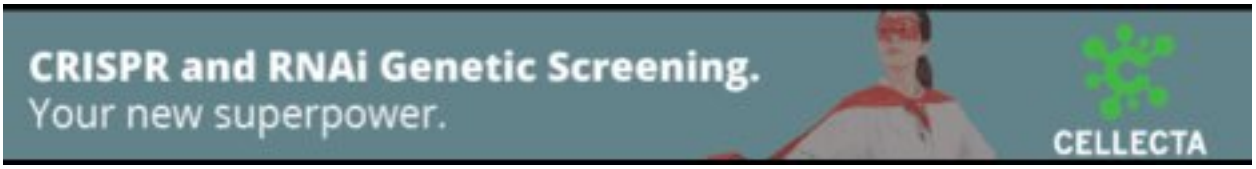
Punctuated duplication seeding events during the evolution of human chromosome 2p11

Julie E. Horvath, Cassandra L. Gulden, Rhea U. Vallente, et al.

Genome Res. 2005 15: 914-927
Access the most recent version at doi:10.1101/gr.3916405



Letter

Punctuated duplication seeding events during the evolution of human chromosome 2p11

Julie E. Horvath,[1,5] Cassandra L. Gulden,[1] Rhea U. Vallente,[1,4] Marla Y. Eichler,[1] Mario Ventura,[2] John D. McPherson,[3] Tina A. Graves,[3] Richard K. Wilson,[3] Stuart Schwartz,[1] Mariano Rocchi,[2] and Evan E. Eichler[1,6,7]

[1]*Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA;* [2]*Sezione di Genetica, DAPEG, University of Bari, 70126 Bari, Italy;* [3]*Washington University School of Medicine Genome Sequencing Center, St. Louis, Missouri 63108, USA;* [4]*Washington State University School of Molecular Biosciences, Pullman, Washington 99164, USA;* [5]*Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina 27708, USA;* [6]*Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA*

Primate genomic sequence comparisons are becoming increasingly useful for elucidating the evolutionary history and organization of our own genome. Such studies are particularly informative within human pericentromeric regions—areas of particularly rapid change in genomic structure. Here, we present a systematic analysis of the evolutionary history of one ~700-kb region of 2p11, including the first autosomal transition from pericentromeric sequence to higher-order α-satellite DNA. We show that this region is composed of segmental duplications corresponding to 14 ancestral segments ranging in size from 4 kb to ~115 kb. These duplicons show 94%–98.5% sequence identity to their ancestral loci. Comparative FISH and phylogenetic analysis indicate that these duplicons are differentially distributed in human, chimpanzee, and gorilla genomes, whereas baboon has a single putative ancestral locus for all but one of the duplications. Our analysis supports a model where duplicative transposition events occurred during a narrow window of evolution after the separation of the human/ape lineage from the Old World monkeys (10–20 million years ago). Although dramatic secondary dispersal events occurred during the radiation of the human, chimpanzee, and gorilla lineages, duplicative transposition seeding events of new material to this particular pericentromeric region abruptly ceased after this time period. The multiplicity of initial duplicative transpositions prior to the separation of humans and great-apes suggests a punctuated model for the formation of highly duplicated pericentromeric regions within the human genome. The data further indicate that factors other than sequence are important determinants for such bursts of duplicative transposition from the euchromatin to pericentromeric regions.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. AY954301–AY954363.]

Human pericentromeric and subtelomeric regions, much like the majority of the Y chromosome, have long been viewed by many as "genetic wastelands" (Skaletsky et al. 2003) due to the fact that they are composed of large complex blocks of heterochromatic sequences and contain few genes (Donze and Kamakaka 2002). Recent studies suggest that understanding these transition regions will provide us a more complete picture of human genome architecture and the relationship of chromosome structure and function (She et al. 2004a). Despite recent advances in genome sequencing and the finishing of human euchromatin (International Human Genome Sequencing Consortium [IHGSC] 2004), the structure of these regions remains largely incomplete (Eichler et al. 2004). Sequence gaps are particularly enriched within pericentromeric regions, and most chromosome sequences fall short of bridging classically defined (Manuelidis 1978; Willard and Waye 1987; Willard 1991) heterochromatic sequences and euchromatin.

[7]Corresponding author.
E-mail eee@gs.washington.edu; fax (206) 685-7301.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3916405. Article published online before print in June 2005.

More recently, a handful of laboratories have extended efforts to include heterochromatic transition regions (Bailey et al. 2001; IHGSC 2001; Schueler et al. 2001; Rudd and Willard 2004; She et al. 2004a). From these and other efforts, we now understand that more than half of all human chromosomes contain segmentally duplicated sequences, primarily found in pericentromeric or subtelomeric regions. A noticeable reduction in transcription is observed within the most proximal 1 Mb portion of the duplication region, suggesting that some heterochromatic properties extend beyond α-satellite DNA. These duplications range in size from 1 kb to more than half a megabase and typically originate from euchromatic regions of the genome (She et al. 2004a). A few pericentromeric duplications have been characterized in detail, although the mechanism for their dispersal is still largely unknown (Guy et al. 2000, 2003; Ji et al. 2000; Bailey et al. 2001; Horvath et al. 2001; Samonte and Eichler 2002). A highly nonrandom distribution of duplications within pericentromeric regions has been noted with both quiescent and active regions of duplication for specific human chromosomes (She et al. 2004a).

Limited comparisons of pericentromeric regions among

15:914–927 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

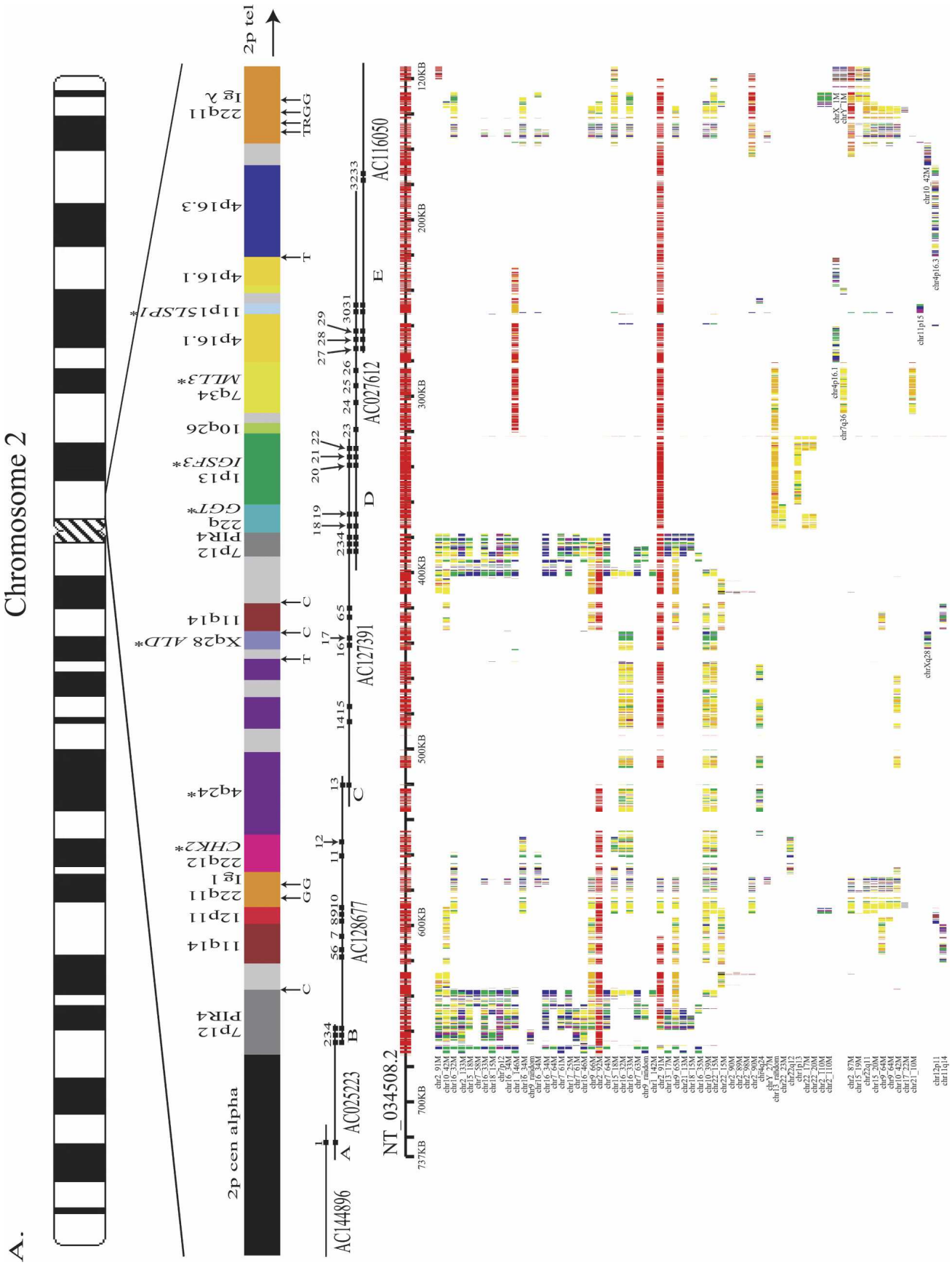


Figure 1. (Continued on next page)

