MBE Advance Access published May 30, 2003

Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications.

Horvath, J.E. [1], Gulden, C.L. [1], Bailey, J. A. [1], Yohn, C. [1], Mcpherson, J.D. [2], Prescott, A. [3], Roe, B. A. [3], de Jong, P. J[4], Ventura, M. [5], Misceo, D. [5], Archidiacono, N[5]., Zhao, S. [6], Schwartz, S. [1], Rocchi, M. [5], and Eichler, E. E[1]†.

[1]Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, OH 44106. [2]Washington University School of Medicine Genome Sequencing Center, St. Louis, Missouri 63108, USA. [3]Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, P.O. Box 26901, Oklahoma City, OK 73190, USA. [4]Children's Hospital Oakland Research Institute, BACPAC Resources, 747 52nd Street, Oakland, CA 94609-1809, USA. [5]Sezione di Genetica, DAPEG, University of Bari, Via Amendola 165/A 70126 Bari, Italy and [6]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

†Corresponding author:

Evan Eichler, Ph.D.
Department of Genetics
Case Western Reserve University
BRB720, 10900 Euclid Ave.
Cleveland, OH 44106
Phone: (216) 368-4883
Fax: (216) 368-3432
e-mail: eee@po.cwru.edu

Keywords: Pericentromeric DNA, segmental duplications, genome architecture, non-human primates, genome evolution.

Running head: PIR4 expansion in primate genomes.

Copyright (c) 2003 Society for Molecular Biology and Evolution

ABSTRACT

Despite considerable advances in sequencing of the human genome over the last few years, the organization and evolution of human pericentromeric regions have been difficult to resolve. This is due, in part, to the presence of large, complex blocks of duplicated genomic sequence at the boundary between centromeric satellite and unique euchromatic DNA. Here, we report the identification and characterization of a ~49 KB repeat sequence that exists in more than 40 copies within the human genome. This repeat is specific to highly duplicated pericentromeric regions with multiple copies distributed in an interspersed fashion among a subset of human chromosomes. Using this interspersed repeat (termed PIR4) as a marker of pericentromeric DNA, we recovered and sequence-tagged 3 Mb of pericentromeric DNA from a variety of human chromosomes as well as non-human primate genomes. A global evolutionary reconstruction of the dispersal of PIR4 sequence and analysis of flanking sequence supports a model in which pericentromeric duplications initiated before the separation of the great ape species (>12 mya). Further, analyses of this duplication and associated flanking duplications narrow the major burst of pericentromeric duplication activity to a time just before the divergence of the African great ape and human species (5-7 mya). These recent duplication exchange events substantially restructured the pericentromeric regions of hominoid chromosomes and created an architecture where large blocks of sequence are shared among non-homologous chromosomes. This report provides the first global view of the series of historical events that have reshaped human pericentromeric regions over recent evolutionary time.

INTRODUCTION

With the sequence of many organisms complete or nearly so, comparative work between species promises to expand our knowledge of genome organization and evolution. Pericentromeric regions are particularly interesting because these regions demarcate the transition between the heterochromatic alpha satellite DNA at the centromere and the euchromatic gene-containing chromosome arm sequences. Further, such regions are sites of rapid evolutionary turnover, reduced gene expression and suppressed genetic recombination (Eichler 2001a; Mahtani and Willard 1998; Yan et al. 2002). An understanding of the genetic and functional properties requires a detailed understanding of the sequence structure.

Pericentromeric organization

In general, resolution of the organization and evolution of these regions has been hampered by unusual constellations of repetitive sequences when compared to other regions of the genome. Sequence analysis of *Drosophila melanogaster* pericentromeric regions indicated that they are mainly composed of simple satellite sequences, transposons, retroposons, and rRNA genes (Adams et al. 2000) (Sun, Wahlstrom and Karpen 1997). Similarly, *Arabidopsis thaliana* pericentromeric regions are largely composed of retroelements, transposons, microsatellites and various classes of middle repetitive DNA (Copenhaver et al. 1999; The Arabidopsis genome initiative 2000). In addition to simple satellites and retroposons, directed analyses of human pericentromeric regions on chromosomes 2, 10 and 16 reveal a preponderance of partial gene duplications

(Horvath et al. 2000) (Horvath, Schwartz and Eichler 2000; Jackson et al. 1999) (Guy et al. 2000; Loftus et al. 1999). Although the occurrence of mobile genetic elements and duplicated sequence within pericentromeric regions is a common property shared by these distant species, the structure of the human genome appears to be unique in the proportion and extent of these blocks of duplications which may be as large as a few Mb in size (Hattori et al. 2000) (Dunham et al. 1999; Jackson et al. 1999) (Bailey et al. 2002a; Bailey et al. 2001; Bailey et al. 2002b; Crosier et al. 2002; IHGSC 2001).

Human pericentromeric duplications

The structure of these large mosaic blocks of duplication is complex. For nearly half of human chromosomes, an estimated zone of duplication extends from the satellite-repeat sequence to the unique euchromatic region (Bailey et al. 2001; IHGSC 2001). These regions are composed of a mosaic of duplicated genomic segments that originate from diverse areas of the genome. A large number of partial and whole gene duplications have been recently characterized in detail (Horvath et al. 2001; Jackson et al. 1999). These segmental duplications share conserved exon-intron structure and have been termed duplicons (Eichler 2001b). In most cases, the duplicons originate from an ancestral expressed locus, range in copy number from 2-15, and show an interchromosomal distribution restricted largely to pericentromeric regions. Comparative analyses of a few regions indicate that these transposed duplicated segments are found only in humans and closely related non-human primates (Horvath et al. 2000) (Eichler et al. 1997; Eichler et al. 1996; Regnier et al. 1997) (Orti et al. 1998) (Zimonjic et al. 1997) (Arnold et al. 1995). With the exception of these few anecdotal studies focused on individual

duplicated segments, a global synopsis of this property of genome evolution and chromosome structure has been lacking. The molecular basis for the duplicative transposition bias toward pericentromeric regions is unknown.

Pericentromeric specific repeat sequences

In addition to duplicated gene segments, a variety of primate-specific degenerate repeat sequences have been identified between the duplicons (Horvath, Schwartz and Eichler 2000) (Eichler, Archidiacono and Rocchi 1999) (Eichler et al. 1997; Guy et al. 2000). The fact that they demarcate the transition between unrelated pericentromeric genic duplication events and that, in at least one case, they existed prior to the evolutionary transfer of the duplicated segments which has been taken as circumstantial evidence that these repeats may play a role in the duplication process (Eichler, Archidiacono and Rocchi 1999). Unlike the genic duplications described above, these pericentromeric interspersed repeat sequences (PIRs) do not exhibit obvious exon/intron structure. They, therefore, do not appear to be derived from ancestral gene sequences that have been transposed from non-pericentromeric regions of the genome. Several types of pericentromeric repeat sequences have been described including CAGGG, GGGCAAAAGCCG and chAB4 repeats (Eichler et al. 1997; Eichler et al. 1996) (Eichler, Archidiacono and Rocchi 1999; Horvath, Schwartz and Eichler 2000) (Assum et al. 1991; Wohr, Fink and Assum 1996). Unlike satellite sequences, these sequences are not composed of repetitive tandem arrays. In some cases, the underlying sequence structure of the interspersed repeats is reminiscent of degenerate subtelomeric repeat tracts (Flint et al. 1997) (Riethman et al. 2001). Indeed, telomeric associated-repeats

have occasionally been reported in close proximity to these sequence elements (Eichler, Archidiacono and Rocchi 1999). In addition, the pericentromeric interspersed repeats often exist at multiple locations within the same chromosome, separated by tens to hundreds of kb of intervening duplicated sequence.

Here we characterize a novel pericentromeric interspersed repeat, termed PIR 4, that is specific to the genomes of humans and apes. This element represents one of the most abundant recent segmental duplications within the human genome. Among humans this repeat occurs on more than half of all chromosomes; it is found in association with other segmental duplications; and it is restricted almost exclusively to pericentromeric regions. The purpose of this study was to take advantage of the multichromosomal and pericentromeric distribution of this interspersed repeat, using it as a marker to 1) recover additional sequence from these intractable regions of the genome 2) map existing sequences generated as part of the HGP that were ambiguously placed and 3) reconstruct the series of evolutionary events that occurred in the distribution of this repeat among primate chromosomes. Our analysis provides a global snapshot of the dynamic evolutionary history of these regions and the series of non-homologous sequence exchanges that created the architecture of contemporary human chromosomes.

MATERIALS AND METHODS

Computational Analyses

To identify all sequenced copies of PIR4 in the genome, BLASTN sequence similarity searches were initially performed against both nr (non-redundant) and htgs (high

throughput genomic sequence) divisions of GenBank using masked (RepeatMasker version 07/13/2002, A. F. A. Smit and P. Green, (http://ftp.genome.washington.edu/RM/RepeatMasker.html) using representative PIR4 sequence from AC002038.1 (coordinates 140, 121-161,973). All sequenced accessions were then searched against each other to identify the longest copy that was AC073318 (positions 71,401-120,576). Repeatmasked sequence from AC073318 was used as query against both nr and htgs divisions of GenBank and identified 170 finished and working draft GenBank accessions containing at least 1 kb and >90% sequence identity to the query sequence.

Of the 170 accessions containing PIR4 there were only 37 that were distinct finished copies and could be used for further analyses. These 37 GenBank accessions were analyzed using our previously described algorithm (Bailey et al. 2001) which is designed to capture large genomic alignments despite the presence of retroposon-induced large insertions and deletions. Here, the PIR4 reference (AC073318) was compared to each of the 37 finished clones after the high copy repeats identified by RepeatMasker (version 07/13/2002) were spliced out and then a pairwise comparison using gap BLAST was generated (Altschul et al. 1990). For each alignment, repeats were subsequently reinserted, the end-points were heuristically trimmed, and optimal global alignments were generated using the program ALIGN (Myers and Miller 1988). Based on these alignments, we extracted a PIR4 sequence for each sequence accession based on the orientation and extent of the putative ancestral locus, AC073318, where overlapping sequences compared to the reference sequence were removed in favor of the higher

scoring alignment within the clone's sequence. These extracted segments served as the basis for constructing an optimal global alignment for all PIR4 pairs of sequence. We limited our analysis to alignments ≥10 kb (a total of 25 GenBank accessions). We estimated the number of substitutions/site/year (substitution rate) by correcting the divergence for multiple substitutions using Kimura's two-parameter model (Kimura 1980).

To study the characteristics of other duplicons flanking the PIR4 sequences, we performed a second all-by-all BLASTN comparison of the 25 accessions that included the entire GenBank accession. We defined flanking alignments as alignments within 7 kb (the full length of L1 element insertion) of PIR4 sequence. Alignment statistics were only calculated for the non-PIR4 alignment portions. For each clone comparison, we selected the largest global alignment (minus PIR4) that was ≥10kbp. To compare the divergence of PIR4 to the largest flanking alignment we calculated the difference (K_{PIR4} - $K_{flanking}$). In this case, a positive (K_{PIR4} - $K_{flanking}$) value indicates that PIR4 is more divergent while a negative (K_{PIR4} - $K_{flanking}$) value reflects a more divergent flanking sequence. A value at or near zero indicates that both PIR4 and flanking duplicons were equally divergent and likely duplicated at or near the same timepoint in evolution.

Identification of duplicons within accessions (Figure 1 a, b, c) was performed using a Repeat Masked accession against the EST division of GenBank. All ESTs exhibiting exon/intron structure to the given accession were searched against the Unigene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene). A representative EST (for

each cluster with more than one EST hit to the given accession) and all ESTs without Unigene hits were subsequently queried against the nr and htgs divisions of GenBank. The accession with a 100% match to the query EST was considered the ancestral locus of the duplication which then was used in comparisons to the original PIR4 accession to determine the extent of overlap and percent identity of this segment as shown in Fig. 1.

