0.1445	[0.0094]	37.7	[2.5]
chr13:33482174-33479327	13q13.3	NBEA (BCL8)	Yes	2665	0.0657	[0.0051]	17.1	[1.3]
chr2:166048329-166047349	2q24.3	ND	Yes	980	0.026	[0.0049]	6.8	[1.3]
chr3:197883141-197881708	3q29	PAK2	Yes	1783	0.0477	[0.0049]	12.4	[1.3]
chr22:28618704-28619790	22q12.2	MTMR3	Yes	1077	0.0847	[0.0087]	22.1	[2.3]
chr2:94949176-94950408	2q11.1	ND	Yes	1193	0.1395	[0.0118]	36.4	[3.1]
chr8:43187959-43186644	8p11.1	ND	Yes	1239	0.1244	[0.011]	32.5	[2.9]
chr3:77757145-77754264	3p12.3	ND	Yes	2837	0.0795	[0.0052]	20.7	[1.4]
chr15:24657812-24658781	15q12	GABRA5	Yes	950	0.1071	[0.0114]	27.9	[3]
chr17:29732735-29730351	17q11.2	NF1	No	2102	0.0653	[0.0055]	17	[1.4]

[a] Coordinates from the ancestral locus used in phylogenetic analysis, from build34 (July 2003) of the human genome assembly. Ancestral position determined by mouse-human synteny (Methods and materials) and examination of minimaly evolutionary shared segments (MESS).
[b] Band position according to UCSC Genome Browser (build34).
[c] Gene content identified by previous reports in the literature, or examination of the ancestral interval in the UCSC Genome Browser (build34). ND = Not determined.
[d] Mouse net support – unambiguous "ancestral" locus identified based on mouse-human synteny (Methods and materials).
[e] Length is defined as the number of intron sites evaluated by pairwise analysis of the chromosome 15 and ancestral loci.
[f] k = number of single bp changes/bp of the alignment of chromosome 15 and ancestral sequences, presented with the associated standard error.
[g] Time of divergence in millions of years (My). Divergence time was calculated using the formula R = k/2T. R, the rate of nucleotide change, was determined by the alignment of human and chimpanzee sequence from two duplicons in the 15q11 contig (Methods and materials). The associated standard error is indicated in the adjacent column.

which are approximately a megabase apart within the human 15q11 contig, show a similar interchromosomal distribution to chromosomes 2, 18 and 21. In addition to gaining insight into the individual history of each duplicon, the results of the phylogenetic analysis provide further support for the identification of ancestral loci. Until large-scale comparative sequence for these regions is obtained from non-human primate species, such complex movements will remain untested hypotheses. Definitive phylogenies which can parsimoniously track the evolutionary history of the basic elements of the duplication mosaic will require directed comparative studies within these regions. Our analyses, however, clearly indicate that such regions represent a rich resource for understanding the natural pattern of primate genetic variation.

References

Arnold N, Wienberg J, Emert K, Zachau H: Comparative mapping of DNA probes derived from the Vk immunoglobulin gene regions on human and great ape chromosomes by fluorescence in situ hybridization. Genomics 26:147–156 (1995).

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

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: Recent segmental duplications in the human genome. Science 297:1003–1007 (2002a).

Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, Archidiacono N, Schwartz S, Rocchi M, Eichler EE: Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. Am J Hum Genet 70:83–100 (2002b).

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE: Hotspots of mammalian chromosomal evolution. Genome Biol 5:R23 (2004).

Barber JC, Cross IE, Douglas F, Nicholson JC, Moore KJ, Browne CE: Neurofibromatosis pseudogene amplification underlies euchromatic cytogenetic duplications and triplications of proximal 15q. Hum Genet 103:600–607 (1998).

Brun ME, Ruault M, Ventura M, Roizes G, De Sario A: Juxtacentromeric region of human chromosome 21: a boundary between centromeric heterochromatin and euchromatic chromosome arms. Gene 312:41–50 (2003).

Cheung J, Wilson MD, Zhang J, Khaja R, MacDonald JR, Heng HH, Koop BF, Scherer SW: Recent segmental and gene duplications in the mouse genome. Genome Biol 4:R47 (2003).

Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, et al: Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature 409:953–958 (2001).

Crosier M, Viggiano L, Guy J, Misceo D, Stones R, Wei W, Hearn T, Ventura M, Archidiacono N, Rocchi M, Jackson MS: Human paralogs of KIAA0187 were created through independent pericentromeric-directed and chromosome-specific duplication mechanisms. Genome Res 12:67–80 (2002).

Dyomin VG, Chaganti SR, Dyomina K, Palanisamy N, Murty VV, Dalla-Favera R, Chaganti RS: BCL8 is a novel, evolutionarily conserved human gene family encoding proteins with presumptive protein kinase A anchoring function. Genomics 80:158–165 (2002)

Eichler EE, Lu F, Shen Y, Antonacci R, Jurecic V, Doggett NA, Moyzis RK, Baldini A, Gibbs RA, Nelson DL: Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. Hum Mol Genet 5:899–912 (1996).