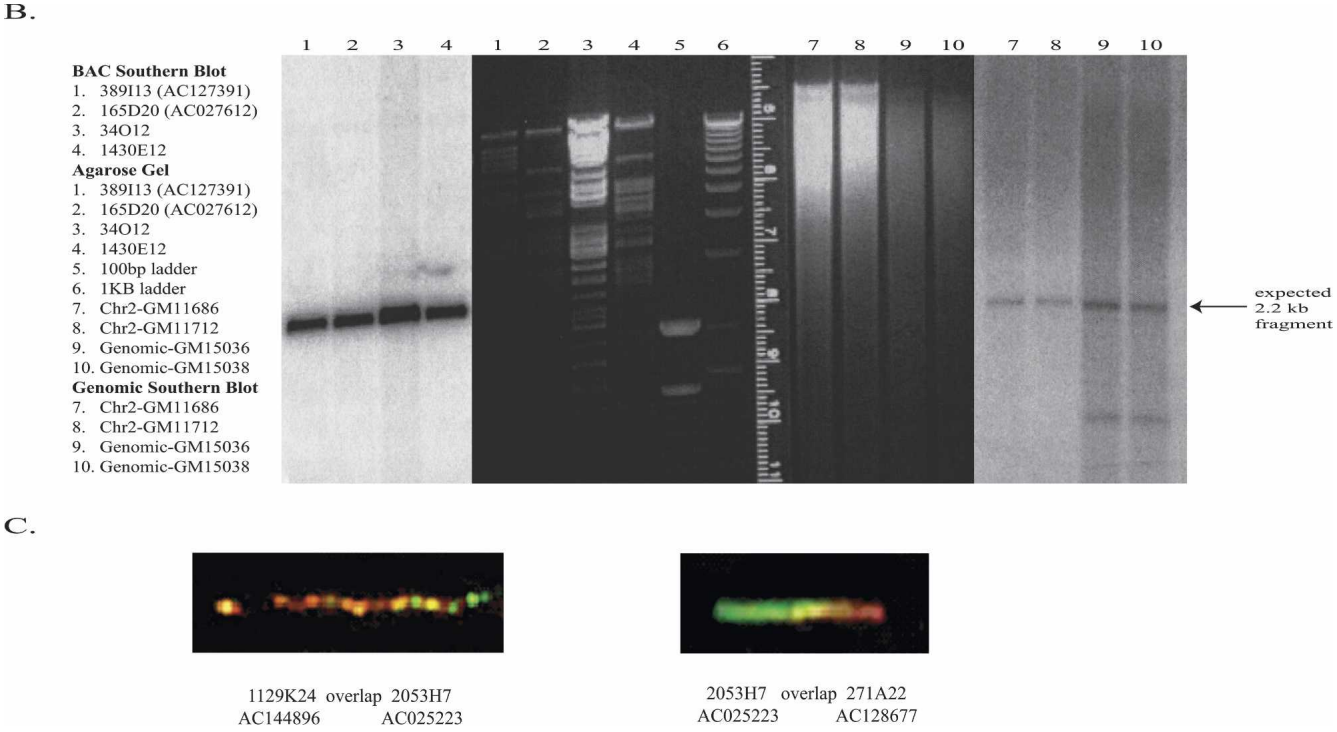


Figure 1. 2p11 Duplicon architecture. (*A*) A schematic representation of the duplicon architecture (colored bars) is shown in reference to an ideogram of chromosome 2 and ~700-kb BAC minimal tiling path. The black bar represents α-satellite sequence (~175 kb), while light gray bars denote various pericentromeric-specific interspersed repeats (PIRs). Other enriched pericentromeric repeat sequences are indicated: C=CAAAAAG repeat, G=CAGGG, R=REP522, and T=TAR1 repeats (Smit 1996). Below the BAC tiling path are results of database searches using this entire sequence (represented by NT_034508) against the human genome (build34, July 2003). All pairwise alignments (>5 kb and >90%) to this segment are shown to other regions of the genome as indicated by the chromosome number and approximate position in megabases (ancestral loci are denoted by cytogenetic band position). A color scheme encodes the average percentage sequence identity for each alignment block (red, 99%; orange, 98–99%; yellow, 97–98%; green, 96–97%; blue, 95–96%; indigo, 94–95%; and violet, 90–94%). (*B*) Sequence overlaps were confirmed by Southern analysis between BAC clone and genomic DNA. An example of validation is shown for overlap D (between AC127391 [R11–389I13] and AC027612 [R11–165D20]). A PCR-generated probe (165D20–6n7) (Supplemental Table 2) was hybridized. The expected 2.2-kb band is observed in multiple overlapping BACs (389I13, 165D20, 34O12, and 1430E12) in addition to the chromosome 2 hybrid and genomic DNA samples. Note: An additional lower band is observed in the genomic DNA samples compared with the monochromosomal hybrid DNA samples, indicating that at least one additional copy of the GGT1 duplicon exists within the human genome. (*C*) Extended fiber FISH validating overlap (in yellow) of the three most proximal BACs in a chromosome 2 hybrid cell line (GM11712). Results in a second chromosome 2 hybrid line (GM11686) and total human cell lines showed similar results (data not shown).

closely related primates suggest extraordinary dynamism where duplication, deletion, and rearrangement of large segments of DNA occur at an unprecedented scale (Eichler et al. 1996, 1997; Regnier et al. 1997; Zimonjic et al. 1997; Orti et al. 1998; Horvath et al. 2000b, 2003; Crosier et al. 2002). These findings have suggested that the actual number of "chromosomal rearrangements" among primates far exceed expectations based on the comparison of primate karyotypes. Limited phylogenetic analyses of a small number of segmental duplications (Eichler et al. 1997; Orti et al. 1998; Horvath et al. 2000b; Luijten et al. 2000) support a two-step model for their origin whereby initial rounds duplicate portions of the euchromatin to a specific pericentromeric "acceptor region." Subsequent duplication events move larger blocks of duplication (often made of several blocks of initial duplication) among the acceptor regions.

In an effort to provide insight into these complex regions of our genome, we conducted a detailed molecular evolutionary analysis of a 700-kb pericentromeric region of human chromosome 2p11. This human chromosome is particularly remarkable since it contains a large number of highly identical inter- and intrachromosomal segmental duplications. It is also noteworthy as the only chromosome to have emerged in the human lineage as a result of a chromosome fusion (Ijdo et al. 1991; Fan et al. 2002). There were two main objectives of this research: (1) to characterize the organization of the 2p11 pericentromeric region up to and including higher-order α-satellite repeats and (2) to assess the evolutionary origin and the timing of the duplication events in primate evolution. Our previous pilot analysis of 2p11 indicated that this type of organization was a property common to many pericentromeric regions. Therefore, 2p11 provides a model for the organization of many human pericentromeric regions containing interchromosomal duplications, and gives us insight into the general mechanism for their formation.

Results

Sequence, assembly, validation, and annotation of the 2p11 pericentromeric region

We constructed a physical map and sequenced 700 kb of the most proximal portion of the short arm of human chromosome 2. The presence of high-identity duplications to multiple regions of the human genome complicates sequence and assembly of these regions (She et al. 2004b). The organization and represen-

Table 1. 2p11 Duplicon sequence properties

Ancestral location	Duplicon name	Build 34 alignments	Ancestral chromosome	Build 34 begin	Build 34 end	Paralogous chr2	Build 34 begin	Build 34 end	Alignment length (bp)	Percent identity	Repeat
7p12	pir4	50	chr7	52931091	52953785	chr2	91716578	91738986	23046	94.0	
11q14	11q	7	chr11	89568972	89571044	chr2	91699117	91701188	2091	95.1	
11q14	11q	9	chr11	89519459	89539156	chr2	91679227	91699086	21099	94.1	
12p11	12p	8	chr12	27307033	27314333	chr2	91673194[a]	91679108	7350	94.9	*Alu*Sq
22q11	Igλ	32	chr22	20948276	20976056	chr2	91651635	91673445[a]	28411	95.8	
22q12	*CHK2*	13	chr22	27396207	27415189	chr2	91629693[b]	91651630	22155	95.2	
4q24	4q24	5	chr4	104340868	104356278	chr2	91612384	91629698[b]	17360	95.6	
4q24	4q24	5	chr4	104329872	104338171	chr2	91606549[a]	91612376	8365	95.8	
4q24	4q24	5	chr4	104322271	104328805	chr2	91600337	91606554[a]	6691	95.8	L1PA15, *Alu*Sg
4q24	4q24	7	chr4	104276520	104319279	chr2	91551291	91593954	42964	95.8	
4q24	4q24	6	chr4	104266506	104276513	chr2	91531378	91541381	10067	96.4	
Xq28	*ALD*	5	chrX	151473852	151483609	chr2	91513314	91523079	9842	95.7	
11q14	11q	9	chr11	89525311	89539156	chr2	91499241	91512571	14527	95.2	
11q14	11q	7	chr11	89568972	89571044	chr2	91497139	91499210	2091	95.1	
7p12	PIR4	50	chr7	52931091	52955315	chr2	91457832	91481757	24575	94.0	
22q11	*GGT1*	5	chr22	23337169	23343213	chr2	91449527[a]	91455557	6051	96.7	
22q11	*GGT1*	5	chr22	23322945	23332600	chr2	91441101	91449696[a]	9669	97.2	*Alu*Sg
1p13	*IGSF3*	2	chr1	116476821	116495886	chr2	91422050	91441097	19108	97.5	
1p13	*IGSF3*	4	chr1	116495873	116512648	chr2	91402369[a]	91419084	16828	97.1	
10q26	10q26	4	chr10	127182300	127187515	chr2	91397399	91402574[a]	5259	97.0	*Alu*Sc
7q36	*MLL3*	7	chr7	151300710	151330205	chr2	91360615[a]	91390058	29629	97.1	
4p16.1	4p16.1	2	chr4	9531944	9552677	chr2	91340328	91360773[a]	21014	95.3	*Alu*Y
11p15	11p15	2	chr11	1859940	1863883	chr2	91328768	91332730	3973	94.8	
11p15	11p15	2	chr11	1871783	1872884	chr2	91327670	91328767	1101	96.8	
7q36	*MLL3*	2	chr7	151386594	151392806	chr2	91318140	91324329	6226	96.9	
4p16.1	4p16.1	2	chr4	9512955	9532432	chr2	91301404	91318132	95141	99.2	
4p16.3	4p16.3	1	chr4	4248063	4268960	chr2	91280163[a]	91300737	21155	95.1	
4p16.3	4p16.3	1	chr4	4274738	4306539	chr2	91248931	91280164[a]	32160	95.2	L1MB2
22q11	Igλ	36	chr22	20936002	20979439	chr2	91200069	91238134	44479	97.2	

