Phylogenetic Screening of the Human Genome: Identification of Differentially Hybridizing Repetitive Sequence Families

Joyce A. Lloyd,* Allen N. Lamb,† and S. Steven Potter*

*Children's Hospital Research Foundation; and †Department of Molecular Biology and Biochemistry, Wesleyan University

The ϕ-screen, a method of phylogenetic screening, can be employed to detect repetitive sequence families that differentially hybridize between closely related species. Such differences may involve sequence divergence or variations in copy number, including total presence versus absence of a family of repeated DNA. We present the results of a ϕ-screen comparing the human genome to that of the prosimian, *Galago crassicaudatus*. Three human repetitive families that are divergent or not present in galago have been detected. One of these families is described in detail; it is similar among the anthropoids but is present in a lower copy number and/or divergent form in prosimians. The family is clearly related to the transposon-like human element (THE) described by Paulson et al. (1985). THEs have long terminal repeats reminiscent of retroviruses but are unique in that they have no sequence similarity to known mammalian retroviruses. The sequence of a solo long terminal repeat, found unassociated with THE internal sequence, is presented. This family member, THE p2, is bordered by a 5-bp target-site repeat and is interrupted by the insertion of an Alu element. A solo THE element sequenced by Wiginton et al. (1986) contains an insertion of Alu at precisely the same position as does THE p2.

Introduction

At least 15% of most eukaryotic genomes is composed of interspersed repetitive DNA sequences (Jelinek and Schmid 1982), the function of which is largely unknown. Many of these repetitive sequences are transposable elements, with the ability to preferentially replicate and insert into new positions in the genome, thus increasing their copy number. Transposable elements can generate DNA rearrangements, including (1) placing new promoters in front of genes that they insert next to (Roeder and Fink 1983), (2) causing inversions and deletions of flanking sequence owing to recombination between elements (Roeder and Fink 1983), (3) cooperating to levitate large blocks of DNA to new positions in the genome (Paro et al. 1983), and (4) producing mutations by inserting into genes (Collins and Rubin 1982; Bender et al. 1983; Kuff et al. 1983; Doring and Starlinger 1984). Transposable elements, therefore, generate genetic variation for natural selection to act on. Indeed, transposable elements may provide a prime force in molding genomes during evolution (Campbell 1983; Syvanen 1984; Temin 1985).

1. Key words: phylogenetic screening method, evolution of THE family, primate transposon, Alu element, DNA sequence, repeated sequence.

Address for correspondence and reprints: Dr. S. Steven Potter, Children's Hospital Research Foundation, IDR Building, Fourth Floor, Elland and Bethesda Avenues, Cincinnati, Ohio 45229.

© 1987 by The University of Chicago. All rights reserved.
0737-4038/87/0402-4201$02.00

85
While the fact that transposable elements can cause such DNA rearrangements makes them generally interesting, the study of human transposable elements is particularly important, for practical reasons. Transposable elements can be polymorphic in chromosomal position and therefore might facilitate the rapid isolation of restriction-fragment-length polymorphisms (RFLPs) useful in diagnosing genetic disease. Transposable elements probably also play a role in causing some human diseases, owing to their ability to generate mutations. Furthermore, human transposable elements may prove useful as gene transfer vectors, à la the P-elements in *Drosophila* (Rubin and Spradling 1982).

One way of screening for transposable elements in eukaryotic genomes has been to characterize randomly cloned repetitive sequences. Although useful for better understanding the basic structure of a genome, this method does not specifically select for transposable elements that are currently active. Another successful approach has been to look for repetitive elements in alphoid satellite DNA (Grimaldi and Singer 1983; Potter and Jones 1983; Potter 1984), assuming that their presence in alphoid satellite indicates that their insertion is recent, since alphoid satellite tends to “correct” itself by unequal crossover (Smith 1973, 1976) and/or gene conversion.

Yet another method is the recently described phylogenetic DNA-screening procedure, the Φ-screen, which compares the repetitive components of two related species and identifies differentially hybridizing sequence families that might represent active, recently acquired transposable elements (Wichman et al. 1985). We define differentially hybridizing sequence families as those that (1) have diverged sufficiently so that they do not cross-hybridize between species under our conditions or (2) vary greatly in copy number between species, even to the extent of being present in one species and totally absent in another. The rationale for the Φ-screen procedure is based on the observation that P-elements have apparently invaded the genome of *D. melanogaster* very recently in evolutionary time; they are not found in old laboratory stocks of *D. melanogaster* (Kidwell 1979) and are apparently absent in some closely related *Drosophila* species (Lansman et al. 1985). Our laboratory has previously employed the Φ-screen technique to compare the white-footed mouse (*Peromyscus leucopus*) and the house mouse (*Mus domesticus*) genomes, finding a white-footed mouse retrotransposon family absent in the house mouse genome (Wichman et al. 1985).