Hybridization

The RPCI-11 human BAC library (segments 1, 2, 4, 5), RPCI-43 chimpanzee BAC library, RPCI-41 baboon BAC library (segments 1 and 2), CHORI-253 orangutan (segment 1) and cosmid libraries LLNL-01AH, LLNL-02AE, LLNL-07Y, LLNL-09P, LA13NC01, LA14NC01, LA15NC01, LA16NC01, LLNL-18AD, LLNL-21Q and LLNL-22N (Table 1) were hybridized with a PCR-generated probe using forward primer 32 and reverse primer 49 (see Table 2) amplified from 2p11 BAC DNA (AC002038). Known PUC false positives were removed from all BAC positive lists before PCR analysis. Since no false positive lists exist for the cosmid libraries all positives were used in PCR assays and only those amplifying with 32N49 were used in further analyses. The RPCI-41 (segment 2) baboon BAC library was also hybridized using a long-range PCR-generated probe using forward primer (62) and reverse primer (64) (see Table 2). This hybridization yielded no positives. Hybridization probes were purified from pooled PCR product using Qiagen's QIAquick® PCR purification kit (250) according to the manufacturer's recommendations. Twenty-five to fifty nanograms of purified product was random-hexamer labeled with [α-^{32}P] dCTP using Amersham's Megaprime kit

according to the manufacturer's recommendations. All membranes were blocked with 1mg sonicated salmon sperm DNA (Stratagene, La Jolla, CA). High-density arrayed BAC and cosmid membranes were hybridized at 65ºC for at least 16 hours in 25mL hybridization solution (0.25M $NaPO_4$, 0.25M NaCl, 5% SDS, 10% PEG, 1mM EDTA) heated to 65ºC. Membranes were washed three times (for 30 minutes each) at 65ºC in wash solution (0.05M $NaPO_4$, 0.5% SDS, 1mM EDTA) at room temperature for the first wash and heated to 65ºC for the second and third washes. Genomic southern blots were performed using 5µg PstI digested genomic DNA from two chimpanzees, two bonobos, one baboon, one orangutan, two gorillas and two humans. Genomic DNA was transferred to Zeta-Probe® membranes (BIO-RAD). "Genomic blots" were hybridized at least 16 hours in QuikHyb® (Stratagene, La Jolla, CA) at 65ºC and then washed four times (1 minute at room temperature, 1 minute at room temperature, 15 minutes at room temperature, 15 minutes at 65ºC) in 2X SSC/0.1% SDS and then four times (10 minutes each at 65ºC) in 0.1XSSC/0.1%SDS.

PCR and Sequencing

The BAC and cosmid clones used for PCR analysis were grown from single colony isolates in 5mL overnight cultures. The DNA was isolated using Qiagen (Valencia, CA) QIAwell 8 DNA isolation kit and resuspended in water and 1/25 (BAC) or 1/250 (cosmid) of the total volume (~15ng) was used in subsequent PCR assays. Gibbon genomic DNA was isolated from cell lines using PUREGENE® DNA isolation kit (Gentra systems, Minneapolis, MN) according to the manufacturer's recommendations and 100ng was used as a template in PCR assays. Long-range gibbon PCR products (~1

kb) were amplified from primate genomic DNA (*Hylobates Lar* and *Hylobates klossii*), subcloned into PGEM®-T easy cloning vector using the Promega Rapid ligation kit according to the manufacturer's recommendations, transformed into XL1-Blue supercompetent cells (Stratagene, La Jolla, CA), and screened by PCR to identify transformants containing full-length inserts. Positive transformants were then amplified with three sets of forward and reverse primers (32N49, 120N123, and 124N129, see Table 3). All PCR conditions entailed a 2 minute initial denaturation at 95ºC, followed by 35 cycles of: 95ºC for 30 seconds, 55ºC for 30 seconds and 72ºC for 45 seconds followed by a final extension at 72ºC for 7 minutes and then a 4ºC hold. Long-range PCR using primers 62N64 were amplified as above with a 60 second extension time at 72ºC for each of 35 cycles. PCR products were directly sequenced using the forward and reverse primers following a modified dye-terminator sequencing protocol (Horvath et al. 2000). To remove single-stranded DNA and deoxynucleotide triphosphates from the PCR after the cycling steps were completed, 8uL of PCR product was treated with 1.50 U exonuclease I and 0.30 U shrimp alkaline phosphatase (Amersham Corporation) at 37ºC for 5 minutes and then heat inactivated at 72ºC for 15 minutes followed by a 4ºC hold. Cycle sequencing conditions were performed in 8µL: 5µL exonuclease I/shrimp alkaline phosphatase treated PCR product, 1µL primer (20µM), and 2µL dichlororhodamine dye-terminator reaction mix (ABI). All fluorescent traces were analyzed using an Applied Biosystems PRISM® 377 DNA Sequencing System (Perkin-Elmer Applied Biosystems, Norwalk, CT) and the quality of the sequence data was assessed with PHRED/PHRAP/CONSED software (http://genome.wustl.edu).

Phylogenetic Analysis

Fasta formatted sequence files for all BAC and cosmid sequences were created after comparison of both forward and reverse sequences of each PCR product using CONSED. Fasta formatted sequence files from accessions used to generate the 1kb and 3kb trees were extracted from the most updated GenBank accession and coordinates are listed in supplemental tables 1 and 2, respectively. Multiple pairwise alignments were generated using CLUSTALW (version 1.82) (Higgins, Thompson and Gibson 1996). For the 1kb tree, 945 bp of PIR4 sequence was generated from 67 human, 2 chimpanzee, 4 orangutan, and 4 gibbon loci (Figure 2). For the 3kb tree, 3000 bp of sequence was extracted from 32 human accessions (Supplemental Figure 1). Phylogenetic analyses were performed using MEGA (Molecular Evolutionary Genetic Analysis) version 2.1 (http://www.megasoftware.net/) (Kumar et al. 2001). Neighbor-joining analysis was used with complete deletion parameters and bootstrap (1000 iterations) to provide confidence of each branchpoint in the phylogenetic trees. We chose to use the neighbor-joining method (although minimum evolution was also used and yielded a tree with similar results) because we were interested in calculating divergence times between sequence taxa and neighbor joining methods were amenable to this task. Also, maximum likelihood and parsimony methods are too cpu-intensive with 77 taxa. Determination of orthologous chimpanzee and orangutan BAC sequences was conducted by BAC end sequence placement with respect to NCBI, build31, (November 2002). Chimpanzee BAC 145P5 end sequences and orangutan BACs 220O24 and 1J18 end sequences placed at positions orthologous to human AC073318 on build31. We were unable using build31 to determine the orthologous placement of chimpanzee BAC 109O6. Because the rates

of nucleotide substitution vary for pseudogenic sequences, the rate of nucleotide substitution was calibrated based on orthologous PIR4 sequence comparisons between human and primate sequences using a divergence of 18 mya for Gibbon-Human and 6 mya for Chimpanzee-Human divergence. Duplication timing events were calculated using the equation T=K/2r (Li 1997). We conducted relative rate tests to determine whether molecular clock estimates would be valid. Relative rate tests were performed in MEGA using human AC073318 and chimpanzee BAC 145P5 or orangutan BAC 220O24 in comparison to orangutan BAC 220O24 or gibbon as an outgroup, respectively. The relative rate test (Tajima's test) was not rejected as chi squared values ranged between 0.5 and 0.27 with a probability of 0.819 to 0.602, respectively.

Fluorescent *in situ* Hybridization

Human metaphase chromosomes (Table 4) were prepared as described previously (Horvath, Schwartz and Eichler 2000) and hybridized with BAC DNA isolated using the Nucleobond® DNA isolation kit from Clontech (Palo Alto, CA) according to manufacturer's recommendations. Human and primate metaphase chromosomes (Figure 3) from *H. sapiens*, *P. troglodytes*, *G. gorilla*, *P. pygmaeus*, *H. lar* and *M. fascicularis* were prepared from lymphoblastoid lines as described previously (Horvath, Schwartz and Eichler 2000).

RESULTS

Identification and Characterization of PIR4 within the Human Genome

PIR4 (pericentromeric interspersed repeat number 4) was initially identified during the sequence characterization of a clone that mapped to the pericentromeric region of human chromosome 2p11 (AC002038 (Horvath, Schwartz and Eichler 2000)). Using reiterative DNA searches against GenBank (see Materials and Methods) a putative full-length copy of this element (49kb) was subsequently identified on chromosome 7 (AC073318). Sequence similarity searches of GenBank (12/15/02) revealed highly identical (90-99% sequence identity) copies of this element (>1kb in length) on more than 170 human accessions representing at least 44 distinct loci. The majority of these genomic sequences (70%) were not assigned to a chromosome within the working draft assembly of the human genome (NCBI, build 31, November 2002). Sequence similarity searches of expressed sequence databases revealed a single significant HSP (E=2e-47) for a potential unprocessed EST that did not have mRNA support (BQ082091.1). Other than this single EST, there was no evidence that the region was transcribed or that it possessed ancestral exon/intron structure. Sequence content analysis of the 49 kb element revealed repeat and GC-content only slightly lower (34.4% GC and 42.5% repeat content) than the genomic average (IHGSC 2001). Among the interspersed repeat classes only LTR and LINE content were slightly (albeit not significantly) increased (LTR; 10.24% vs. 8.29% genome average, and LINE; 24.73% vs. 20.42% genome average). Overall, there is no obvious sequence property of this element that would easily account for its proliferation within the genome.

PIR4 copy number estimates

A variety of methods (Southern analysis, FISH, library depth-of-coverage) were used to estimate copy number in the genome. Initially, a "unique" 300 bp PCR (32N49) amplicon was designed specific to the repeat and screened against a 25.5 fold redundant human BAC library (RPCI-11, segments 1, 2, 4, and 5). We obtained 768 strongly hybridizing positives, suggesting there were at least 30 copies of this element in the human genome (Table 1). An independent analysis of depth of coverage using whole-genome shotgun sequence data (Bailey et al. 2002a) showed a 40 (1972/47.2) fold excess of sequence read depth when compared to unique regions of the genome (http://humanparalogy.cwru.edu/). Of all segmental duplications characterized within the human genome, only DNA sequence corresponding to rDNA duplications from acrocentric chromosomes surpassed PIR4 in both depth of coverage and degree of sequence identity. PCR analysis of a monochromosomal hybrid DNA panel confirmed copies of PIR4 on chromosomes 1, 2, 7, 9, 10, 13, 14, 15, 16, 17, 21, 22 and Y (Figure 5a). Subsequent sequence analysis of the amplicons, in many cases, revealed the presence of "heterozygous" sequence signatures (Figure 5b). Since each DNA sample was derived from a monochromosomal somatic cell hybrid source, it is likely that multiple copies of the element are present on many chromosomes. To improve the copy number estimate and to recover genomic clones specific for each chromosome, cosmid libraries from flow-sorted human chromosomes 1, 2, 7, 9, 13, 14, 15, 16, 18, 21 and 22 were hybridized with amplicon 32N49 (Table 1). Based on the depth of coverage for each library, the results suggested that most human chromosomes contained multiple copies of the PIR4 repeat sequence (mean=4 copies per chromosome for chromosomes

with PIR4) with chromosomes 1 and 9 being particularly enriched (estimated 7- 9 copies).

Mapping PIR4 sequences to chromosomes

Two approaches were undertaken to determine the location of PIR4 sequences in the human genome. First, 17 RPCI-11 BACs were individually probed against metaphase spreads of human chromosomes (Table 3, Figure 4). Chromosomes 1, 2, 7, 9-17, 21 and 22 were observed in more than half of all BAC FISH experiments, consistent with monochromosomal hybrid PCR data. Occasionally, two signals were observed on the same chromosome (chromosome 2 Figure 4a, 4c and chromosome 9 Figure 4b and 4c). In both of these cases, the non-centromeric signal (9q12 and 2q21) corresponded to the site of an ancient vestigial centromere recently euchromatized as a result of evolutionary chromosomal rearrangements within the human lineage (Baldini et al. 1993). To exclude the possibility that duplicated sequences flanking PIR4 were primarily responsible for these cross-hybridization results, these experiments were repeated using a chromosome 22 cosmid clone (N20B5, AC093314) which had been sequenced in its entirety and was found to contain PIR4 as its sole duplicon, as well as common repeats that could be easily blocked by $C_{ot}1$-DNA that hybridizes to almost all chromosomes identified in Table 2 (Figure 3a).

Although FISH data confirmed a pericentromeric map location for the vast majority of PIR4-containing BAC clones, it was impossible using this approach to unambiguously assign a specific BAC clone to its chromosome of origin. The variable copy number of

the repeat within specific chromosomes, furthermore, made assignment based on signal intensity unreliable (Table 3 and Figure 4). As a secondary means of resolving the chromosomal location of PIR4 containing clones, we implemented a sequence-based strategy (termed paralogous sequence tagging) (Horvath, Schwartz and Eichler 2000) that depends on the identification and characterization of paralogous sequence variants (PSVs) specific to chromosomes with PIR4. Since the BAC libraries were constructed from two chromosomal haplotypes (maternal and paternal) sequence variants between two BACs may be due to either allelism (one maternal variant and one paternal variant at the same locus) or paralogy (two variants at different loci as the result of a duplication event). In contrast to BAC libraries, cosmid libraries constructed early in the Human Genome Sequencing Project at Lawrence Livermore and Los Alamos National Laboratories were constructed from a single flow-sorted chromosome (from a somatic cell line containing a single human chromosome) and represent in theory a single haplotype, thereby excluding allelism as a possible source for the variation (Trask et al. 1991). Thus, sequence variants between two cosmid clones from the same library identify paralogous, not allelic, copies. Sequence identity matches between BAC and cosmid sequence signatures further allow the assignment of large-insert BAC clones to specific pericentromeric regions. We reasoned that clones identified from each of the cosmid libraries could, therefore, be informative as a mapping resource for these intractable, highly duplicated areas of the genome.