Ancestral positions of all duplicons > 5 kb in total are indicated.
[a]Chr2 end position of one duplicon overlaps with the chr2 begin position of the adjacent duplicon due to termination within a repeat element (LINE or SINE).
[b]Chr2 begin and end positions overlapped by 5 bp.
Putative ancestral chromosomal location and copy number (build34) of all duplicons found within 2p11 are shown. The percentage identity between the ancestral and duplicated copy on chromosome 2p11 was calculated. Occasionally, a duplicon termination point occurs within a high copy repeat (as indicated on the far right) and appears to overlap another duplicon termination point as a result.

Table 2. Duplicon junctions

		JUNCTION 1						JUNCTION 2					
Map location	Duplicon name	Donor chromosome	Donor begin position	Repeat type	Acceptor chromosome	Acceptor begin position	Repeat type	Donor chromosome	Donor end position	Repeat name	Acceptor chromosome	Acceptor end position	Repeat type
7p12	PIR4	chr7	52931091	none	chr2	91716578	none	chr7	52953785	L1MEb	chr2	91738986	none
11q14	11q	chr11	89568972	L2	chr2	91699117	*Alu*	chr11	89539156	*AluY*	chr2	91699086	*Alu*
12p11	12p	chr12	27307033	*Alu*Sx	chr2	91673194	*Alu*Sq	chr12	27314333	none	chr2	91679108	(TA)n[a]
22q11	*Igλ*	chr22	20948276	*Alu*Sq	chr2	91651635	*Alu*Jb L1MB7[a]	chr22	20976056	L1	chr2	91673445	*Alu*Sq
22q12	*CHK2*	chr22	27396207	*Alu*Jb	chr2	91629693	none	chr22	27415189	none	chr2	91651630	*Alu*Jb[a]
4q24	4q24	chr4	104340868	*Alu*Sp	chr2	91612384	*Alu*Sp[a]	chr4	104276513	L1M4	chr2	91541381	L1M4
Xq28	*ALD*	chrX	151473852	none	chr2	91513314	none	chrX	151483609	L3b	chr2	91523079	*AluYc*[a]
11q14	11q	chr11	89525311	L1ME	chr2	91499241	*Alu*Yc5	chr11	89571044	*Alu*Sp	chr2	91499210	*Alu*Yc5
7p12	PIR4	chr7	52931091	none	chr2	91457832	none	chr7	52955315	none	chr2	91481757	none
22q11	*GGT1*	chr22	23337169	*AluY*	chr2	91449527	*Alu*Sg	chr22	23332600	*Alu*Sg	chr2	91449696	*Alu*Sg
1p13	*IGSF3*	chr1	116476821	none	chr2	91422050	none	chr1	116512648	*Alu*Sx[a]	chr2	91419084	none
10q26	10q26	chr10	127182300	*Alu*Sx	chr2	91397399	*Alu*Sg/x	chr10	127187515	*Alu*Sc	chr2	91402574	*Alu*Sc
7q36	*MLL3*	chr7	151300710	*AluY*	chr2	91360615	*AluY*[a]	chr7	151330205	*Alu*Sg	chr2	91390058	*Alu*Sx
4p16.1	4p16.1	chr4	9531944	*AluY*	chr2	91340328	LTR16B	chr4	9552677	LTR16B	chr2	91360773	*AluY*
11p15	*LSP1*	chr11	1859940	none	chr2	91328768	none	chr11	1872884	none	chr2	91328767	none
7q36	*MLL3*	chr7	151386594	L1MEc	chr2	91318140	*Alu*Sx[a]	chr7	151392806	*Alu*Sx	chr2	91324329	none
4p16.1	4p16.1	chr4	9512955	HERVE[a]	chr2	91301404	*Alu*Sx[a]	chr4	9532432	L2	chr2	91318132	*Alu*Sx[a]
4p16.3	4p16.3	chr4	4248063	*Alu*Sp	chr2	91280163	L1MB2[a]	chr4	4306539	A rich	chr2	91280164	L1MB2[a]
22q11	*Igλ*	chr22	20936002	none	chr2	91200069	none	chr22	20979439	L1MDb	chr2	91238134	L1MDa

[a]indicates position of repeat end.
"A rich" signifies a run of nine or more A nucleotides.
The type of repeat within 5 bp on either side of each duplicon junction was assessed using RepeatMasker. Corresponding donor (ancestral) and acceptor (2p11) junctions were assigned based on the sequence alignment. *Alu*S and *Alu*Y repeats are shaded gray. None, no repeat element detected within the range.

Table 3. Comparative FISH results

Duplicon location	Duplicon name	Probe type	Probe name	HSA	PTR	GGO	PPY	PHA	MFA
7p12[b]	PIR4	22 cosmid	N20B5	1p, 1q, 2cen, 7cen, 9p, 10q, 13p, 14p, 15p, 16p, 17p, 18p, 21q, 22q	1p, 2p, 2q, 7cen, 10cen, 16cen	16p	7q	no signal	nd
11q14	11q	11 cosmid	25122	11q	11	11	11, 14, 21	no signal	nd
12p11	12p	nd	nd	nd	nd	nd	nd	nd	nd
22q11	Igλ	nd	nd	nd	nd	nd	nd	nd	nd
22q12	*CHK2*	22 cosmid	N5e10	16p, 22	16p, 22	16p, 16q, 22	21, 22	no signal	nd
4q24[a]	4q24	RPCI-11 BAC	289C17	2p11, 4q24, 14p11, 10q11, 16p11, 22q11, Yq11	4q24, 10q11, 16p11, 22q11	4q24, 10q11, 16p11	4q24	nd	4q24
xq28[a]	*ALD*	LR-PCR probe	ALD9.7	2p11, 10q11, 16p11, 22q11, Xq28	10q11, 16p11, 22q11, Xq28	2p11, 10q11, 14q11, 16p11(2x), Xq28	Xq28	nd	Xq28
22q11	*GGT1*	22 cosmid	N24g22	22q	22q	2p, 2q, 9, 21, 22q	2q, 22q	22	nd
1p13	*IGSF3*	1 cosmid	AH27b17	1p, 2, 13	1, 2p	1, 2p, 9, 13, 15, 18, 21, 22	1	nd	1
10q26	10q26	nd	nd	nd	nd	nd	nd	nd	nd
7q36	*MLL3*	7 cosmid	Y54g1	1q, 2q, 7qter, 13 (1 signal), 21	7 ter	9	7q, 7qter	no signal	nd
4p16.1	4p16.1	RPCI-11 BAC	751L19	4p16.1	nd	nd	4p16.1	4p16.1	nd
11p15	*LSP1*	nd	nd	nd	nd	nd	nd	nd	nd
4p16.3	4p16.3	RPCI-11 BAC	265O12	4p16.3	nd	nd	4p16.3	4p16.3	nd

Cosmid, BAC and long-range PCR (LR-PCR) probes corresponding to each duplicon were hybridized to metaphase chromosome preparations of a variety of primate species. FISH results are indicated for HSA (*H. sapiens*), PTR (*P. troglodytes*), GGO (*G. gorilla*), PPY (*P. pygmaeus*), PHA (*P. hamadryas*), and MFA (*M. fascicularis*). An "nd" signifies that hybridization was not done because duplicon length was too short or the copy number was too great to obtain reliable estimates by FISH. A report of "no signal" meant that no hybridization signal was detected after FISH. All signals are pericentromeric unless noted. All chromosomal designations indicated are with respect to the human phylogenetic group (McConkey, 2004). Information for duplicons 4q24, and ALD was previously published in Horvath et al., 2000b,[a] and PIR4 results were previously published in Horvath et al., 2003.[b]

tation of human chromosome 2 was, therefore, validated by several independent methods, including analysis of sequence overlaps (see tiling path in Fig. 1A), genomic Southern blot analysis (Fig. 1B), two-color FISH experiments (Fig. 1C), and paralogous sequence tagging of monochromosomal DNA (Supplemental Table 1; Supplemental Methods). The assembled sequence included some of the largest (175 kb) contiguous transition sequence into human α-satellite DNA. Several lines of evidence indicate that we have successfully traversed higher-order sequences from chromosome 2 (Supplemental Methods; Supplemental Fig. 1A,B).