In this report we describe the details of the Φ-screen protocol (see Results below) and its application to the screening of 500 human clones, representing ~5,000 kb or 1/600 of the human genome. We have identified three categories of differentially hybridizing sequences in humans by comparing human DNA to that of the prosimian *Galago crassicaudatus*. A detailed characterization of a human transposon-like family is presented. This family either is not amplified to as great a degree in galago and other prosimians or is in an apparently divergent form. We present the sequence of a human family member into which a second transposable element, an Alu element, has transposed. We discuss some insights into a basic understanding of the organization of the human genome gained from the use of the Φ-screen and also the application of the Φ-screen technique to evolutionary and molecular problems.

**Materials and Methods**

**Materials**

Enzymes were purchased from Bethesda Research Laboratories (Klenow), N.E. Biolabs (restriction enzymes) and Life Sciences, Inc. (reverse transcriptase). Isotope
was from New England Nuclear. Agarose was purchased from Sigma, acrylamide and bis from BioRad, nitrocellulose from Schleicher and Schuell, and GeneScreenPlus from New England Nuclear.

The $\phi$-Screen

The construction of a human library from HeLa cell DNA in pBR322 has been previously described (Potter and Jones 1983). DNA minipreparations of randomly chosen individual clones were made using the rapid-boil method of Holmes and Quigley (1981). HeLa clone DNAs were digested with three 6-base recognition-site restriction enzymes (BamHI, EcoRI, and HindIII) and run on 1% agarose gels. The gels were blotted to nitrocellulose using the method of Southern (1975) with the sandwich-blot variation (Maniatis et al. 1982). Hybridizations of the Southern blots were done at 55 C in 4 X SSCP (1 X SSCP = 20 mM sodium phosphate, pH 6.8, 120 mM NaCl, 15 mM NaCitrate) and 1 X Denhardt's solution (Denhardt 1966); filters were presoaked and washed in the same solution at 55 C. The filters were exposed to Kodak XAR-5 film at −70 C using DuPont Cronex Lightning Plus intensifying screens.

DNA Sequencing

The DNA sequencing of the p2 element was performed using both the Maxam and Gilbert (1980) partial chemical cleavage method (G, C, C + T, A > C reactions) and the Sanger enzymatic dideoxy-chain termination method (using either DNA PolI large fragment or reverse transcriptase) (Sanger et al. 1977).

Dot-Blot Hybridizations

The primate DNAs were spotted on a GeneScreenPlus membrane. Dot blots were presoaked and hybridized in 1 M NaCl, 5 X Denhardt's, 100 µg salmon sperm DNA/ml, 0.5% sodium dodecyl sulfate (SDS) at 50 C. They were washed in 2 X SSC (1 X SSC = 150 mM NaCl, 15 mM sodium citrate) and 1.0% SDS for 5 min at room temperature and then in 2 X SSC and 1.0% SDS for 2 h, with three to four changes, at 50 C.

Sequence analysis was performed using the SEQ program developed by members of the MOLGEN group at Stanford University.

Results

The $\phi$-screen procedure was used to isolate human differentially hybridizing repetitive DNA sequences. Five hundred HeLa clones, encompassing ~5,000 kb of human DNA, were randomly selected from a pBR322 library. Repetitive sequences present in these human clones, and absent or less evident in Galago crassicaudatus DNA, were identified in the following manner. DNA minipreparations of the randomly selected clones were digested with BamHI, HindIII, and EcoRI restriction enzymes and electrophoresed on 1% agarose gels. The gels were sandwich blotted to obtain duplicate Southern blots. One blot was then hybridized to nick-translated total genomic HeLa DNA, the other to labeled galago DNA. Repetitive sequences in the cloned DNAs appeared as bands on autoradiograms, with the intensity of the band approximately reflecting the copy number of the sequence in the probe DNA, i.e., in the genome. From previous work with mouse clones in this laboratory, the technique is known to easily detect sequences repeated ≥1,000 times per haploid genome (Wichman et al. 1985). Comparison of duplicate autoradiograms reveals that most bands that
appear when HeLa clones are hybridized to $^{32}$P–HeLa DNA probe are also evident when hybridized to galago probe; but a few are not. These differentially hybridizing bands represent sequences that are either very rare in one of the two primates or divergent between the two primates (fig. 1). Typically, the radioactivity on the blots was then washed off with a 0.5 M NaOH solution and they were hybridized with the opposite probe to guard against artifacts.