To implement this approach, we selected 205 BACs (RPCI-11) and 176 cosmids for sequence analyses. Each clone was PCR amplified using oligonucleotides specific to the

PIR4 repeat and all PCR products were directly sequenced to obtain a catalogue of sequence signatures that distinguish various contigs of clones. A total of 67 distinct cosmid and BAC-derived sequence signatures were identified. A BAC sequence signature was considered distinct if the number of sequence differences was greater (2 differences/252 bp) than that expected for allelic variation. Only high-quality sequence differences present on both forward and reverse sequencing of the PCR amplicon were considered in this analysis. Twenty-one of these cosmid signatures matched a BAC signature allowing for chromosomal assignment of the BAC and providing an anchor for future sequence assembly. However, 20 BAC signatures were left unassigned to any chromosome and 27 cosmid signatures had no evidence of BAC sequence support suggesting extensive levels of allelic variation for these pericentromeric loci. Using the collection of experimentally derived sequence signatures, sequence similarity searches were performed against both the non-redundant (nr) and high throughput genomic sequences (htgs) divisions of GenBank. Forty-nine of the 67 variants matched an accession in the database (zero or one variant in 252bp), nineteen of which could be unambiguously assigned to a chromosome. In contrast, 20 BAC/cosmid signatures were not represented within GenBank suggesting considerable under-representation of this segment within the current genome assembly and additional sequence tag information (total= 945 bp) was obtained for future sequence comparisons. Further, analysis of the most recent assembly of the human genome (NCBI, build 31, November 2002) using these PIR4 sequences revealed that only 20 of the estimated 49 database copies of PIR4 were currently represented within this assembly, confirming considerable under-representation of these sequences within human genome assemblies. In total, our

analysis allowed the unambiguous chromosomal assignment of 19 distinct PIR4 loci. Sixteen RPCI-11 BAC clones (AC127362, AC127380, AC127381, AC127384, AC127387, AC127389, AC127391, AC127701, AC128674, AC128676, AC128677, AC129338, AC129778, AC129779, AC129782, AC092854) and 3 chromosome 22 cosmids (AC093314, AC103582, AC093091) were placed in the sequence queue. Notwithstanding, half of the clones characterized in this study could not be assigned to a specific chromosome. This may be due to extreme levels of allelic variation, structural heteromorphism or clone gaps within existing libraries.

Analysis of PIR4 flanking sequences

Since FISH analyses indicated that PIR4 occurred exclusively in pericentromeric regions, we tested more directly its association with satellite DNA (classical centromeric DNA markers). A subset (306) of PIR4-containing RPCI-11 BACs was selected for end-sequence analysis. These sequences were then searched against GenBank revealing that at least one end sequence placed within centromeric satellite DNA for 83 of these BACs (27.1%), a significantly higher proportion than expected based on random sampling of human BAC end sequences (<1% satellite repeats). This association with satellite DNA was further supported by analysis of existing Human Genome Project data. Twenty of the thirty-seven (54%) distinct BAC clones, for which finished sequence was available, contained at least 1kb (and most often more than 10 kb) of centromerically associated satellite sequences including HSATII, CER, ALR and GAATG/CATTC (Repeatmasker designations of centromeric DNA). These data are consistent with PIR4 sequences lying

within the euchromatin/heterochromatin transition zone in close proximity to human centromeres.

Similarly, the segmental duplication content within the vicinity of PIR4 loci was assessed by comparing the sequences of the thirty-seven large-insert PIR4 BAC clones that had been completely sequenced and a comparison of the flanking genomic sequences to the segmental duplication database of the human genome (Bailey et al. 2002a). With the exception of alpha satellite containing clones, only AC073318 contained PIR4 as its sole duplication element (Figure 1a). The organization of most clones showed complex patterns of segmental duplications (both inter and intrachromosomally) with the PIR4 sequence most often associated with a larger block of duplicated sequence (Figure 1b and 1c). This organization of duplications embedded within duplications is consistent with the previously proposed two-step model for the origin of pericentromeric duplications (Eichler et al. 1997) (Horvath, Schwartz and Eichler 2000). Based on this analysis, it therefore is not surprising that these clones had ambiguous chromosomal assignments in the public build31. Further, the lack of unique sequence in the vicinity of PIR4 and the high degree of sequence identity among the duplicates indicates that most of the available PIR4 containing sequences within GenBank could not be mapped using traditional methods.

PIR4 as a marker of pericentromeric duplications

Based on the multichromosomal distribution and the pericentromeric specificity of PIR4, we reasoned that this interspersed repeat might serve as an informative phylogenetic

marker to reconstruct the series of evolutionary events that have restructured these regions of the human genome. Moreover, since most of the PIR4 elements were associated with larger blocks of segmental duplication the PIR4 elements might also provide insight into these larger secondary duplication events. This assumes that PIR4 sequences have not been preferential targets of gene conversion and therefore represent "neutral" markers of pericentromeric evolution. To test this assumption, the pairwise genetic distance between each finished copy of PIR4 within GenBank was calculated (Fig. 7). Here, ≥10 kb of aligned sequence was compared to the 25 copies of PIR4 for a total of 234 comparisons. Next, we examined the largest flanking sequence excluding PIR4 and calculated the genetic distance between these duplicated flanks. We then compared the genetic distance of the PIR4 element to the genetic distance of the flanking duplicated material as the difference of these two estimates (See Supplemental Figure 2). A difference of zero (identity) between K values would suggest that both PIR4 and flanking sequences had diverged equally and arose at approximately the same time in evolution. A negative K value would suggest that the PIR4 copies were more similar than the flanking DNA and had therefore undergone conversion events. Because we assessed only flanking (within 7 kb of PIR4) and not nearby duplications (>7 kb away) our sample size was small (18) and we likely excluded some duplications that could have been separated from PIR4 due to secondary rearrangement events. However, since nearly half (7/18) of the PIR4 elements showed genetic distances consistent with those of the flanking duplications (a difference of 0.005 changes/bp or less than 1% difference) many of the PIR4 elements act as a marker of pericentromeric DNA.

Comparative Primate Analysis of PIR4

In order to provide evolutionary points of reference in our analysis of PIR4, we employed complementary molecular and cytogenetic approaches among representative non-human primate species. Colony hybridizations were performed using the 32N49 amplicon as a probe against the chimpanzee (RPCI-43), orangutan (CHORI 253) and baboon (RPCI-41) BAC libraries (Table 1). Numerous BAC clones were identified within the orangutan and chimpanzee libraries, suggesting multiple (albeit a reduced number of) PIR4 copies. Interestingly, hybridization experiments against the baboon BAC library (10.8 X coverage) failed to yield a single positive. Subsequent hybridizations using a larger amplicon as well as Southern hybridization experiments against genomic DNA provided no evidence of PIR4 within the baboon. To determine the accuracy of the copy number estimates for the chimpanzee and orangutan hybridizations, DNA was isolated from all positive BACs and PCR products corresponding to the 32N49 amplicon were sequenced. All sequences were compared within each species to identify the contiguous sets of clones linked by a common set of sequence variants. Through these studies, the 57 amplifying chimpanzee BAC sequences could be grouped into 20 distinct sequence classes (one or more differences within 252 bp sampled) while the 25 orangutan BAC clones fell into only 4 sequence classes.

Since this molecular evidence points to multiple copies of PIR4 in chimpanzee and orangutan, two sets of comparative FISH experiments were undertaken to determine the copy number and distribution of PIR4 sequences on these primate chromosomes. In the first study, a human chromosome 22 cosmid probe, N20B5, which contained a single

copy of the PIR4 sequence was probed against chromosome spreads of chimpanzee and orangutan metaphases (Figure 3a). Multiple pericentromeric signals were observed on chimpanzee chromosomes (I, IIp, VII, X and XVI with respect to the human phylogenetic group designations). In contrast to human and chimpanzee metaphases, a single robust signal was observed in orangutan metaphases, corresponding to phylogenetic group VII. Since the absence of signal on orangutans might presumably be due to sequence divergence, a reciprocal set of experiments was conducted using orangutan BACs as probes on both human and orangutan chromosomes. A representative orangutan BAC from each of the four sequence classes was assessed. Two of the orangutan BACs (CHORI-253 220o24 and CHORI-253 1j18) yielded identical results, hybridizing to a single locus on chromosome VII in both human and orangutan (Figure 3b, BAC 1J18). In contrast, orangutan BAC 346B14 hybridized to a single locus in orangutan (chromosome VII, Figure 3c) but multiple chromosomes in human (1, 2, 7, 14, 16, 17, 21, 22) whereas 321D4 hybridized to two discrete but nearby loci on chromosome VII in orangutan and multiple loci in humans (2, 7, 14, 16, data not shown). BAC-end sequencing and subsequent similarity searches of the orangutan PIR4 containing BAC clones revealed that they mapped to two different positions within the human chromosome 7 reference sequence.

Phylogenetic analyses of PIR4 sequences

As the final step in our analysis, a phylogenetic tree was generated using MEGA2 to compare 67 human, 2 chimpanzee, 4 orangutan, and 4 gibbon loci (Figure 2) (Kumar et al. 2001). At least two major clades could be distinguished. One clade (termed A)

consists almost entirely of human sequences from many different chromosomes. This clade is further stratified into relatively chromosome-specific subgroups of PIR4 (see chromosome 1, 2, 7 and 9) as well as an acrocentric chromosome subclade (13, 14 and 21). In contrast, clade B consists of human, chimpanzee, orangutan, and gibbon sequences as well as the putative ancestral human sequence on chromosome 7 (AC073318). With the exception of chromosome 7, very little evidence of chromosome-specific amplification is observed within this clade. It should be noted that chromosomes 2, 7, 13, 16, and 22 have representative sequences in both clade A and B. In order to increase confidence of the two separate clades on the 1kb tree, we generated a 3kb tree from a subset of the accessions (Supplemental Figure 1). This increased bootstrap support from 80% to 100% for the existence of two clades. In total, these data suggest a rapid dispersal of PIR4 sequences over a narrow window of primate evolution followed by more recent chromosome-specific duplication events. To examine this in more detail, another phylogenetic tree was constructed from a shorter multiple sequence alignment (252bp) incorporating an additional 18 distinct chimpanzee BAC sequences. These chimpanzee sequences distributed throughout clade A and B showing, in general, closer phylogenetic relationship to other human loci rather than other chimpanzee sequences (data not shown). Thus, it is likely that PIR4 sequences populated the hominoid genome prior to the divergence of the two lineages.

DISCUSSION

In the absence of a robust genome assembly near centromeric regions, we conducted a global analysis of half of all human pericentromeric regions using a single element

(PIR4) providing insight into the biology of these complex regions of our genome. Within the human genome, we estimate approximately 40 copies of this (20-40 kb) element, which share, on average, 95.2 % sequence identity (range 90.2-99.5% sequence identity) (Table 1). The available data suggest that PIR4 represents one of the most prolific and highly homologous segmental duplications within the human genome. Cytogenetic and molecular evidence confirm that the repeat localizes almost exclusively to the pericentromeric regions of more than half of all human chromosomes (1p, 2p, 7p, 9p, 9q, 10q, 13q, 14q, 15q, 16p, 17q, 18q, 21q, 22q and Ypcen, Table 1, Table 3, Figure 4). With the exception of the ancestral copy on AC073318 from 7p12 and a few copies flanking alpha satellite DNA, all PIR4 elements map within 100kb of other duplicated segments. PIR4 elements themselves almost always are a component of a larger duplication block with a more limited pericentromeric distribution pattern. Interestingly, the evolutionary age of the PIR4 sequences was often consistent with the evolutionary age of the duplicated flanking sequencing. This suggested that phylogenetic analysis of PIR4 would not only be valuable in reconstructing the evolutionary history of the repeat but would also provide insight into the series of large-scale duplications which have reshaped hominoid pericentromeric regions.