We annotated the duplication content by using a variety of computational methods. Seven regions with conserved exon/intron structure were identified within the 2p11 sequence although none contained a complete complement of exons as predicted by the full-length transcript. In each case, the full-length gene mapped to another region of the genome. These were termed duplicons (segmental duplications where the ancestral origin can be determined). Since this search for ancestral duplicons was not limited to sequences outside of defined pericentromeric regions (5 Mb around the centromere), we identified two additional duplicons (*GGT1* and *IGSF3*) that were not identified previously (She et al. 2004a). The 2p11 duplicons included *CHK2* (checkpoint kinase 2) from 22q12, an unknown gene from 4q24, *ALD* (adrenoleukodystrophy) from Xq28, *GGT* (γ-glutamyltransferase 1) from 22q11, *IGSF3* (immunoglobulin superfamily 3) from 1p13, *MLL3* (myeloid/lymphoid leukemia 3) from 7q36, and *LSP1* (lymphocyte-specific protein 1) from 11p15. With the exception of *LSP1*, none of these segments showed any evidence of transcription based on sequence similarity searches of human EST databases.

To identify the putative boundaries of each duplication, we examined all underlying pairwise alignments for the entire region by using PARASIGHT (http://humanparalogy.gs.washington.edu/parasight). This allowed us to obtain the minimally shared segment for each region (Bailey et al. 2002) and facilitated the identification of seven more putative duplicons (PIR4, 11q14, 12p11, λ immunoglobulin (Igλ), 10q26, 4p16.1, and 4p16.3) (Fig. 1A; Table 1) within 2p11. Five of these were previously identified by mouse synteny mapping, but two (PIR4 and Igλ) were excluded due to their location within a pericentromeric region (She et al. 2004a). All 14 of the identified duplicons represent duplicated segments from seven different human chromosomes, exhibit 94%–98.5% identity to the putative ancestral loci, and range in size from <4 kb to >115 kb. Three of these duplicons (*IGSF3*, *GGT1*, and *LSP1*) were shown previously to exist on chromosome 2 by FISH, but detailed analyses into their genomic organization or evolutionary histories were lacking (May et al. 1993; Tassone et al. 1995; Saupe et al. 1998; Ruault et al. 1999).

Previous studies have suggested that GC-rich and *Alu* repeat elements are enriched at the boundaries of duplication (Eichler et al. 1999; Horvath et al. 2000a; Chen and Li 2001) and implicated these as playing a role in the process of segmental duplication (Bailey et al. 2003). In this study we were able to distinguish both donor and acceptor loci (phylogenetically and by comparative FISH). Based on sequence comparison to the ancestral locus, we were able to define 38 donor and acceptor boundaries. Analysis of duplicon termini in 2p11 (Fig. 1A) indicates that GC-rich repeat sequences (CAGGG, CAAAAG, TAR, and REP522) (Smit 1996) occur within 1 kb for at least five of 19 of the acceptor regions. No enrichment of these elements was noted in the vicinity of the donor regions. If we narrowed the junctions to a 5-bp window (Table 2), we found that 15 of 38 (39%) of the donor boundaries and 16 of 38 (42%) acceptor regions show the presence of an *Alu* S or Y repeat sequence at the junction. This *Alu* enrichment is consistent with previous reports and suggests that

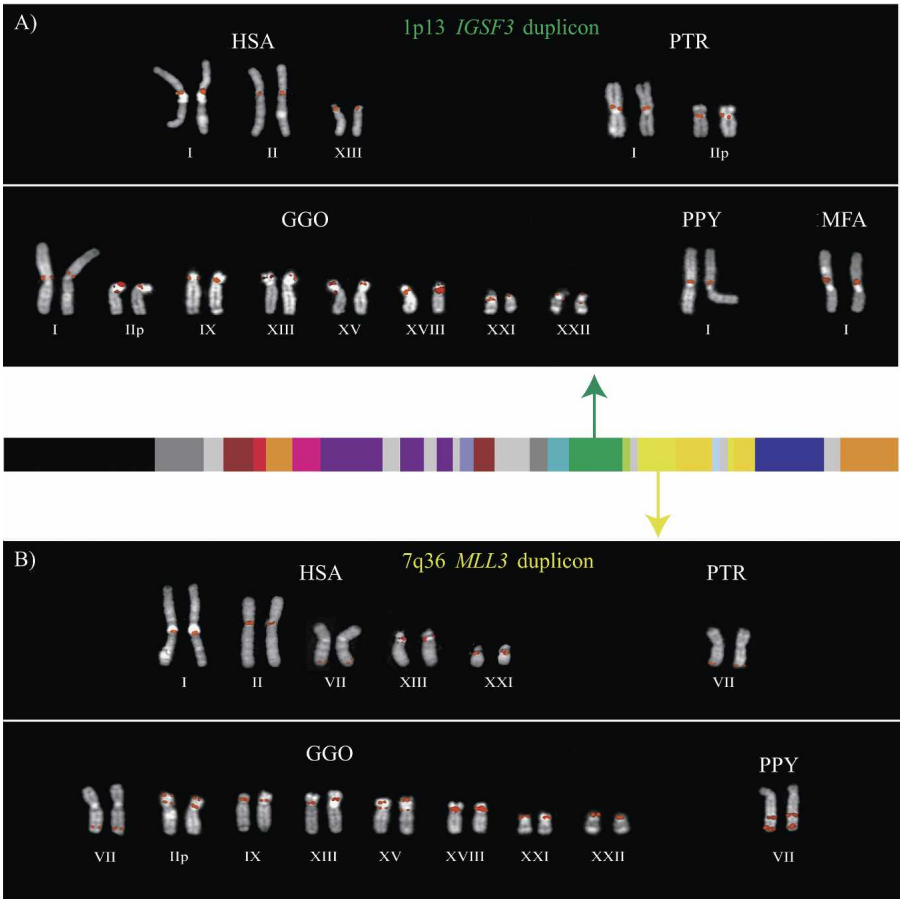


Figure 2. Comparative primate FISH of individual duplicons. Two examples of comparative metaphase FISH experiments for the (*A*) IGSF3 (dark green) duplicon from 1p13 and the (*B*) *MLL3* duplicon (in yellow) from 7q36 are shown. Extracted metaphases for five primates are shown after hybridization with probes corresponding to the two duplicons: HSA indicates *H. sapiens*; PTR, *P. troglodytes*; GGO, *G. gorilla*; PPY, *P. pygmaeus*; and MFA, *M. fascicularis*. Both sets of experiments show multiple signals among humans and the great-apes with a single signal in the Old World monkey macaque. These results are consistent with the phylogenetic and comparative genomic hybridization experiments that suggest a duplication of the ancestral locus <23 Mya. All chromosomal designations are with respect to the human phylogenetic group (McConkey 2004).

Alu repeats have played an important role in initializing pericentromeric seeding events while GC-rich elements contribute to the pericentromeric swapping. At present, there is, however, only indirect evidence for such associations.

Evolutionary analysis of 2p11 duplications

A three-pronged approach was used to reconstruct the evolutionary history of this region. Each of the 14 duplicons (defined above) was treated independently in this analysis. Comparative FISH was used to delineate the origin, dispersal, and copy number variation among closely related primate species. Screening of genomic libraries from nonhuman primates was used as a mapping approach to refine ancestral locations of each duplicon based on comparison of the clone ends to the human genome sequence (see below). Phylogenetic analysis of sequence from each duplication was then used to reconstruct the likely order and timing of the individual duplications during the past 25 million years (Myr) of human genome evolution.

We performed comparative FISH against metaphase chromosomes of four hominoid species (*Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, and *Pongo pygmaeus*) and one Old World monkey representative (*Papio hamadryas* or *Macaca fascicularis*). Genomic probes were prepared for all duplicons >15 kb in size, and hybridization results are summarized in Table 3 (for a representative set of experiments, see Fig. 2). In general, our FISH results indicate a reduction in copy number as probes are hybridized to orangutan and baboon. Interestingly, in several cases, no signals were observed among baboon or macaque. Although not all probes are single copy in orangutan, these results verify many of the putative duplicon ancestral positions as predicted by the origin of the expressed gene (see results for 4q24, Xq28, *IGSF3*, and *MLL3* in Table 3). Reciprocal experiments were conducted with baboon BACs representing each duplicon on baboon and human metaphase chromosomal spreads. Duplicons 11q, 12p, 4q24, *ALD* (from Xq28), and *IGSF3* (from 1p13) were verified to be ancestral loci based on the observation of a single signal in baboon (data not shown).