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

Eichler E, Archidiacono N, Rocchi M: CAGGG repeats and the pericentromeric duplication of the hominoid genome. Genome Res 9:1048–1058 (1999).

Fantes JA, Mewborn SK, Lese CM, Hedrick J, Brown RL, Dyomin V, Chaganti RS, Christian SL, Ledbetter DH: Organisation of the pericentromeric region of chromosome 15: at least four partial gene copies are amplified in patients with a proximal duplication of 15q. J Med Genet 39:170–177 (2002).

Footz TK, Brinkman-Mills P, Banting GS, Maier SA, Riazi MA, Bridgland L, Hu S, et al: Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere. Genome Res 11:1053–1070 (2001).

Golfier G, Chibon F, Aurias A, Chen XN, Korenberg J, Rossier J, Potier MC: The 200-kb segmental duplication on human chromosome 21 originates from a pericentromeric dissemination involving human chromosomes 2, 18 and 13. Gene 312:51–59 (2003).

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

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

Guy J, Hearn T, Crosier M, Mudge J, Viggiano L, Koczan D, Thiesen HJ, Bailey JA, Horvath JE, Eichler EE, Earthrowl ME, Deloukas P, French L, Rogers J, Bentley D, Jackson MS: Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. Genome Res 13:159–172 (2003).

Hillier LW, Fulton RS, Fulton LA, Graves TA, Pepin KH, Wagner-McPherson C, Layman D, et al: The DNA sequence of human chromosome 7. Nature 424:157–164 (2003).

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

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

Horvath JE, Bailey JA, Locke DP, Eichler EE: Lessons from the human genome: transitions between euchromatin and heterochromatin. Hum Mol Genet 10:2215–2223 (2001).

Horvath JE, Gulden CL, Bailey JA, Yohn C, McPherson JD, Prescott A, Roe BA, De Jong PJ, Ventura M, Misceo D, Archidiacono N, Zhao S, Schwartz S, Rocchi M, Eichler EE: Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications. Mol Biol Evol 20:1463–1479 (2003).

IHGSC: Initial sequencing and analysis of the human genome. Nature 409:860–921 (2001).

ISCN (1985): An international system for human cytogenetic nomenclature: Report of the standing committee on human cytogenetic nomenclature. Birth Defects 21:1–117 (1985).

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, Spalluto C: Characterisation of the heterochromatin/euchromatin boundary at 10q11 and identification of novel transcripts by repeat induced instability. Am J Hum Genet 65 (suppl):A56 (1999).

Kehrer-Sawatzki H, Schwickardt T, Assum G, Rocchi G, Krone W: A third neurofibromatosis type 1 (NF1) pseudogene at chromosome 15q11.2. Hum Genet 100:595–600 (1997).

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: Evolution's cauldron: duplication, deletion and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100:11484–11489 (2003).

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

Kumar S, Tamura K, Nei M: MEGA: Molecular Evolutionary Genetic Analysis, version 1.0. (Pennsylvania State University, University Park 1993).

Liu G, Zhao S, Bailey JA, Sahinalp SC, Alkan C, Tuzun E, Green ED, Eichler EE: Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. Genome Res 13:358–368 (2003).

Orti R, Potier MC, Maunoury C, Prieur M, Creau N, Delabar JM: Conservation of pericentromeric duplications of a 200-kb part of the human 21q22.1 region in primates. Cytogenet Cell Genet 83:262–265 (1998).

Regnier V, Meddeb M, Lecointre G, Richard F, Duverger A, Nguyen VC, Dutrillaux B, Bernheim A, Danglot G: Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. Hum Mol Genet 6:9–16 (1997).

Ritchie RJ, Mattei MG, Lalande M: A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. Hum Mol Genet 7:1253–1260 (1998).

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF: Genomic and genetic definition of a functional human centromere. Science 294:109–115 (2001).

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: Human-mouse alignments with BLASTZ. Genome Res 13:103–107 (2003).

Smit A, Green P: RepeatMasker (ftp.genome.washington.edu/RM/RepeatMasker.html) (1999).

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, et al: Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793 (2003).

Thompson JD, Higgins DG, Gibson TJ: 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 (1994).

Tomlinson IM, Cook GP, Carter NP, Elaswarapu R, Smith S, Walter G, Buluwela L, Rabbitts TH, Winter G: Human immunglobulin VH and D segments on chromosomes 15q11.2 and 16p11.2. Hum Mol Genet 3:853–860 (1994).

Zhang Z, Schwartz S, Wagner L, Miller W: A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214 (2000).

Zimonjic D, Kelley M, Rubin J, Aaronson S, Popescu N: Fluorescence in situ hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. Proc Natl Acad Sci USA 94:11461–11465 (1997).