PIR4 ancestral sequence

Several lines of evidence point to chromosome 7 as the ancestral origin of PIR4. First, it is the only chromosome commonly hybridizing to human, chimpanzee and orangutan metaphase spreads (Figure 3a-c). Second, it is the only locus for which a clear orthologue can be identified within each great ape species examined (Figure 2). This is

supported both by phylogeny as well as BAC-end sequence analysis. Third, it is one of the only PIR4 containing loci (GenBank AC073318) devoid of other segmental duplications. Characterization of numerous other segmental duplications (Eichler et al. 1996) (Eichler et al. 1997) (Regnier et al. 1997) (Crosier et al. 2002; Zimonjic et al. 1997) (Horvath et al. 2001) suggest that the progenitor loci most often occur outside of the pericentromeric duplication zone surrounded by unique sequence. Subsequent duplicative transposition events become associated with other pericentromeric duplications. Finally, size estimates of PIR4 from AC073318 indicate that it represents the largest and most complete copy (49 kb). Other copies of PIR4 have become truncated with respect to this locus perhaps as a result of deletion of secondary progenitors prior to subsequent rounds of duplication (Figure 6). For example, AC093787 from chromosome 2 contains 44.2kb of PIR4 while AC025223 from chromosome 2 has 17.3kb and AC073210 and AC104057 from chromosome 7 contain 16.2kb and 26.3kb, respectively. While the extent of PIR4 rearrangement with respect to the putative ancestral locus (AC073318) is not always a good indicator of degree of nucleotide sequence identity, it is noteworthy that similar deletion patterns share monophyletic origins consistent with the placement within the phylogenetic tree (see AC128674, AC127384, AC0024500, AC006359 for an example, Fig. 6). These data not only validate the phylogeny but provide insight into different trajectories of evolutionary duplication, where irreversible deletion/rearrangement events tagged a progenitor copy and its descendants. Finally, the phylogenetic data are consistent with a strict division of PIR4 sequences into two clades, an ancestral (clade B) which contains the putative human donor locus as well as

representatives from each great-ape species and derivative clade (clade A) which contains only chimpanzee and human copies of this element.

PIR4 duplication timing

Since gibbon and orangutan lineages contain only four copies of PIR4 while chimpanzee and human have 20 and 40 copies, respectively, the data suggest that a major transpositional burst of PIR4 sequences likely occurred prior to the divergence of the African great ape and human lineage (5-8 million years ago). In order to determine a more precise estimate of when the burst of PIR4 duplications occurred, we first calculated the specific neutral substitution rate for this duplication since previous molecular clock estimates among primates have varied greatly ($1X10^{-9}$ mutations/site/year for human-chimpanzee vs. $2X10^{-9}$ mutations/site/year for human-lemur comparisons (Liu et al. 2003)). Using gibbon sequences as an outgroup (which diverged from the lineage leading to humans approximately 18mya) (Goodman 1999), and the average K value (0.068) between all gibbon and all human sequences, the rate of neutral substitution for this repeat is $1.89X10^{-9}$ (+/- $0.17X10^{-9}$) (r=k/2T). Similarly, calculating the rate based on chimpanzee 145P5 and its inferred orthologue AC073318 gave a result of $1.75X10^{-9}$ (+/-$0.7X10^{-9}$) (based on a separation of 6 million years between human and chimpanzee (Goodman 1999)) while the rate determined using the orthologous orangutan BACs (1J18 and 220O24) is $1.5X10^{-9}$ (+/-$0.25X10^{-9}$). The higher substitution rates of $1.89X10^{-9}$ and $1.75X10^{-9}$ seen for this repeat agree with the $2.1X10^{-9}$ estimated for the Old World Monkey comparison of the CAGGG (Eichler, Archidiacono and Rocchi 1999) and could indicate that pericentromeric repeats and other sequences

devoid of genes may have different rates of substitution than previously determined for non-coding sequences (Li and Tanimura 1987).

Using the average rate (between $1.5X10^{-9}$, $1.75\ X10^{-9}$, and $1.89X10^{-9}$) of $1.71X10^{-9}$ mutations/site/year, the most divergent human sequences in Figure 2 (AC127701B and AC073318) have a K value of 0.087 suggesting approximately 25 million years of change between them. Interestingly, the phylogenetic tree in Figure 2 has some sequences clustered together (for example, 7cos43b6 and 7cos33f1), further suggesting recent intrachromosomal duplication or conversion events. This suggests that while some PIR4 copies have existed for over 20Myr of evolution, others have arisen recently and the process of PIR4 duplication may be ongoing. However, the mean genetic distance between all human sequences is 0.047 (Kimura's estimate) suggesting that many duplications occurred 14 million years ago (just before the divergence of humans from our Great Ape ancestors) (Figure 7). Surprisingly, sequences from chromosomes 2, 7, 13, 16 and 22 are found in both clades of the tree suggesting very different evolutionary histories exist on the same chromosome.

Consequences of PIR4 duplications

Of the pericentromeric regions identified in this study, most harbor multiple copies of PIR4 (Table 1). Based on available data within GenBank, it appears that intrachromosomal copies of PIR4 are separated by at least 100-150 kb, as BAC clones rarely contain two distinct elements. Based on their high identity and close proximity within pericentromeric regions, PIR4 elements have the potential to undergo gene

conversion. This is supported in part by our analysis of cosmid and BAC PSV signatures that we were sometimes unable to match to one another suggesting that PIR4 elements have rapidly diverged between individuals or have been effectively deleted within the population (see supplemental table 3). In some cases, such as chromosome 9p/9q12 and 2p/2q21, individual copies of PIR4 may be separated by multiple Mb as evidenced by distinct metaphase FISH signals. This organization is presumably due to recent evolutionary centromeric rearrangements that have occurred within these two specific chromosomes (Baldini et al. 1993). The organization of these intrachromosomal copies of PIR4 is reminiscent of low-copy repeat (LCRs) sequences that have been implicated in chromosomal instability associated with more than two dozen genomic disorders. It is possible that intrachromosomal PIR4 sequences separated by 100's of kb could similarly facilitate non-homologous recombination events leading to secondary deletions, duplications and inversions (Stankiewicz and Lupski 2002) (Bailey et al. 2002a). Such dynamic mutational events, if they exist, might account for the considerable heteromorphism observed for these regions of the genome (Buiting et al. 1992) (Barber et al. 1998) (Barber et al. 1999). Although the clinical and evolutionary significance of such germ line/somatic instability is unknown, it is noteworthy that many of the same pericentromeric regions containing PIR4 duplications (1, 2, 8, 9, 14, 15, 16, 17, 18, 21, 22) are regions associated with common breakpoints in solid tumor cell lines suggesting that the presence of PIR4 may be associated with somatic instability (Padilla-Nash et al. 2001). Finally, the unusual architecture of PIR4 repeats on chromosome 2 and 9 could help explain the high frequency of large-scale inversions. Chromosome 9 inversion events are the most common karyotype variation seen in humans while chromosome 2

inversion events are the second most commonly diagnosed event (Kaiser 1984). Although PIR4 has not yet been directly implicated in these common rearrangements, its existence in many regions of instability necessitates a more thorough investigation of the genomic architecture.

Based on BAC end sequencing data as well as large-scale sequencing of PIR4-containing clones, we estimate that ~25% of PIR4 copies abut large tracts (>10 kb) of satellite repeat sequences (alpha, HSATII, etc). Such repetitive sequences have been postulated to play a pivotal role in the recent non-homologous exchanges that have dynamically shaped human pericentromeric regions (Horvath, Schwartz and Eichler 2000) (Guy et al. 2000; Mashkova et al. 1998) as they often demarcate the boundaries of large-scale interchromosomal duplications. The proximity of PIR4 sequences to blocks of satellite may have contributed to their proliferation within the human genome. Of the chromosomes known to contain pericentromeric duplications (Bailey et al. 2001) (Cheung et al. 2001), detailed pericentromeric analyses have only been conducted for chromosomes 2, 10, 16 and the completely sequenced chromosomes 14, 20, 21 and 22 (Horvath, Schwartz and Eichler 2000; Horvath et al. 2000) (Deloukas et al. 2001; Jackson et al. 1999) (Dunham et al. 1999; Hattori et al. 2000) (Heilig et al. 2003). Our analysis predicts that many duplicon-rich pericentromeric regions, such as chromosomes 1, 5, 7, 9, 11, 12, 13, 15, 17, 18, and Y, still remain uncharacterized with respect to the full extent of their duplicated architecture. Interestingly, even among chromosomes that have been deemed completed (21, 22 and 14), our analysis has identified additional clones that have

not yet been sequenced. Presumably, these clones map centromerically to the most proximal sequence within the sequence assembly.

Using PIR4 to fill genome gaps

Utilizing PIR4 as a marker of pericentromeric DNA, we have used paralogous-sequence tagging to begin to successfully map ~40% of these relatively intractable regions of the genome. In addition, our analysis recovered additional candidate clones for targeted sequencing. As part of a collaboration with the Washington School of Medicine Genome Sequencing Center, we have submitted an additional 15 RPCI-11 BAC clones whose sequence signature did not match an accession within the NCBI database (at least 3 variants over 950bp of sequence analyzed) (AC127362, AC127380, AC127381, AC127384, AC127387, AC127389, AC127391, AC127701, AC128674, AC128676, AC128677, AC129338, AC129778, AC129779, AC129782). In collaboration with Oklahoma's Advanced Center for Genome Technology, we have sequenced one RPCI-11 BAC clone (AC092854) as well cosmid clones from chromosome 22 (AC093314, AC103582, AC093091). These clones have effectively added over 2Mb of human pericentromeric sequence to GenBank, although their integration into the final human genome assembly is still ongoing. In cases where chromosome-assigned pericentromeric clones have been dropped during the assembly process, we are working with the sequence community to ensure that such clones are reincorporated into the minimal tiling path of the final human genome sequence. While it is unlikely that complete closure of these regions will be achieved by the finish target date (2003), these sequences should

provide valuable anchor points from which to seed future mapping, sequencing and assembly.

Is the additional effort within these regions warranted? Although biological and evolutionary arguments may be easily mustered, the primary motivation of the Human Genome Project has been to identify all genes within the context of its genomic sequence (Collins et al. 1998). Many pericentromeric regions have been recalcitrant to closure due to their unusual duplication architecture. Pericentromeric genes embedded within these highly duplicated regions have been difficult to identify because of a lack of available sequence, difficulties in assembly of underlying genomic DNA and/or ambiguities of paralogous gene annotation. Furthermore, pericentromeric regions have been operationally classified as heterochromatic DNA, since they are located in the vicinity of centromeres. As such, they are considered gene-poor genomic environments. While heterochromatin is typically devoid of transcription presumably due to its compact nature(Donze and Kamakaka 2002) (Dillon and Festenstein 2002), several recent studies have challenged the notion that DNA sequence in the vicinity of heterochromatic DNA is transcriptionally silent. For example, a mammalian artificial chromosome study indicated that a gene placed in close proximity to and between centromeric and telomeric satellites can still be readily expressed (Bayne et al. 1994). Within Drosophila, essential genes such as the MAP-kinase were recovered embedded within satellite sequences (Adams et al. 2000). Similarly, recent articles by Crosier et al. (2002) and Bailey et al. (2002) provide strong evidence of human transcripts from pericentromeric regions on chromosomes 2 and 22. Our own analysis of 89 GenBank accessions containing (>5 kb)

PIR4 reveals that 31 of these genomic sequences contain at least one transcript (exon-intron structure over at least 2 exons >99% identity to an EST). Seven out of these 31 accessions also contain tracts of satellite sequence (>3 kb). These transcripts map to 19 different Unigene clusters that have been assigned to chromosomes 2, 7, 9, 10, 16 and 22. Although these data do not prove the existence of pericentromerically located genes associated with PIR4 in humans, they do suggest transcriptional potency of these genomic regions. This underscores the importance of complete human genome sequence and assembly up to the higher order alpha satellite arrays in order to provide a comprehensive transcription and, ultimately, a gene map of the human genome.

ACKNOWLEDGEMENTS

We thank Laurie Christ for technical assistance and Dr. Norman Doggett for kindly providing access to chromosome 16 cosmid filters and clones. This work was supported by grants NIH GM58815, NIH HG002385, DOE ER62862 to EEE, and NIH HG02152 to BAR, the financial support of Telethon, CEGBA (Centro di Eccellenza Geni in campo Biosanitario e Agroalimentare), MIUR (Ministero Italiano della Universita' e della Ricerca), European Commission (INPRIMAT, QLRI-CT-2002-01325) to N.A. and M. R. JEH was supported in part by NIH GM08613, Genetics Training grant and JAB was supported by NIH Career Development Program in Genomic Epidemiology of Cancer (CA094816) and Medical Scientist Training Grant.

LITERATURE CITED

Adams, M.D. S.E. Celniker R.A. Holt et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-2195.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Molec. Biol.* **215:** 403-410.

Arnold, N., J. Wienberg, K. Emert, and H. Zachau. 1995. 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.

Assum, G., T. Fink, C. Klett, B. Lengl, M. Schanbacher, S. Uhl, and G. Wohr. 1991. A new multisequence family in human. *Genomics* **11:** 34-41.

Bailey, J.A., Z. Gu, R.A. Clark et al. 2002a. Recent segmental duplications in the human genome. *Science* **297:** 1003-1007.