Since FISH experiments did not always yield a reliable signal in orangutan or baboon, we conducted genomic library hybridizations as a secondary means to refine the ancestral origin more precisely. A PCR probe (for location, see Fig. 1A; for sequence, see Supplemental Table 2) was designed within each duplicon and was used to screen large-insert genomic BAC libraries from orangutan (CHORI-253) and baboon (RPCI-41). Based on the genomic coverage and the number of positively hybridizing BACs, we estimated the copy number for each duplicon within each primate species (Table 4; Supplemental Methods). With the exception of the Igλ segment (which maps to a tandem gene cluster), the PIR4 segment (which was not identified in the baboon), and the *LSP1* duplicon (which apparently has undergone an independent duplication expansion), 11 out of the 14 duplicons mapped to a single locus in either orangutan or baboon (Table 4). Orangutan and baboon BACs corresponding to each single site were end-sequenced, and the sequences were aligned to the human genome reference sequence by using BLAST (build 34, NCBI, July 2003) (Supplemental Tables 3, 4). With the exception of orangutan *IGSF3* BACs, primate BAC end-sequences from each duplicon corresponded to human sequence located at the putative ancestral location.

To provide a more precise estimate of duplication timing, we performed a phylogenetic analysis based on primate comparative sequencing of each duplicon as described previously (Horvath et al. 2003). By utilizing PCR assays designed to noncoding 2p11 human reference sequence, orangutan and baboon BACs were PCR amplified, and the products were directly sequenced with multiple primer pairs within each duplicon. We constructed a neighbor-joining phylogenetic tree for 11 of the duplicons where complete sequence information could be obtained (Fig. 3). Genetic distances are indicated in Table 5 and were used to calculate the ancestral nucleotide substitution rate specifically for

Table 4. Summary of BAC hybridization results

		CHORI-253			RPCI-41		
Duplicon location	Duplicon name	Orangutan clone no.	Estimated copy no.	No. of Seq variants	Baboon clone no.	Estimated copy no.	No. of Seq variants
7p12	PIR4	25	3.9	4	0	0.0	0
11q14	11q	8	1.3	1+	6	1.2	1
12p11	12p	10	1.9	2	5	1.0	1+
22q11	Igλ	275	43.0	nd	230	44.2	nd
22q12	*CHK2*	6	0.9	1+	5	1.0	1
4q24	4q24	11	1.7	1+	7	1.3	1+
Xq28	*ALD*	5	0.8	1	8	0.8	1
22q11	*GGT1*	6	0.9	1	12	2.3	1
1p13	*IGSF3*	2	0.3	1+	3	0.6	1
10q26	10q26	9	1.4	1+	4	0.8	1
7q36	*MLL3*	5	0.8	2+	3	0.6	1+
4p16.1	4p16.1	6	0.9	1	7	1.3	1+
11p15	*LSP1*	19	1.6	3+	26	5.0	3
4p16.3	4p16.3	7	1.1	1+	7	1.3	1+

A "+" after sequence variant number denotes another copy exists with one to four differences out of 500bp (99.8%–99.2% identity). PCR-generated probes (Supplemental Table 2) corresponding to each duplication (except Igλ which was not done (nd) due to high copy number) were hybridized to orangutan (CHORI-253) and baboon (RPCI-41) BAC libraries. The number of positives obtained and the fold coverage of the library (based on the segments screened) were used to estimate the copy number of each duplication in each species (no. positives/coverage, expected copies). Only one segment of each library was screened with the exception of the 11p15 duplicon in orangutan and Xq28 and 7p12 in baboon, where both library segments were screened. All BAC positives were PCR amplified with the hybridization primer pair, and then PCR products were directly sequenced. The PCR product sequences were compared to determine the number of different sequence variants for each duplicon within each species (Seq variants).

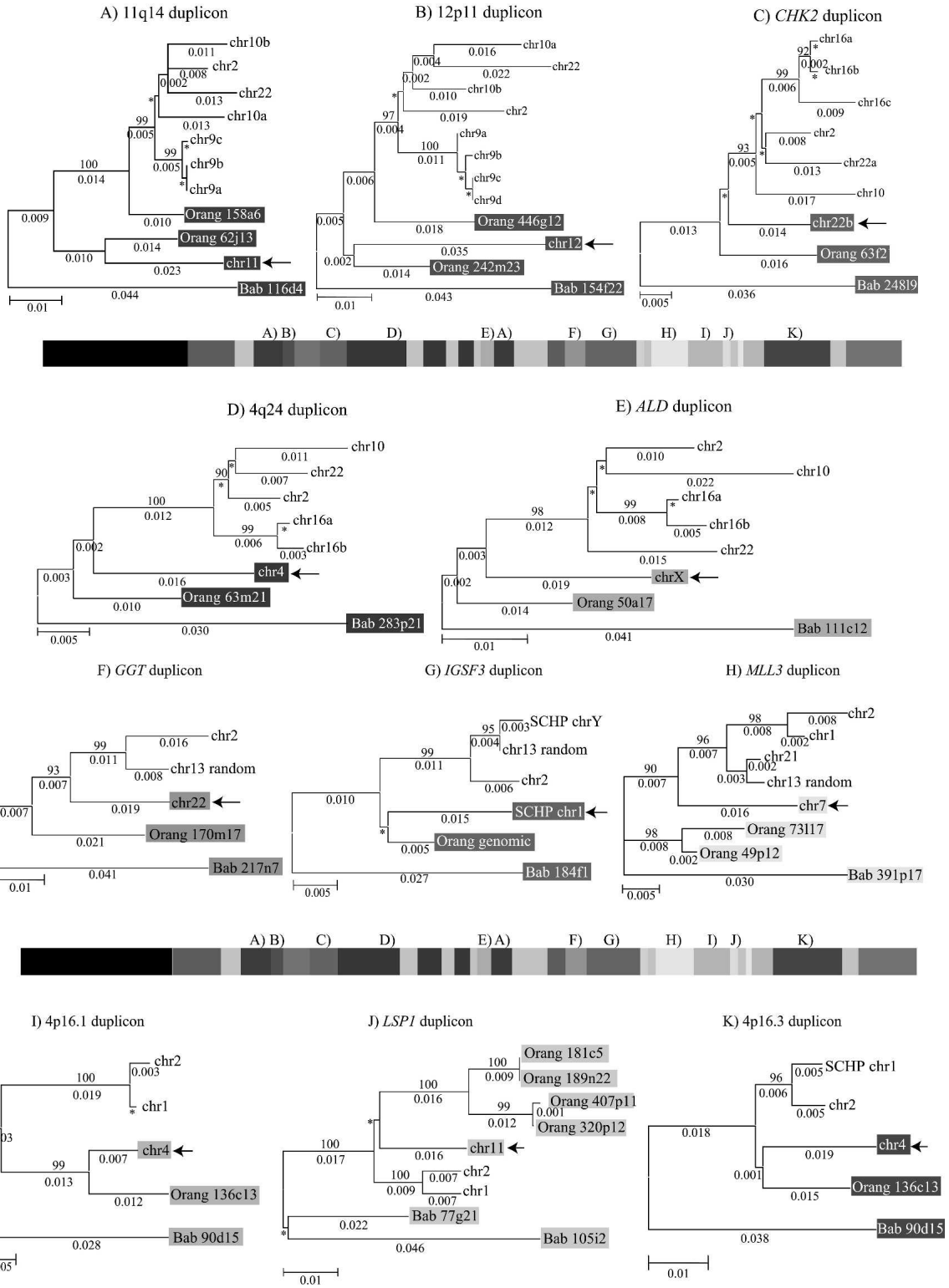


Figure 3. Phylogenetic trees for 2p11 duplicons. A neighbor-joining tree was constructed for each individual duplicon as shown *above* and *below* the gray schematic of the 2p11 duplicons (*A–K*). See Figure 1 for corresponding colored boxes. Gray boxes outline ancestral human, orangutan (Orang), and baboon (Bab) sequence taxa within the phylogenetic trees. Ancestral human sequences are also marked with an arrow. Branch lengths are proportional to the number of nucleotide changes between taxa and are indicated *below* each respective branch. An asterisk *next to* or *below* a branch length indicates a branch length of 0.001. Bootstrap values >90 from 1000 replicates are indicated *above* each corresponding branch. Sequence data from baboon and orangutan outgroups were obtained from large-insert BAC clones (CHORI-253 and RPCI-41) or total genomic DNA.