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

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

Baldini, A., T. Ried, V. Shridhar, K. Ogura, L. D'Aiuto, M. Rocchi, and D.C. Ward. 1993. An alphoid DNA sequence conserved in all human and great ape chromosomes: evidence for ancient centromeric sequences at human chromosomal regions 2q21 and 9q13. *Hum Genet* **90:** 577-583.

Barber, J.C., I.E. Cross, F. Douglas, J.C. Nicholson, K.J. Moore, and C.E. Browne. 1998. Neurofibromatosis pseudogene amplification underlies euchromatic cytogenetic duplications and triplications of proximal 15q. *Hum Genet* **103:** 600-607.

Barber, J.C., C.J. Reed, S.P. Dahoun, and C.A. Joyce. 1999. Amplification of a pseudogene cassette underlies euchromatic variation of 16p at the cytogenetic level. *Hum Genet* **104:** 211-218.

Bayne, R.A., D. Broccoli, M.H. Taggart, E.J. Thomson, C.J. Farr, and H.J. Cooke. 1994. Sandwiching of a gene within 12 kb of a functional telomere and alpha satellite does not result in silencing. *Hum Mol Genet* **3:** 539-546.

Buiting, K., V. Greger, B.H. Brownstein, R.M. Mohr, I. Voiculescu, A. Winterpacht, B. Zabel, and B. Horsthemke. 1992. A putative gene family in 15q11-13 and 16p11.2: possible implications for Prader-Willi and Angelman Syndromes. *Proc Natl Acad Sci USA* **89:** 5457-5461.

Cheung, V.G., N. Nowak, W. Jang et al. 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. The BAC Resource Consortium. *Nature* **409:** 953-958.

Collins, F.S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. 1998. New goals for the U.S. Human genome project: 1998-2003. *Science* **282:** 682-689.

Copenhaver, G.P., K. Nickel, T. Kuromori et al. 1999. Genetic definition and sequence analysis of Arabidopsis centromeres. *Science* **286:** 2468-2474.

Crosier, M., L. Viggiano, J. Guy et al. 2002. Human paralogs of KIAA0187 were created through independent pericentromeric-directed and chromosome-specific duplication mechanisms. *Genome Res* **12:** 67-80.

Deloukas, P. L.H. Matthews J. Ashurst et al. 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865-871.

Dillon, N. and R. Festenstein. 2002. Unravelling heterochromatin: competition between positive and negative factors regulates accessibility. *Trends Genet* **18:** 252-258.

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

Dunham, I., N. Shimizu, B.A. Roe et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489-495.

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

Eichler, E.E. 2001a. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* **17:** 661-669.

Eichler, E.E. 2001b. Segmental duplications: what's missing, misassigned, and misassembled- and should we care? *Genome Res* **11:** 653-656.

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

Eichler, E.E., F. Lu, Y. Shen et al. 1996. Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum Molec Genet* **5:** 899-912.

Flint, J., K. Thomas, G. Micklem, H. Raynham, K. Clark, N. Doggett, A. King, and D. Higgs. 1997. The relationship between chromosome structure and function at a human telomeric region. *Nature Genet.* **15:** 252-257.

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

Guy, J., C. Spalluto, A. McMurray 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.

Hattori, M., A. Fujiyama, T.D. Taylor et al. 2000. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405:** 311-319.

Heilig, R., R. Eckenberg, J.L. Petit et al. 2003. The DNA sequence and analysis of human chromosome 14. *Nature* **421:** 601-607.

Higgins, D.G., J.D. Thompson, and T.J. Gibson. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266:** 383-402.

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

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

Horvath, J.E., J.A. Bailey, D.P. Locke, and E.E. Eichler. 2001. Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum Mol Genet* **10:** 2215-2223.

IHGSC. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

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

Jackson, M.S., M. Rocchi, G. Thompson 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.

Kaiser, P. 1984. Pericentric inversions. Problems and significance for clinical genetics. *Hum Genet* **68:** 1-47.

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., K. Tamura, I.B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244-1245.

Li, W. 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Li, W.H. and M. Tanimura. 1987. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326:** 93-96.

Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13:** 358-368.

Loftus, B., U. Kim, V. Sneddon et al. 1999. Genome duplications and other features in 12 Mbp of DNA sequence from human chromosome 16p and 16q. *Genomics* **60:** 295-308.

Mahtani, M.M. and H.F. Willard. 1998. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res* **8:** 100-110.

Mashkova, T., N. Oparina, I. Alexandrov, O. Zinovieva, A. Marusina, Y. Yurov, M.H. Lacroix, and L. Kisselev. 1998. Unequal cross-over is involved in human alpha satellite DNA rearrangements on a border of the satellite domain. *FEBS Lett* **441:** 451-457.

Myers, E.W. and W. Miller. 1988. Optimal alignments in linear space. *Comput Appl Biosci* **4:** 11-17.

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

Padilla-Nash, H.M., K. Heselmeyer-Haddad, D. Wangsa et al. 2001. Jumping translocations are common in solid tumor cell lines and result in recurrent fusions of whole chromosome arms. *Genes Chromosomes Cancer* **30:** 349-363.

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

Riethman, H.C., Z. Xiang, S. Paul, E. Morse, X.L. Hu, J. Flint, H.C. Chi, D.L. Grady, and R.K. Moyzis. 2001. Integration of telomere sequences with the draft human genome sequence. *Nature* **409:** 948-951.

Stankiewicz, P. and J.R. Lupski. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18:** 74-82.

Sun, X., J. Wahlstrom, and G. Karpen. 1997. Molecular structure of a functional *Drosophila* centromere. *Cell* **97:** 1007-1019.

The Arabidopsis genome initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Trask, B.J., G. van den Engh, M. Christensen, H.F. Massa, J.W. Gray, and M. Van Dilla. 1991. Characterization of somatic cell hybrids by bivariate flow karyotyping and fluorescence in situ hybridization. *Somat Cell Mol Genet* **17:** 117-136.

Wohr, G., T. Fink, and G. Assum. 1996. A palindromic structure in the pericentromeric region of various human chromosomes. *Genome Res* **6:** 267-279.

Yan, C.M., K.W. Dobie, H.D. Le, A.Y. Konev, and G.H. Karpen. 2002. Efficient recovery of centric heterochromatin P-element insertions in Drosophila melanogaster. *Genetics* **161:** 217-229.

Zimonjic, D., M. Kelley, J. Rubin, S. Aaronson, and N. Popescu. 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 USA* **94:** 11461-11465.

Table 1. Genomic Distribution of PIR4 in Representative Primates

Library	Library name	coverage	Positives	PCR amplified	Estimated copy number	# PSVs
Human BAC	RPCI-11*	25.5x	768	205	30.1	43+
1 cosmid	LL01	4x	95**	43	10.8	9
2 cosmid	LL02	3.9x	40**	19	4.9	3
7 cosmid	LL07	4.8x	32**	16	3.3	7
9 cosmid	LL09	5.6x	43**	37	6.6	7
13 cosmid	LA13	7x	16**	15	2.1	4
14 cosmid	LA14	5x	33**	29	5.8	6
15 cosmid	LA15	6x	5**	5	0.8	1
16 cosmid	LA16	5.9x	51**	25	4.2	8
18 cosmid	LL18	5.7x	5**	0	0	0
21 cosmid	LL21	7.4x	8**	4	0.5	1
22 cosmid	LL22	17.1x	39**	24	1.4	4
Chimp BAC	RPCI-43	3.5x	82	57	23.4	20+
Orangutan BAC	CHORI-253	6.4x	33	25	3.9	4
Baboon BAC	RPCI-41	10.4x	0	0	0	0

A summary of the number of positives identified by radioactive colony hybridization of human, chimpanzee, orangutan and baboon BAC and human cosmid libraries using a 300 bp PCR product as a probe (amplicon 32N49; see Table 2). All positively hybridizing clones were screened by PCR (with the exception of RPCI-11 where only a subset of the positive BACs were chosen for further analyses). Copy number estimates for the BAC libraries were calculated by dividing the number of positives by the fold coverage of the library. Copy number estimates for the cosmid libraries were estimated by dividing the number of cosmids PCR amplifying by the fold coverage of the library. Direct sequencing was performed on all PCR products (Materials and Methods). Based on the genomic coverage of each library and the number of positive BACs or positive cosmids that successfully amplified by PCR, an estimate of the copy number (expected) of PIR4 in each library was calculated. The number of PSVs (paralogous sequence variants) observed was determined by comparison of all directly sequenced PCR products to determine the number of distinct variants represented in each library. Occasionally a BAC sequence would have heterozygous peaks suggesting there were two copies of PIR4 within a single BAC. *Analyzed segments 1, 2, 4, 5. **number of true positives unknown because no false positive list available. + some BACs have at least two copies of PIR4 which could not be distinguished by direct sequencing. PSV (Paralogous Sequence Variant)-A distinct sequence signature when compared to other paralogous copies.

Table2: PCR oligonucleotide sequences

32	F	CAGTATCTTCACATTCTCTCCCTGTCC
49	R	GAAAGAAGCAAGAGTGCGCTAAAC
62	F	TCCTCTCAGGTGGGAGAATTGTTG
64	R	CCACCAGTTGACAGGCAAAGTTCT
120	F	GTGCTTGAGGTAAATAGGAGAAAC
123	R	CCACAGAAAAGACTCAAGACCACC
124	F	GTACTCCAAATCAGTACTGCTCAC
125	F	GTTTAGCGCACTCTTGCTTCTTTC
129	R	GGGAGCTCTTTAATAACATAAAC

All oligonucleotides were designed based on the 2p11 BAC reference sequence (AC002038). Sequences are presented in 5' to 3' orientation; Both forward 'F' and reverse 'R' oligonucleotides for each assay are presented.

Table 3: FISH localizations of PIR4 positive BACs

BAC	Library	GenBank Accession	Map Location	Metaphase FISH locations																	
101B6	CIT	AC002038	2p11	1cen	2cen*	4q24	---	7pcen	---	9p/qcen	10cen	---	---	---	16p11	17q11	---	---	---	21cen	22q11	Ycen
2i21	RPCI-11	none	unk	1cen	2cen	---	---	7cen*	---	9p/qcen	10cen	13q11	14pcen	15q11	16cen	17cen	18cen	19cen	20cen	21cen	22cen	---
2053H7	CIT	AC025223	2	1cen	2cen*	---	---	7cen	---	9cen	10cen	13cen	14cen	15cen	16cen	17cen	18cen	19cen	---	21cen	22cen	---
168j1	RPCI-11	AC034151	2	1qh	2pcen*	4q24	---	7cen	---	9cen	10cen	13cen	14cen	15cen	16pcen	---	---	---	---	21cen	22cen*	Ycen
165d20	RPCI-11	AC027612	2	1cen/qh*	2cen*	---	---	7cen/qter	---	9cen	---	13cen	14cen	15cen	16cen	---	---	---	---	21cen	22 doublet	---
28o7	RPCI-11	AC129782	1	1p32/ 1qh*	2p11/2q14	---	---	7cen	---	9p11	---	13cen	14cen	15cen	16p11/q 11	17cen	18cen	---	---	21cen	22cen	---
51a19	RPCI-11	AC129338	7	1cen	2cen	---	---	7cen*	---	---	---	---	---	---	16p11	---	---	---	---	---	---	---
1429e17	RPCI-11	none	1or7	1qh	2p/qcen	---	---	7cen/qter	---	9qh	10qtel	13cen*	14cen	15cen	16cen	17cen	---	---	---	21cen*	22cen*	---
1390m18	RPCI-11	none	1or9	1p/qh*	---	---	---	---	---	---	---	---	---	---	16pcen	---	---	---	---	---	---	---
1386h14	RPCI-11	none	1	1qh*	2cen*	4pter	---	7cen	---	9qh	10cen	13pcen	14pcen	15qcen	16cen	17cen	---	---	---	21cen	22cen	Ycen
3m10	RPCI-11	none	13	---	---	---	---	---	---	---	---	13cen*	14pcen	15cen*	---	---	---	---	20cen	21pcen	22pcen	---
1360m22	RPCI-11	AC127381	15	1qh	2cen/q21	---	---	7cen	---	9qh	10cen	13cen	14cen/ qter	15qcen*	16cen*	17cen	18cen	---	---	21cen	22cen*	---
1391n9	RPCI-11	none	22	1qh	2cen/q21	---	---	7cen	8cen	---	10cen	13cen	14cen	---	16qh*	17cen	18cen	---	---	21cen	22cen*	---
1390a11	RPCI-11	AC127384	16	1p36/ pcen/ qh*	2pcen	---	5cen	7cen	---	---	---	13cen	14cen	---	16pcen	---	---	---	---	21pcen	22pcen	---
1221g12	RPCI-11	AC129778	unk	1qh	2cen/q21*	---	---	7cen	---	9qh/p*	10cen	13cen	14cen	15cen	16pcen	---	---	---	---	21cen	22cen	Ycen
1360o11	RPCI-11	none	unk	1qh	2pcen	---	---	7pcen	---	---	---	---	14cen	---	16pcen*	---	---	---	---	---	22pcen	---
1363e3	RPCI-11	none	2	1qh	2p/qcen*	4q26	---	7cen	---	9qh	10cen	13cen	14cen	15cen	16pcen	---	---	---	---	---	22qcen*	Ycen

CIT D California Institute of Technology, Library D

RPCI-11 Roswell Park Cancer Institute

*Largest signal(s)

To assess genomic distribution of PIR4, 17 RPCI-11 (human) BACs were selected as probes for FISH against human metaphase chromosomes. Accession numbers and chromosomal placement (determined if a BAC PSV matches a cosmid PSV) are indicated when known. Chromosomes 1, 2, 7, 9-17, 21 and 22 were observed in more than half of all BAC FISH experiments.

FIGURE LEGENDS

Figure 1: The genomic organization of sequences flanking PIR4.

Three examples of the duplication architecture surrounding PIR4 loci are shown. For each, the horizontal black line depicts the genomic sequence drawn to scale (tick marks occur every 20 kb along the sequence). Colored boxes above and below the lines demarcate different duplicons/repeat segments as identified by RepeatMasker and BLAST searches (Methods). Black rectangles represent PIR4 loci, gray rectangles represent satellite sequences, and colored rectangles indicate duplicons shared by two or more genomic loci. Slanted black lines between accessions indicate regions containing PIR4 while blue lines indicated regions duplicated between two or more genomic loci. The program PARASIGHT (Bailey, unpublished) was used to generate this output. A) A comparison of ancestral AC073318 sequence to AC127380 indicates that these two accessions have only the 35 kb of PIR4 in common. While PIR4 on AC073318 is flanked by "unique" sequence, AC127380 is flanked by "unique" sequence on one side and a stretch of nearly 160 kb of satellite sequences on the other side. B) Comparing two chromosome 2 loci to ancestral AC073318 indicates that both chromosome 2 loci share more sequence in common than PIR4 alone. Although PIR4 is truncated at the end of both chromosome 2 clones, on one side PIR4 is flanked by a duplicated genomic segment common to both clones. Both of these clones also contain duplicons not shared by the other and AC026273 contains ~40 kb of alpha satellite sequences. Ancestral loci are indicated below each colored rectangle as well as percent identity and length of alignment. Interestingly, AC026273 contains a 100% match over 5 exons to EST BQ651044 (Crosier et al. 2002). C) Two loci from different chromosomes containing

PIR4 indicate that PIR4 is often flanked by duplicons shared by different chromosomes. These two loci on chromosome 13 and 21 share almost 100 kb of sequence that is composed of numerous duplicons including PIR4.

Figure 2: Phylogeny of PIR4.

A neighbor-joining phylogram rooted on gibbon using ~950 bp of PIR4 sequence from human, chimpanzee, orangutan and gibbon sequences was constructed using MEGA (Materials and Methods). Bootstrap values >80% from 1000 replicates are indicated on each respective branch. The branch separating clades A and B has a bootstrap value of 80 for the 1kb tree and 100 for the 3kb tree (supplemental figure 1) indicating high confidence. Sequences with an asterisk indicate those that have been mapped to a chromosome based on cosmid support. Colored boxes behind human sequences indicate chromosomal location, when known. Non-human primate sequences generated from BAC clones for chimpanzee and orangutan and genomic DNA for gibbon are shaded gray. Two clades are readily distinguished (clade A and B). Most non-human primate sequences (orangutan and gibbon) as well as the putative ancestral locus on chromosome 7, map to clade B.

Figure 3: Primate metaphase FISH of PIR4 containing clones.

A) A human chromosome 22 cosmid (N20B5, AC093314) containing only the PIR4 duplicon was hybridized to human (Hsa), common chimpanzee (Ptr), and orangutan (Ppy) metaphase spreads after $C_{ot}1$ blocking. In humans, this probe hybridized to pericentromeric regions of chromosomes 1, 2, 7, 9, 13, 14, 15, 16, 17, 18, 21 and 22

while in chimpanzee it hybridized to pericentromeric regions of chromosomes I, IIp, VII, X and XVI. In contrast, this probe hybridized only to syntenic chromosome VII in orangutan, the putative ancestral locus. All chromosomal designations are with respect to the Human phylogenetic group designations (ISCN 1985). B) Orangutan BAC CHORI-253-1J18 containing PIR4 sequences hybridizes only to chromosome 7 in human and orangutan while C) orangutan BAC CHORI-253-346B14 hybridizes solely to chromosome VII in orangutan but multiple pericentromeric regions (chromosomes 1, 2, 7, 14, 16, 17, 21 and 22) in humans.

Figure 4: Examples of metaphase FISH from PIR4-containing BACs.
FISH was conducted for all BACs listed in Table 2. The results for a subset are shown here. All BACs hybridize to multiple pericentromeric loci. Signal intensity alone is not a good indicator of chromosomal origin since some chromosomes consistently demonstrate large signals (chromosome 1 in panel b) while others show discrete signals flanking the centromere suggesting more than one copy of PIR4 exists on certain chromosomes (chromosome 2 in panel a and c and chromosome 9 in panels b and c).

Figure 5: Genomic distribution of PIR4.
A) PCR analysis using primer pair 32N49 against a monochromosomal somatic cell hybrid panel of DNAs. Chromosomes 1, 2, 7, 9, 10, 13, 14, 15, 16, 17, 21, 22 and Y amplified a product of identical size (~300bp,Gibco-BRL 100bp ladder). B) The chromosome 2 monochromosomal hybrid DNA chromatogram shows 4 putative variant sites suggestive of multiple copies of PIR4 on chromosome 2. Subsequent sequencing of

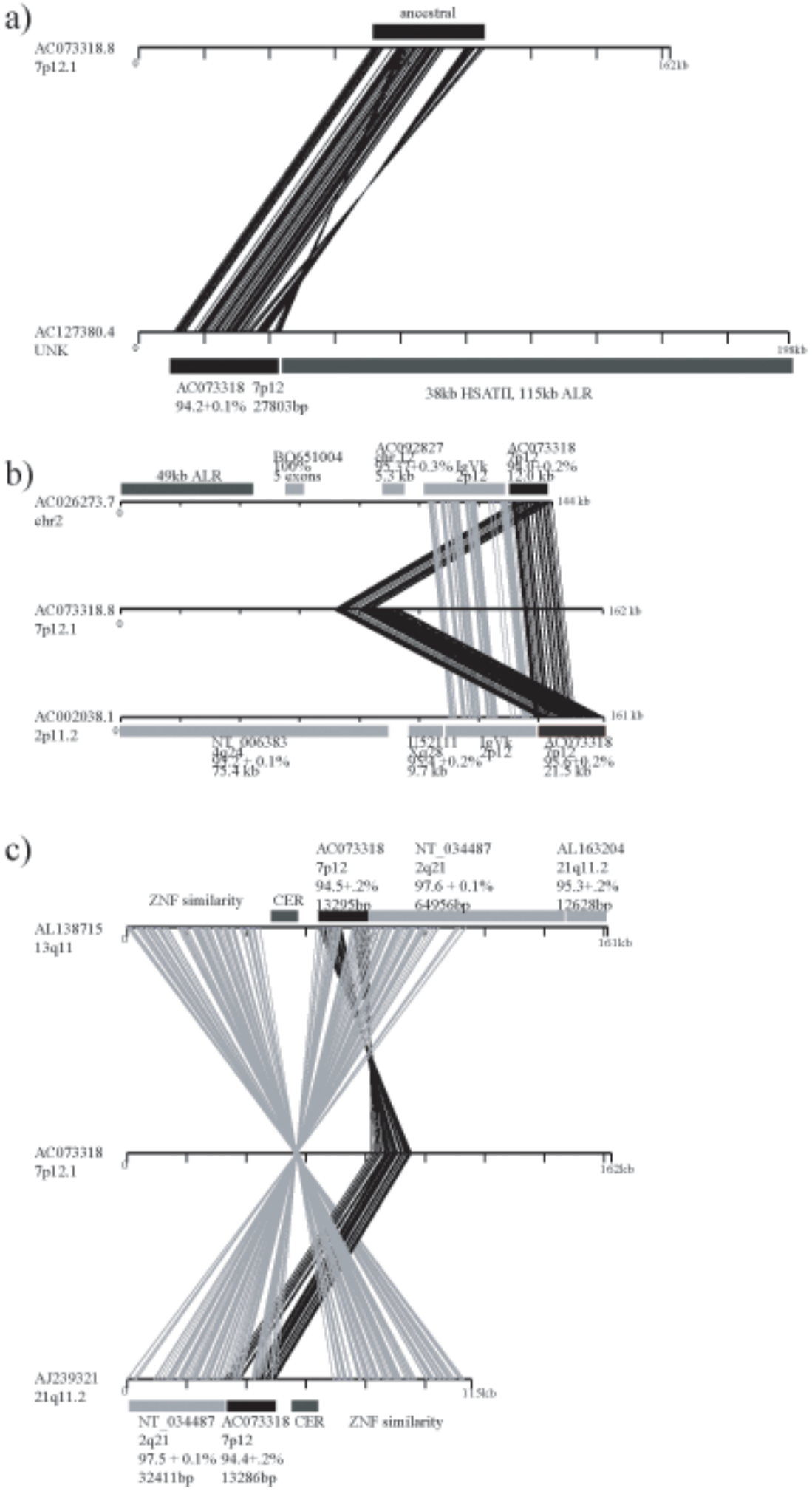
cosmids from the chromosome 2 LLNL02 library (92M18 and 67N8) resolves these variant sites (indicated by letters in red).

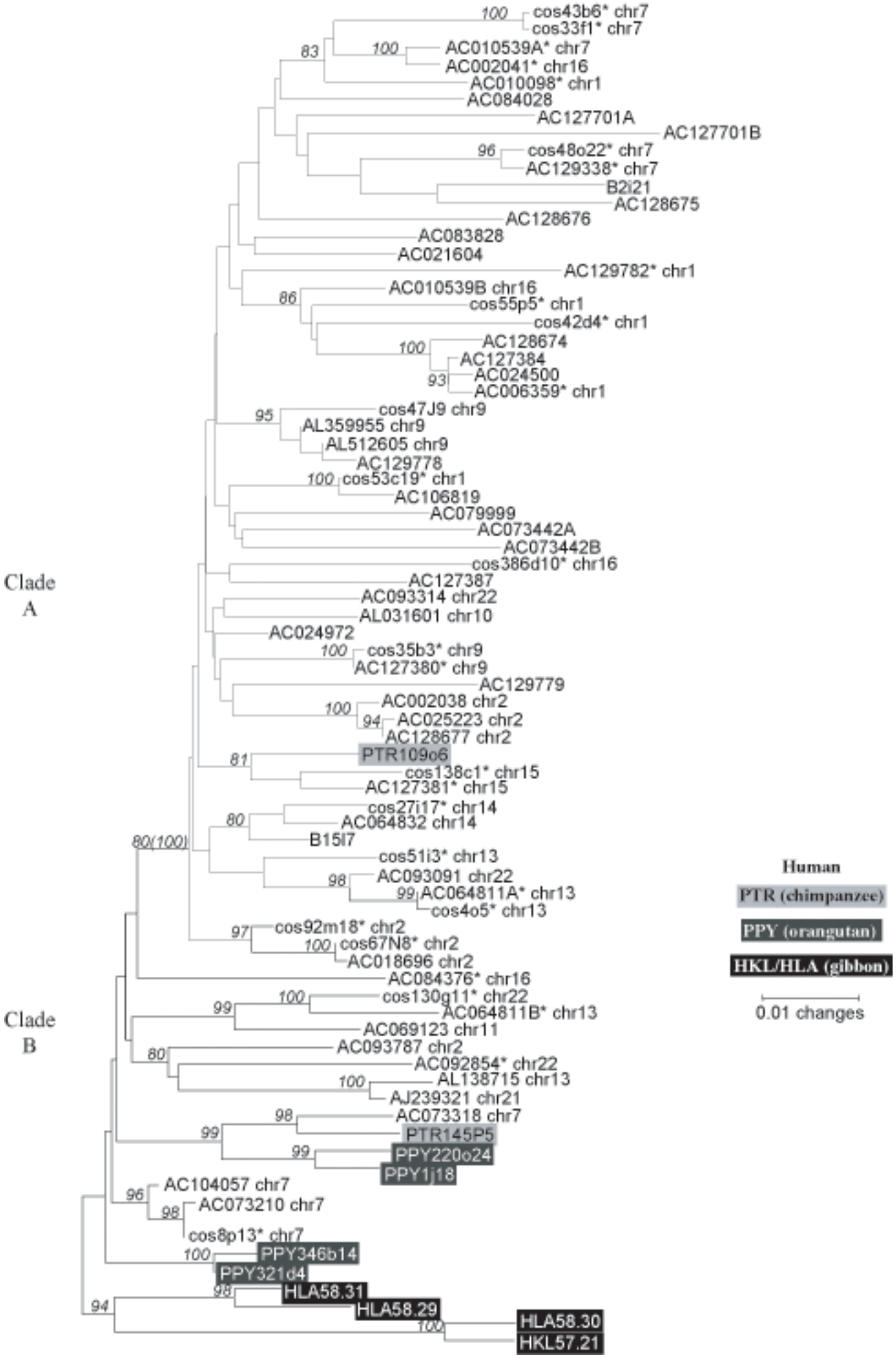
Figure 6: Genomic structure of PIR4 sequences.