Table 5. Genetic distances summary

			Outgroup distances			Human interparalog distances					
Duplicon location	Duplicon name	Base pairs analyzed	Mean K Hum to Bab	Mean K Bab to Para	Mean K Bab to Anc	Mean K Ancest to Para	Mean K all Para	Bab mya	Bab rate (X10^{-9})	Baboon Seed (mya)	Baboon Swaps (mya)
11q14	11q	1074	0.081	0.082	0.078	0.063	0.016	23	1.70	17.9	4.5
12p11	12p	737	0.076	0.075	0.084	0.064	0.027	23	1.83	19.4	8.2
22q12	*CHK2*	804	0.067	0.068	0.064	0.031	0.022	23	1.39	10.6	7.6
4q24	4q24	806	0.055	0.055	0.053	0.036	0.015	23	1.15	15.1	6.3
Xq28	*ALD*	663	0.061	0.074	0.067	0.047	0.028	23	1.46	17.7	10.6
11q14	11q	1074	0.081	0.081	0.078	0.061	0.016	23	1.70	17.3	4.5
22q11	*GGT1*	770	0.079	0.079	0.074	0.041	0.025	23	1.61	11.9	7.3
1p13	*IGSF3*	758	0.053	0.053	0.054	0.031	0.008	23	1.17	13.5	3.5
7q34	*MLL3*	553	0.054	0.053	0.056	0.032	0.015	23	1.22	13.6	6.4
4p16.1	4p16.1	1057	0.052	0.052	0.052	0.041	0.007	23	1.13	18.1	3.1
11p15	*LSP1*	734	0.055	0.054	0.056	0.032	0.009	23	1.22	13.4	3.8
4p16.3	4p16.3	672	0.071	0.068	0.077	0.03	0.01	23	1.67	9.7	3.2

The average number of nucleotide substitutions per site (K, Kimura two-parameter model) and associated standard errors were calculated for three output comparisons: all human sequences to the baboon outgroup (Hum to Bab), all human nonancestral paralogs to the baboon outgroup (Bab to Para), and the ancestral human to the outgroup (Bab to Anc). These were compared with two human interparalog calculations: human ancestral to all human paralogs (Ancest to Para) and the average K of all human paralogs (all Para). Relative to the outgroup distance, these latter two estimates provide information for the initial duplication from the euchromatin to the pericentromeric region (pericentromeric seeding) and the secondary duplication events (pericentromeric swapping). Based on an estimated divergence time of 23 million years between the human and baboon lineages, we calculate the effective nucleotide substitution rates (r = K/2T) for each locus. Standard errors around K values were small (<20% of the K value) with few exceptions (paralog SE for IGSF3, MLL3, 4p16.1, LSP1, and 4p16.3; range between 27% and 40%) and are therefore not known.

each duplicon. These substitution rates range between 1.13 and 1.83×10^{-9} substitutions/site/yr and are generally consistent with estimates from other duplicated segments (Eichler et al. 1999; Liu et al. 2003). These values were then used to calculate seed and swap times corresponding to initial duplication of each donor segment and subsequent dispersal of these segments to other pericentromeric regions.

Ten of the 11 tree topologies are consistent with a major duplication seeding event occurring after the separation of Old World monkey and great-ape lineages (<23 million years ago [Mya]). All 10 phylogenies clearly distinguish two major events: an ancestral event (termed an ancestral duplicative transposition) followed by a series of secondary duplications (pericentromeric swapping) that group all human paralogs. Bootstrap support distinguishing these events ranges from 96–100 (Fig. 3). The *LSP1* duplicon is the only locus that is inconsistent with this model of evolution. In some cases, we observed similarities in the tree topology based on spatial proximity of the ancestral duplicons within 2p11. The first three duplications (for PIR4 tree, see Horvath et al. 2003) nearest the human centromere, for example, show evidence of duplication of the ancestral locus prior to the divergence of the humans and the great-apes from the Old World monkey lineage as evidenced by progenitor duplicates in the orangutan lineage. In general, evolutionary genetic distance estimates between human ancestral and paralogous loci (0.03–0.06) are significantly less than the genetic distance between the human ancestral locus and the corresponding baboon locus (0.05–0.08) (Fig. 4; Table 5). By using locus-specific substitution rates, we calculated that the initial duplication of the ancestral locus occurred between 9 and 19 Mya. Although secondary dispersal events occurred ~3–11 Mya (Fig. 4; Table 5), there is no evidence of a novel ancestral duplicative transposition event having occurred over the past 9 Myr within this region of 2p11.

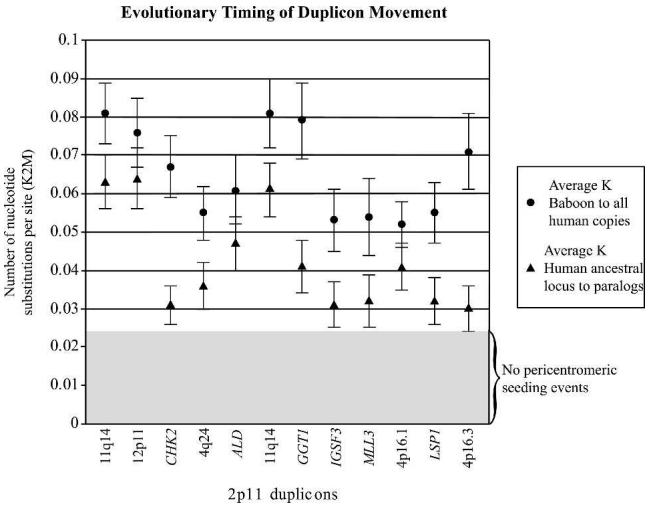


Figure 4. Sequence divergence of 2p11 duplicons. The graph compares the average divergence (substitutions per site, Kimura two-parameter model with standard error measurements) for baboon and all human duplicate copies (circles) to the average divergence for the human ancestral locus to all human pericentromeric copies (triangles). The former provides a locus-specific estimate of the effective number of substitutions since the divergence of Old World monkeys and human lineages (~23 Mya), while the latter provides an estimate of the timing of the initial duplication event. With the exception of *LSP1*, the baboon copy corresponds to a single (nonduplicated) locus. The data are consistent with an initial duplicative transposition of the ancestral locus for all loci after separation of the Old World and human lineages. No duplications from an ancestral locus are observed within this 700-kb region which show <0.03 substitutions/per site. This suggests a cessation of euchromatic colonization of this region ~10 Mya.

Discussion

We present one of the most comprehensive evolutionary analyses, to date, of a human centromeric transition region. We have extended the model of pericentromeric duplication by systematically tracking the origin and timing of a series of duplicons located within a 700-kb pericentromeric region of 2p11 (She et al. 2004a). Our goal was to reconstruct the evolutionary history of this region by using a combination of phylogenetic, genomic, and comparative FISH approaches. Our study provides compelling evidence for an evolutionarily punctuated movement of duplicated material 10–20 Mya for the majority of the 2p11 pericentromeric region. Although we can not preclude the existence of more ancient duplications of euchromatin that have been deleted/diverged before this time period, the identification of more recent ancestral duplicative transpositions should have been trivial to detect. None, however, were identified within this portion of 2p11.

Previous analyses have suggested that pericentromeric regions have been formed via the duplication of euchromatic segments that have colonized pericentromeric DNA over the past 30 Myr of evolution (Eichler et al. 1996; Jackson et al. 1999; Horvath et al. 2000b, 2003; Bailey et al. 2002; Crosier et al. 2002; She et al. 2004a; Locke et al. 2005). This duplicative transposition of euchromatic segments into pericentromeric regions (which we have termed "pericentromeric seeding") has led to the formation of complex mosaics of segmental duplications consisting of juxtaposed duplicons from diverse euchromatic positions. Secondary duplications of larger mosaic blocks (termed "pericentromeric swapping" events) occurred subsequently, leading to differential distribution of these blocks among the great-ape and human pericentromeric regions. Detailed analyses of pericentromeric regions (10p11, 10q11, 15q11, 2p11, and 16p11), as well as more global computational analysis, suggest that this is a general principle of human genome evolution (Jackson et al. 1999; Guy et al. 2000, 2003; Horvath et al. 2000b; Locke et al. 2005). Our extended analysis of 2p11 confirms this two-step model (Fig. 5) but also indicates that most euchromatic seeding events occurred over a more narrow window of evolutionary time than previously appreciated (Guy et al. 2000, 2003; Bailey et al. 2001, 2002).

Results from comparative FISH of 2p11 duplicons indicate that many segments were originally duplicated after the divergence of the human and baboon lineages (~23 Mya), but before the divergence of human and the African great-apes (~8 Mya) (Fig. 2; Table 3). The phylogenetic data agree closely with the comparative FISH data. The genetic distance, for example, between human and baboon sequence ranges from 0.052–0.081, while the evolutionary distance between the human euchromatic ancestral locus and pericentromeric paralogs ranges from 0.03–0.064 (Fig. 4; Table 5). Based on relative rate tests and individual calibration for the substitution rate of each locus, these distances translate into pericentromeric seeding events that occurred 10–20 Mya. As expected, our genomic studies occasionally identified duplicated sequence among the orangutan great-apes (thought to have diverged 12–14 Mya) (Fig. 3). No additional evidence of euchromatic to pericentromeric seeding events could be identified within human 2p11 after the separation of humans