The genomic structure of 47 human PIR4 loci with respect to the putative ancestral locus (AC073318) is depicted using the program PARASIGHT (Bailey, unpublished). Briefly, 60 kb of the ancestral copy of PIR4 (AC073318) is represented by the top horizontal blue line. Regions with sequence similarity (>90% identity) to other PIR4-containing segments are highlighted in red. Regions with significant sequence homology are shown below using colored boxes to indicate sequence identity to AC073318 (see legend) Some accessions (AC093787) share considerable sequence in common (>40 kb for AC073318) while others (AL138715) share much less (13KB with AC073318). The vertical black line through the middle indicates the approximate position of the ~1kb tree (rooted on gibbon) which is provided on the left side of the schematic for comparison.

Figure 7: Genetic distances between PIR4 sequences

The number of substitutions per site (K) among PIR4 elements as a function of the number of pairwise alignments. Based on available finished sequence, 25 PIR4 sequences were extracted. A total of 254 pairwise alignments were performed and the genetic distance for each pairwise was calculated. Each alignment was at least 10 kb in length. Distance estimates were computed using the Kimura two-parameter model.


a)
ancestral
AC073318.8
7p12.1
162kb
AC127380.4
UNK
198kb
AC073318 7p12
94.2+0.1% 27803bp
38kb HSATII, 115kb ALR
b)
49kb ALR
BQ651004
100%
5 exons
AC092827
chr 17
95.3+0.3% IgVk
5.3 kb
2p12
AC073318
7p12
94.3+0.2%
12.0 kb
AC026273.7
chr2
144 kb
AC073318.8
7p12.1
162 kb
AC002038.1
2p11.2
161 kb
NT_006383
4q24
94.2 + 0.1%
75.4 kb
U52111
Xq28
95.4+0.2%
9.7 kb
IgVk
2p12
AC073318
7p12
95.6+0.2%
21.5 kb
c)
AC073318
7p12
94.5+.2%
13295bp
NT_034487
2q21
97.6 + 0.1%
64956bp
AL163204
21q11.2
95.3+.2%
12628bp
ZNF similarity
CER
AL138715
13q11
163kb
AC073318
7p12.1
162kb
AJ239321
21q11.2
115kb
NT_034487
2q21
97.5 + 0.1%
32411bp
AC073318
7p12
94.4+.2%
13286bp
CER
ZNF similarity


100
cos43b6* chr7
cos33f1* chr7
83
100
AC010539A* chr7
AC002041* chr16
AC010098* chr1
AC084028
AC127701A
AC127701B
96
cos48o22* chr7
AC129338* chr7
B2i21
AC128675
AC128676
AC083828
AC021604
AC129782* chr1
AC010539B chr16
86
cos55p5* chr1
cos42d4* chr1
100
AC128674
AC127384
93
AC024500
AC006359* chr1
95
cos47J9 chr9
AL359955 chr9
AL512605 chr9
AC129778
100
cos53c19* chr1
AC106819
AC079999
AC073442A
AC073442B
cos386d10* chr16
AC127387
AC093314 chr22
AL031601 chr10
AC024972
100
cos35b3* chr9
AC127380* chr9
AC129779
100
AC002038 chr2
94
AC025223 chr2
AC128677 chr2
81
PTR109o6
cos138c1* chr15
AC127381* chr15
80
cos27i17* chr14
AC064832 chr14
B15I7
80(100)
cos51i3* chr13
98
AC093091 chr22
99
AC064811A* chr13
cos4o5* chr13
97
cos92m18* chr2
100
cos67N8* chr2
AC018696 chr2
AC084376* chr16
100
cos130g11* chr22
99
AC064811B* chr13
AC069123 chr11
AC093787 chr2
80
AC092854* chr22
100
AL138715 chr13
AJ239321 chr21
98
AC073318 chr7
99
PTR145P5
99
PPY220o24
PPY1j18
96
AC104057 chr7
98
AC073210 chr7
cos8p13* chr7
100
PPY346b14
PPY321d4
98
HLA58.31
HLA58.29
94
100
HLA58.30
HKL57.21
Clade A
Clade B
Human
PTR (chimpanzee)
PPY (orangutan)
HKL/HLA (gibbon)
0.01 changes

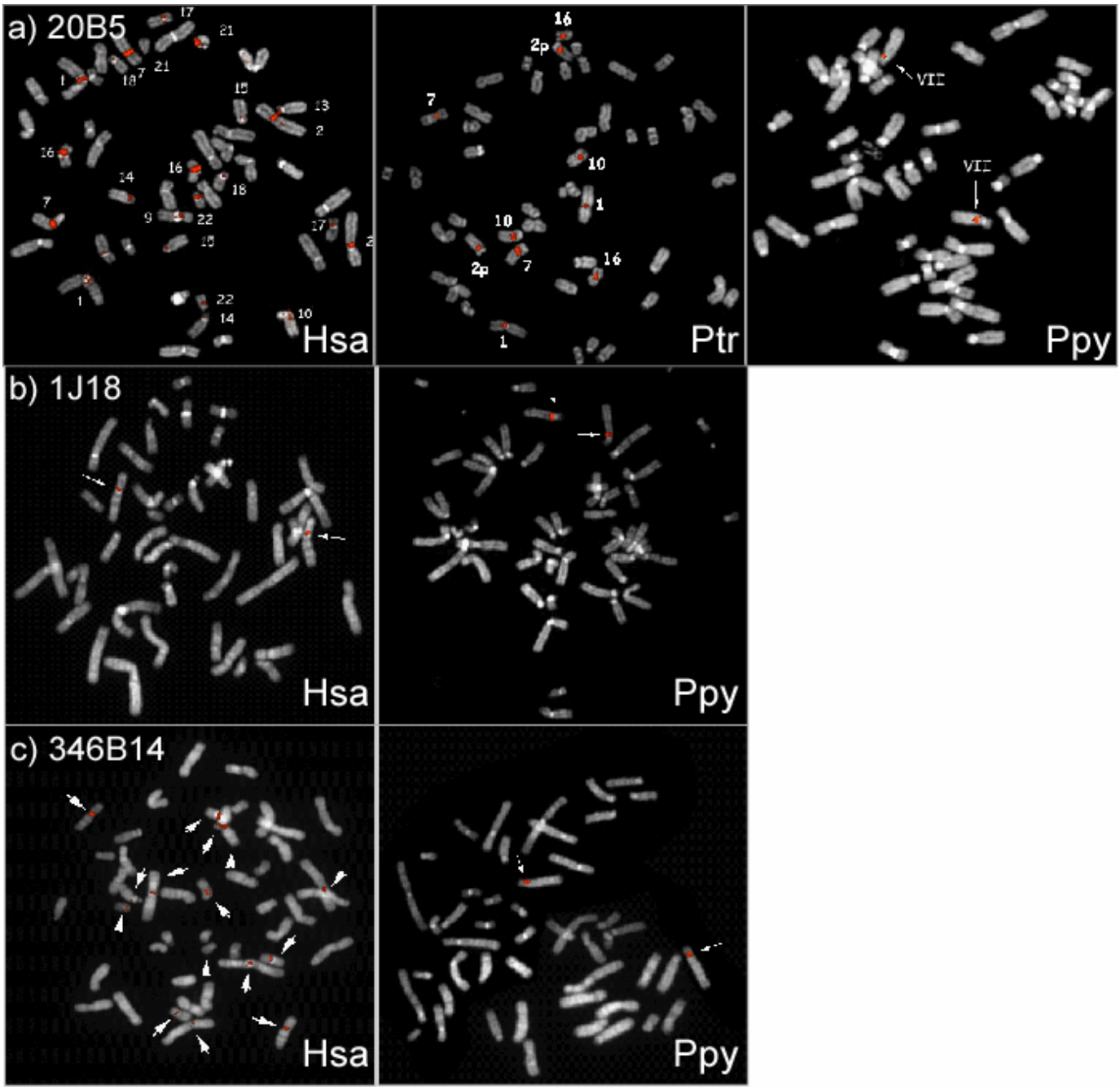


Figure 3: Primate metaphase FISH of PIR4 containing clones.

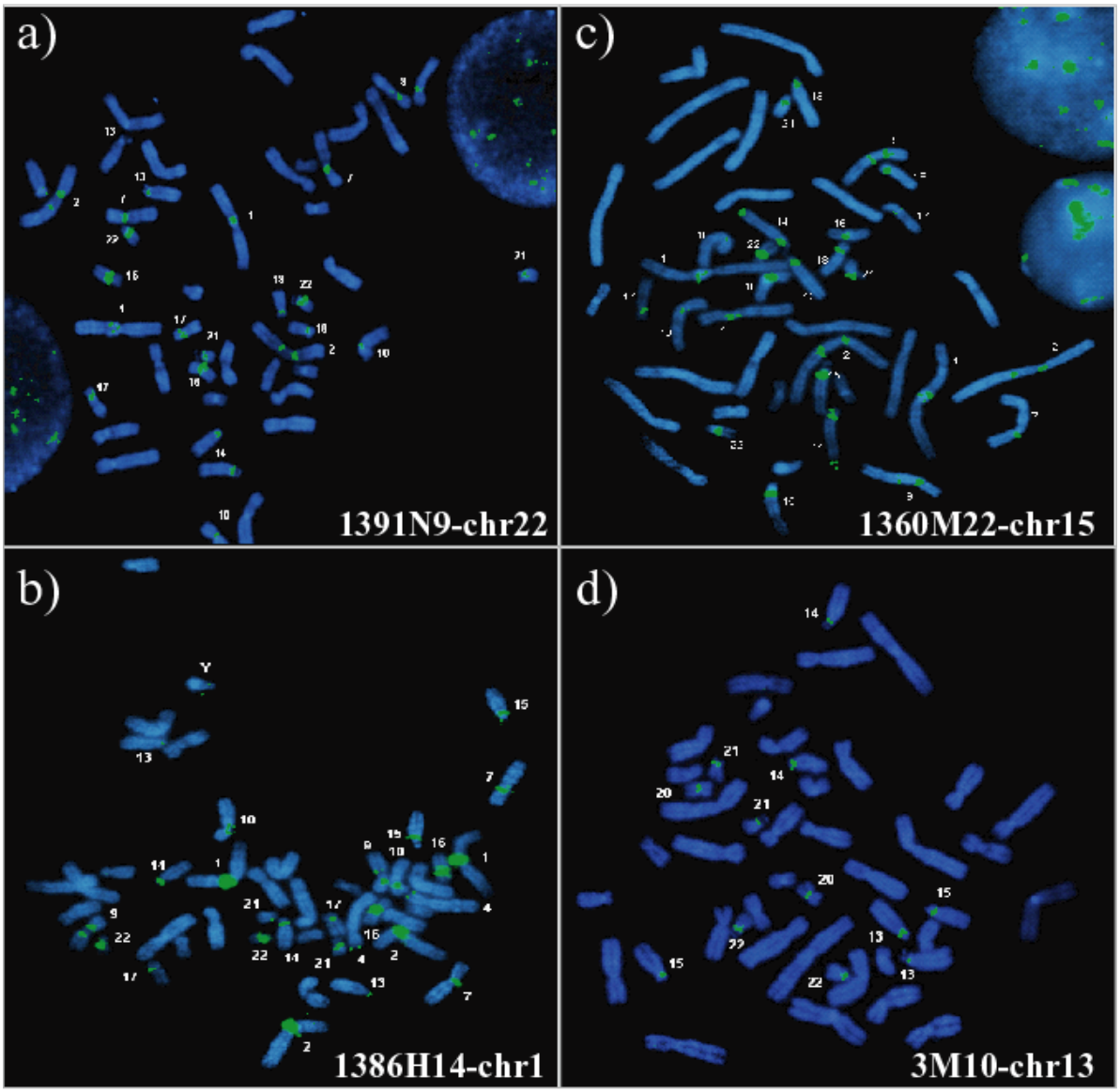
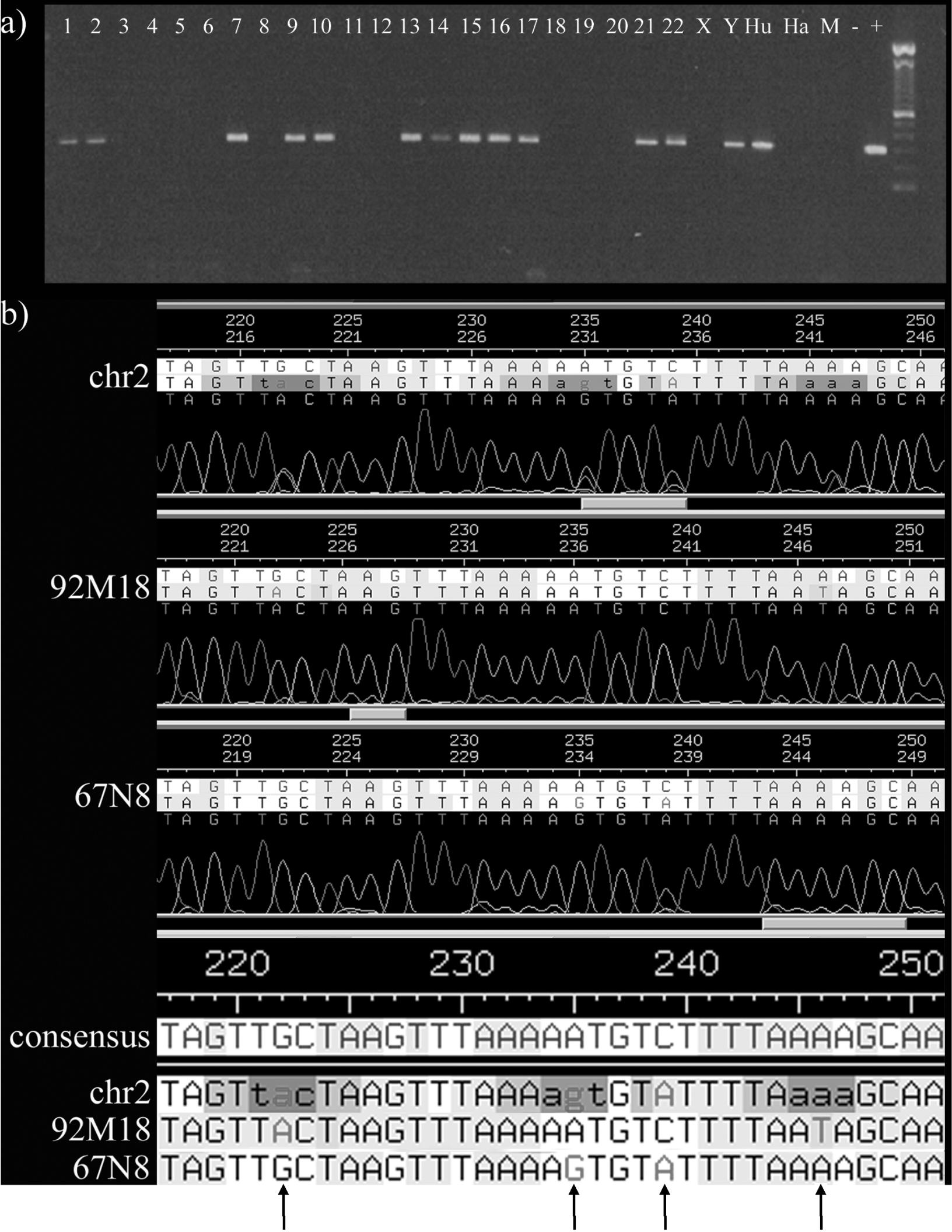
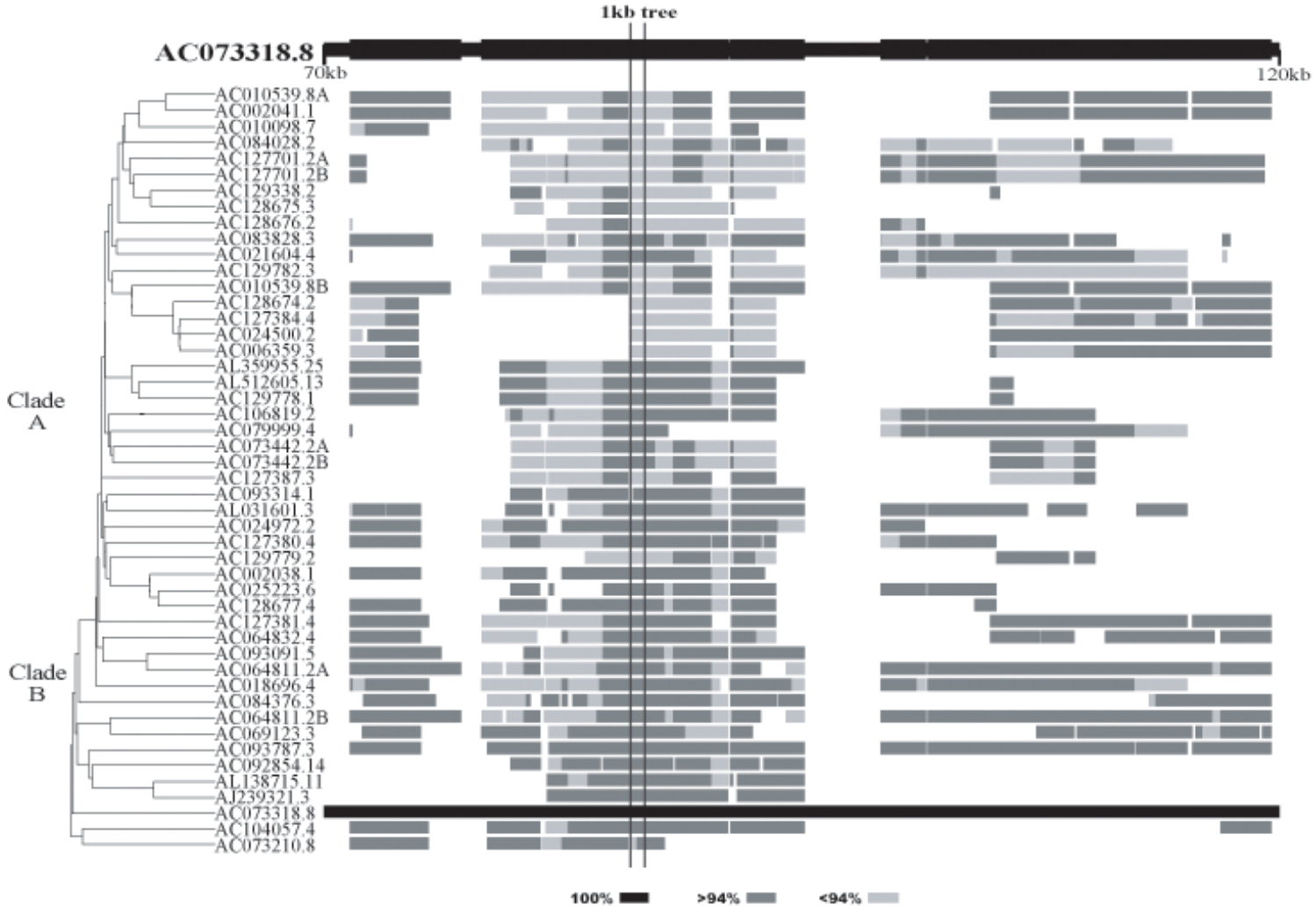


Figure 4: Examples of metaphase FISH from PIR4-containing BACs.


a)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y Hu Ha M - +
b)
chr2
92M18
67N8
consensus
chr2 TAGTtacTAAGTTTAAAAagtGTATTTTAaaaGCAA
92M18 TAGTTACTAAGTTTAAAAATGTCTTTTAATAGCAA
67N8 TAGTTGCTAAGTTTAAAAGTGTATTTTAAAAGCAA


1kb tree
AC073318.8
70kb
120kb
AC010539.8A
AC002041.1
AC010098.7
AC084028.2
AC127701.2A
AC127701.2B
AC129338.2
AC128675.3
AC128676.2
AC083828.3
AC021604.4
AC129782.3
AC010539.8B
AC128674.2
AC127384.4
AC024500.2
AC006359.3
AL359955.25
AL512605.13
AC129778.1
AC106819.2
AC079999.4
AC073442.2A
AC073442.2B
AC127387.3
AC093314.1
AL031601.3
AC024972.2
AC127380.4
AC129779.2
AC002038.1
AC025223.6
AC128677.4
AC127381.4
AC064832.4
AC093091.5
AC064811.2A
AC018696.4
AC084376.3
AC064811.2B
AC069123.3
AC093787.3
AC092854.14
AL138715.11
AJ239321.3
AC073318.8
AC104057.4
AC073210.8
Clade A
Clade B
100%
>94%
<94%

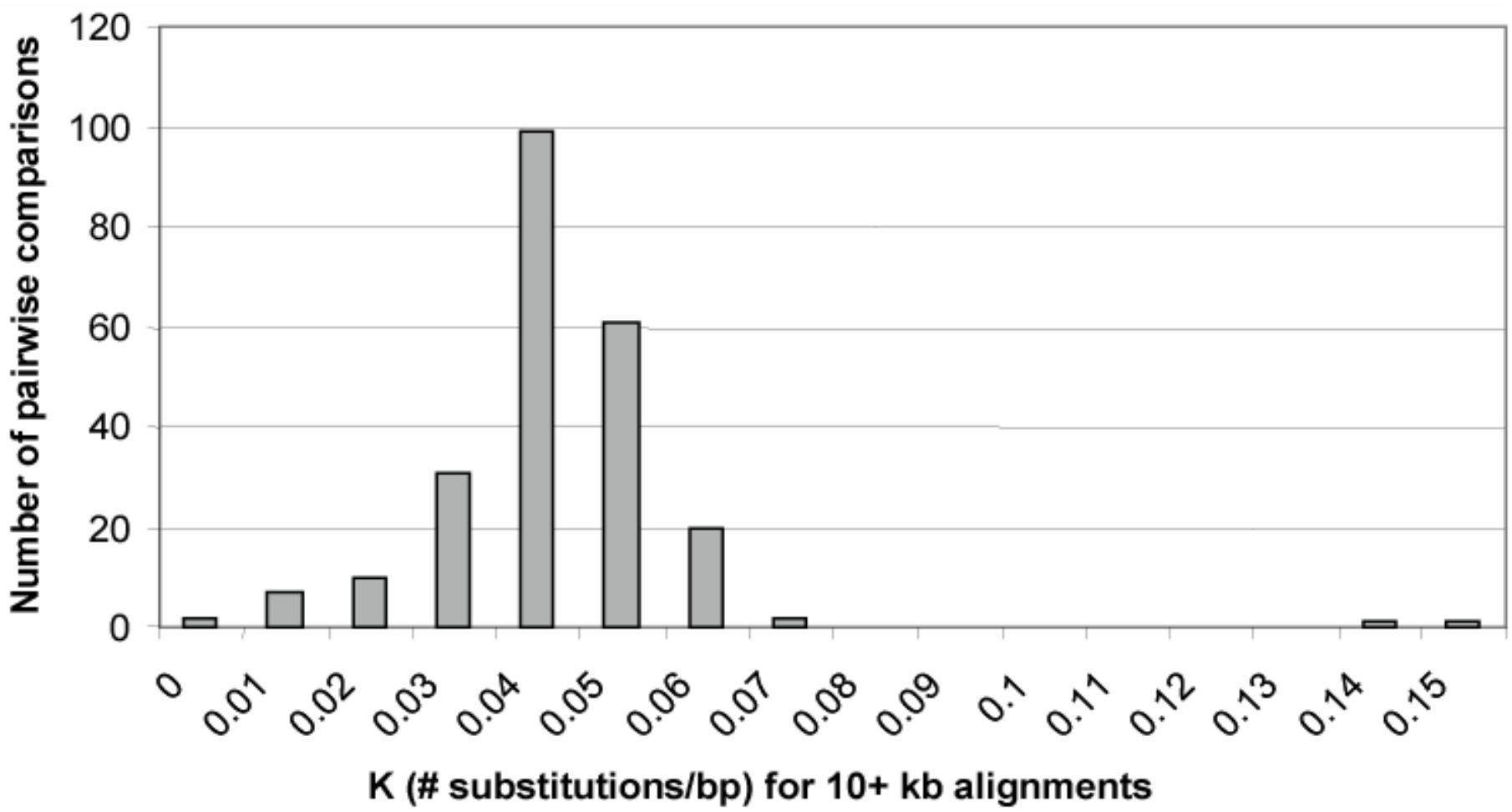


Figure 7: Genetic distances between PIR4 sequences.