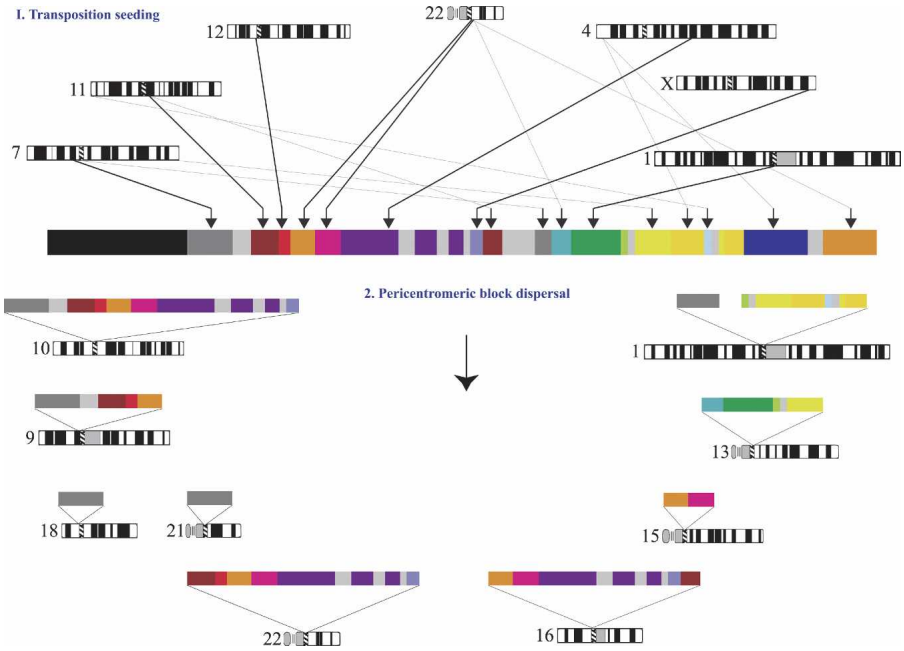


Figure 5. A model for the acquisition and dispersal of 2p11 duplicons. An expanded two-step model is shown to explain the current organization of 2p11. First, a burst of DNA duplicative transposition events occurs in the common ancestor of humans and apes (10–20 Myr), creating a large mosaic region consisting of at least 14 duplicons. During the radiation of humans and African great-apes (4–8 Mya), a series of secondary duplications disperse larger cassettes to other pericentromeric regions, leading to quantitative and qualitative differences of each larger block within different lineages. More recent transposition events suddenly cease or are no longer fixed during this second phase.

from chimps and gorillas, although secondary duplication events (pericentromeric swapping) are readily observed.

It is unclear why pericentromeric seeding events occurred so frequently during this period of human/great-ape evolutionary history. It is also unclear why they suddenly cease, at least in the case of 2p11. One possible scenario may be that certain regions of the genome are permissive to segmental duplication events only at specific periods of time. The permissive nature may relate to evolutionary changes in transcriptional activity or the chromatin configuration of these regions. In such a scenario, one might expect to find pericentromeric regions with younger or older duplicons depending on differences in the chromatin context in which they emerged. A global analysis of several pericentromeric regions confirms that, in general, younger (<8 Mya) pericentromeric seeding events are a relatively rare occurrence in the human genome (Bailey et al. 2002; She et al. 2004a; Locke et al. 2005). This is not to say that pericentromeric-to-pericentromeric duplications have *not* continued to occur more recently. Indeed, there are numerous examples of such pericentromeric swapping events that have emerged since the great-ape/human separation, and a few have been unambiguously shown to be lineage-specific events (Bailey et al. 2002). In addition, other nonpericentromeric regions of the human genome show ample evidence of more recent (<8 Mya) duplicative transposition events into acceptor regions (Johnson et al. 2001; Stankiewicz et al. 2004).

There are several other possible scenarios that may be put forward to explain this punctuated genome restructuring process. For example, it is interesting to note that the "shift" from pericentromeric seeding to pericentromeric swapping coincides with the emergence of higher-order α-satellite DNA (8 Mya) (Haaf and Willard 1998). This change in centromeric higher-order structure may have influenced ectopic recombination events among nonhomologous chromosomes, providing a mechanism for these secondary duplication events.

We cannot rule out the possibility that our view of the duplication process as "punctuated" is obscured by having an incomplete genome. If new seeding events are primarily restricted to the unsequenced p arms of acrocentric chromosomes, we may miss them entirely. There is a small amount of evidence that acrocentric p arms do harbor duplicons (Wohr et al. 1996; Eisenbarth et al. 1999; Hattori et al. 2000; Cserpan et al. 2002); however, their sequence identity attributes do not appear to differ significantly from what has been observed for other pericentromeric regions.

High-quality BAC-based sequence within pericentromeric regions has revealed a remarkable level of evolutionary dynamism. Comparative studies such as these provide valuable information into the evolutionary forces that have reshaped our genomes—forces that likely contribute to contemporary variation and disease. Detailed comparative sequencing of these regions, however, is required to address several of the hypotheses and models that we have put forward. While correct assembly of these regions is often a daunting task, we have demonstrated that such regions can be assembled and sequenced with available genomic resources (Horvath et al. 2000a). Unfortunately, the quicker method of sequence assembly, whole genome shotgun assembly, may preclude such rich evolutionary analyses as complex and duplicated regions will be incorrectly assembled or simply not represented (She et al. 2004b). Targeted comparative studies with large-insert clones from these regions promise to provide valuable insight into the evolution of our species and genome.

Methods

Computational analyses

Duplicon identification was conducted for each individual accession by using RepeatMasker (RepeatMasker version 07/13/2002; A. Smit and P. Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html) sequence as query against the EST division of GenBank. All ESTs showing exon/intron structure to the query accession were used to identify UniGene clusters when available (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene). A representative EST from each UniGene cluster was used as query against nr (nonredundant) and htgs (high throughput genome sequence). All ESTs not belonging to a UniGene cluster were used as query individually. An accession with an identical match to the representative EST was considered the ancestral locus and was used to identify the chromosomal region in build34 for further comparisons of duplicon size and identity (Table 1). Optimal global alignments of BAC overlaps and ancestral loci to each 2p11 paralogous segment were generated by

using the program ALIGN (Myers and Miller 1988). NT_034508 was used (in reverse orientation) for database searches of all paralogous and ancestral loci in build34. These hits were displayed by using PARASIGHT (http://humanparalogy.gs.washington.edu/parasight) (Fig. 1A).

PCR and sequencing

The BAC and cosmid clones used for PCR analysis were grown from single colony isolates in 5 mL overnight cultures. The DNA was isolated by using the Millipore (Millipore) or Perfectprep BAC 96 kit (Eppendorf) and resuspended in water. Approximately 15 ng BAC DNA (1/25 the total volume) and 15 ng of cosmid DNA (1/50 the total volume) were used in subsequent PCR assays. All PCR and sequencing conditions were previously described elsewhere (Horvath et al. 2003). BAC end sequencing reactions were conducted as previously described (She et al. 2004a). Cosmid end sequencing reactions were identical to BAC end reactions except that only 1/12 the total volume of cosmid DNA was used, and only 70 cycles of sequencing were conducted. We assessed the quality of all sequence data using PHRED/PHRAP/CONSED software (http://genome.wustl.edu).

Phylogenetic analysis

FASTA formatted sequences were obtained after comparison of both forward and reverse sequences from each PCR product using CONSED. All primate BAC sequences were searched against build34 to obtain all fully sequenced human copies. Sequence alignments were built by using CLUSTALW (version 1.82) (Higgins et al. 1996), and maximum parsimony, minimum evolution, and neighbor-joining methods were all used to construct phylogenetic trees by using MEGA (Molecular Evolutionary Genetic Analysis) v2.1 (http://www.megasoftware.net/) (Kumar et al. 2001). Although all three methods yielded trees with identical topology, neighbor-joining phylograms are shown because they allow for distance estimates between taxa. Neighbor-joining analysis was used with complete deletion parameters for all duplicon trees (Fig. 3) and pairwise deletion parameters for the α-satellite trees (Supplemental Fig. 1B) with 1000 bootstrap iterations. Tajima's relative rate tests (Tajima 1993) were used in MEGA (Kumar et al. 2001) to determine if rates of nucleotide substitution were constant between the three species (human, orangutan, and baboon). We estimated the number of substitutions/site/year (substitution rate) by correcting the divergence times for multiple substitutions using Kimura's two-parameter model (Kimura 1980). Divergence times of 23 Myr between the human and baboon lineages and 13 Myr between human and orangutan lineages were used. Duplication timing events were calculated by using the equation T=K/2r (Li 1997). The approximate seed time (in millions of years) was determined by multiplying the ancestral to paralog K value by 23 Myr (human to baboon divergence estimate) and dividing by the baboon to paralog K value. Swap times were calculated using the average K of all human paralogs in place of the ancestral to paralog K value.

Acknowledgments

We thank Lawrence Livermore National Labs and the UK HGMP Resource Centre for providing the cosmid library filters and clones. We thank Sean McGrath, Mandeep Sekhon, Andrew Grow, Jason Carter, and Laurie Christ for technical assistance and Dr. Norman Doggett for kindly providing access to chromosome 16 cosmid filters and clones. We thank Huntington F. Willard, Carol Stepien, Stuart Schwartz, Mitch Drumm, and Joe Nadeau for insightful discussions regarding all aspects of this work. We also thank Mary Schueler and Katie Rudd for helpful discussions regarding α-satellite DNA, and Lisa Chadwick for helpful suggestions with this manuscript. Chromosome ideograms for Figure 5 were obtained from the University of Washington Department of Pathology Web site: (http://www.pathology.washington.edu/research/cytopages/idiograms/human/). This work was supported, in part, by NIH grants HG002385 and GM58815 to E.E.E. In addition, we 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. J.E.H. was supported in part by NIH GM08613, Genetics Training grant.

References

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

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

Bailey, J.A., Liu, G., and Eichler, E.E. 2003. An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73:** 823–834.

Chen, F.C. and Li, W.H. 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.

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

Cserpan, I., Katona, R., Praznovszky, T., Novak, E., Rozsavolgyi, M., Csonka, E., Morocz, M., Fodor, K., and Hadlaczky, G. 2002. The chAB4 and NF1-related long-range multisequence DNA families are contiguous in the centromeric heterochromatin of several human chromosomes. *Nucleic Acids Res.* **30:** 2899–2905.

Donze, D. and Kamakaka, R.T. 2002. Braking the silence: How heterochromatic gene repression is stopped in its tracks. *Bioessays* **24:** 344–349.

Eichler, E.E., Lu, F., Shen, Y., Antonacci, R., Jurecic, V., Doggett, N.A., Moyzis, R.K., Baldini, A., Gibbs, R.A., and Nelson, D.L. 1996. 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.

Eichler, E.E., Budarf, M.L., Rocchi, M., Deaven, L.L., Doggett, N.A., Baldini, A., Nelson, D.L., and Mohrenweiser, H.W. 1997. Interchromosomal duplications of the adrenoleukodystrophy locus: A phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.* **6:** 991–1002.

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

Eichler, E.E., Clark, R.A., and She, X. 2004. An assessment of the sequence gaps: Unfinished business in a finished human genome. *Nat. Rev. Genet.* **5:** 345–354.

Eisenbarth, I., Konig-Greger, D., Wohr, G., Kehrer-Sawatzki, H., and Assum, G. 1999. Characterization of an alphoid subfamily located near p-arm sequences on human chromosome 22. *Chromosome Res.* **7:** 65–69.

Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., and Trask, B.J. 2002. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14.1 and paralogous regions on other human chromosomes. *Genome Res.* **12:** 1651–1662.

Guy, J., Spalluto, C., McMurray, A., Hearn, T., Crosier, M., Viggiano, L., Miolla, V., Archidiacono, N., Rocchi, M., Scott, C., et al. 2000. Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Hum. Mol. Genet.* **9:** 2029–2042.

Guy, J., Hearn, T., Crosier, M., Mudge, J., Viggiano, L., Koczan, D., Thiesen, H.J., Bailey, J.A., Horvath, J.E., Eichler, E.E., et al. 2003.

Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. *Genome Res.* **13:** 159–172.
Haaf, T. and Willard, H.F. 1998. Orangutan α-satellite monomers are closely related to the human consensus sequence. *Mamm. Genome* **9:** 440–447.
Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., et al. 2000. The DNA sequence of human chromosome 21: The chromosome 21 mapping and sequencing consortium. *Nature* **405:** 311–319.
Higgins, D.G., Thompson, J.D., and Gibson, T.J. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266:** 383–402.
Horvath, J., Schwartz, S., and Eichler, E. 2000a. 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. 2000b. Molecular structure and evolution of an α/non-α satellite junction at 16p11. *Hum. Mol. Genet.* **9:** 113–123.
Horvath, J.E., Bailey, J.A., Locke, D.P., and Eichler, E.E. 2001. Lessons from the human genome: Transitions between euchromatin and heterochromatin. *Hum. Mol. Genet.* **10:** 2215–2223.
Horvath, J.E., Gulden, C.L., Bailey, J.A., Yohn, C., McPherson, J.D., Prescott, A., Roe, B.A., De Jong, P.J., Ventura, M., Misceo, D., et al. 2003. Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications. *Mol. Biol. Evol.* **20:** 1463–1479.
International Human Genome Sequencing Consortium (IHGSC). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
———. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.
Ijdo, J., Baldini, A., Ward, D.C., Reeders, S.T., and Wells, R.A. 1991. Origin of human chromosome 2: An ancestral telomere–telomere fusion. *Proc. Natl. Acad. Sci.* **88:** 9051–9055.
Jackson, M.S., Rocchi, M., Thompson, G., Hearn, T., Crosier, M., Guy, J., Kirk, D., Mulligan, L., Ricco, A., Piccininni, S., et al. 1999. Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications and unstable sequences with homologies to telomeric and other centromeric locations. *Hum. Mol. Genet.* **8:** 205–215.
Ji, Y., Eichler, E.E., Schwartz, S., and Nicholls, R.D. 2000. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res.* **10:** 597–610.
Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E.E. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413:** 514–519.
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.
Kumar, S., Tamura, K., Jakobsen, I.B., and Nei, M. 2001. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244–1245.
Li, W. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, MA.
Liu, G., Zhao, S., Bailey, J.A., Sahinalp, S.C., Alkan, C., Tuzun, E., Green, E.D., and Eichler, E.E. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13:** 358–368.
Locke, D.P., Jiang, Z., Pertz, L.M., Misceo, D., Archidiacono, N., and Eichler, E.E. 2005. Molecular evolution of the human chromosome 15 pericentromeric region. *Cytogenet. Genome Res.* **108:** 73–82.
Luijten, M., Wang, Y., Smith, B.T., Westerveld, A., Smink, L.J., Dunham, I., Roe, B.A., and Hulsebos, T.J. 2000. Mechanism of spreading of the highly related neurofibromatosis type 1 (NF1) pseudogenes on chromosomes 2, 14 and 22. *Eur. J. Hum. Genet.* **8:** 209–214.
Manuelidis, L. 1978. Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma* **66:** 23–32.
May, W., Korenberg, J.R., Chen, X.N., Lunsford, L., Wood, W.J., Thompson, A., Wall, R., and Denny, C.T. 1993. Human lymphocyte-specific pp52 gene is a member of a highly conserved dispersed family. *Genomics* **15:** 515–520.
McConkey, E.H. 2004. Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. *Cytogenet. Genome Res.* **105:** 157–158.
Myers, E.W. and Miller, W. 1988. Optimal alignments in linear space. *Comput. Appl. Biosci.* **4:** 11–17.
Orti, R., Potier, M.C., Maunoury, C., Prieur, M., Creau, N., and Delabar, J.M. 1998. Conservation of pericentromeric duplications of a 200-kb part of the human 21q22.1 region in primates. *Cytogenet. Cell Genet.* **83:** 262–265.
Regnier, V., Meddeb, M., Lecointre, G., Richard, F., Duverger, A., Nguyen, V.C., Dutrillaux, B., Bernheim, A., and Danglot, G. 1997. 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.
Ruault, M., Trichet, V., Gimenez, S., Boyle, S., Gardiner, K., Rolland, M., Roizes, G., and De Sario, A. 1999. Juxta-centromeric region of human chromosome 21 is enriched for pseudogenes and gene fragments. *Gene* **239:** 55–64.
Rudd, M.K. and Willard, H.F. 2004. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* **20:** 529–533.
Samonte, R.V. and Eichler, E.E. 2002. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* **3:** 65–72.
Saupe, S., Roizes, G., Peter, M., Boyle, S., Gardiner, K., and De Sario, A. 1998. Molecular cloning of a human cDNA IGSF3 encoding an immunoglobulin-like membrane protein: Expression and mapping to chromosome band 1p13. *Genomics* **52:** 305–311.
Schueler, M.G., Higgins, A.W., Rudd, M.K., Gustashaw, K., and Willard, H.F. 2001. Genomic and genetic definition of a functional human centromere. *Science* **294:** 109–115.
She, X., Horvath, J.E., Jiang, Z., Liu, G., Furey, T.S., Christ, L., Clark, R., Graves, T., Gulden, C.L., Alkan, C., et al. 2004a. The structure and evolution of centromeric transition regions within the human genome. *Nature* **430:** 857–864.
She, X., Jiang, Z., Clark, R.A., Liu, G., Cheng, Z., Tuzun, E., Church, D.M., Sutton, G., Halpern, A.L., and Eichler, E.E. 2004b. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431:** 927–930.
Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825–837.
Smit, A.F. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743–748.
Stankiewicz, P., Shaw, C.J., Withers, M., Inoue, K., and Lupski, J.R. 2004. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res.* **14:** 2209–2220.
Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135:** 599–607.
Tassone, F., Xu, H., Burkin, H., Weissman, S., and Gardiner, K. 1995. cDNA selection from 10 Mb of chromosome 21 DNA: Efficiency in transcriptional mapping and reflections of genome organization. *Hum. Mol. Genet.* **4:** 1509–1518.
Willard, H.F. 1991. Evolution of α satellite. *Curr. Opin. Genet. Dev.* **1:** 509–514.
Willard, H.F. and Waye, J.S. 1987. Chromosome-specific subsets of human α satellite DNA: Analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.* **25:** 207–214.
Wohr, G., Fink, T., and Assum, G. 1996. A palindromic structure in the pericentromeric region of various human chromosomes. *Genome Res.* **6:** 267–279.
Zimonjic, D., Kelley, M., Rubin, J., Aaronson, S., and Popescu, N. 1997. Fluorescence in situ hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. *Proc. Natl. Acad. Sci.* **94:** 11461–11465.

Web site references

http://humanparalogy.gs.washington.edu/parasight; PARASIGHT.
http://ftp.genome.washington.edu/RM/RepeatMasker.html; RepeatMasker.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene; UniGene clusters.
http://www.megasoftware.net/; MEGA.
http://genome.wustl.edu; PHRED/PHRAP/CONSED software.
http://www.pathology.washington.edu/research/cytopages/idiograms/human/; Idiogram album from the University of Washington, Department of Pathology.

Received March 14, 2005; accepted in revised form May 3, 2005.