Of 500 clones checked with HeLa versus galago probes, 20 with differences in hybridization were found. Clones containing these differentially hybridizing bands were then subjected to hybridization with probes for previously characterized families such as Alu, alphoid satellite, and L1 (KpnI). Three categories of differentially hybridizing bands were noted; those hybridizing to 5' regions of L1 (13 clones), those hybridizing to alphoid satellite (1 clone), and those not hybridizing to any of these probes (6 clones).

To better understand the repeated sequence content of the human genome, we hybridized a collection of the Southern filters including digests of 300 clone DNAs to probes for two known repeat sequence families. Ninety percent (939 of 1,049) of the total number of bands with repetitive DNA proved to contain Alu elements, and of the remaining segments with repeats 87% were L1. These results indicate that Alu and L1 repeats constitute a great majority of the sequences repetitive enough to be detected by this procedure.

A Human Transposon-like Sequence

p53 is a clone from the pBR322 HeLa library containing a fragment that shows differential hybridization to total HeLa and galago probes but does not hybridize to an L1 probe (data not shown). Restriction mapping of this clone (fig. 2) revealed that the differentially hybridizing region could be localized to the edge of the insert. Preliminary sequence data (not shown) demonstrated that the region did not contain Alu-family or alphoid-satellite homology. The position of the repetitive element at the edge of the clone suggested that this might not be a complete copy but perhaps was truncated by the cloning process. A probe from p53 (see fig. 2A) was therefore used to isolate other hybridizing clones from the pBR322 HeLa library.

One such clone was p2 (see restriction map in fig. 2A). The p53 probe hybridized to a 925-bp PvuII-BamHI fragment of p2, which in turn overlapped a 1.8-kb EcoRI fragment of p2 showing differential hybridization. We also isolated four λ clones that possessed multiple restriction fragments, within individual clones, hybridizing to the p2 1.8-kb EcoRI fragment (data not shown). This internal repetition of sequence within clones suggested to us that all of the clones might contain the terminal repeats of a transposable element. Sequencing of a portion of the p2 clone proceeded by the strategy shown in figure 2B. Partial sequence of p53 and one of the λ clones defined the approximate ends of the p2 element (not shown). It is called THE (for transposon-like human element) p2 because sequence comparisons between it and THE 1-A (a member of the THE family described by Paulson et al. [1985]) permit the inference that they are homologous (see fig. 4). The entire THE 1-A transposon is 2.3 kb in length, including two flanking long terminal repeats (LTRs) of ~350 bp each; it has no sequence similarity to known mammalian retroviruses. Paulson et al. report that the THE LTRs are present in ~30,000 copies per haploid genome, whereas the internal region of the transposon is found only 10,000 times. This predicts a copy number of 10,000 for solo LTRs, i.e., single THE LTRs not associated with internal THE sequence.
FIG. 1.—Phylogenetic screen of the human genome. Duplicate Southern blots are shown of nine randomly selected human cloned DNAs, each digested with EcoRI, HindIII, and BamHI and then hybridized to total genomic DNA probes from galago (sections A and D) and human (sections B and C). The arrows in A and B point to bands in lane 3 that are present with the human probe and absent with the galago probe. Subsequent studies with known probes established that the top band hybridizes to the 5' end of L1 elements and that the bottom band hybridizes to sequences of the THE family. Lane 3 in section C, missing in section D, also represents differential hybridization of the 5' end of the L1 repeat.
Indeed, two such solo LTRs, O4 and O5, were sequenced by Sun et al. (1984). The p2 clone appears to contain a solo THE LTR into which an Alu element has transposed. The sequence of the p2 LTR and inserted Alu is shown in figure 3; THE p2 is surrounded by a 5-bp direct repeat in target DNA (ccttg). The ~150-bp flanking THE p2’s on either side do not show any significant similarity to the internal sequence of THE 1-A. As revealed by the sequence comparison in figure 4, p2 shows 69% identity to the THE 1-A left LTR, 71% identity to the THE 1-A right LTR, 73% identity to O4, and 74% identity to O5. Independently, in sequencing the human adenosine deaminase (ADA) gene, Wiginton et al. (1986) identified a THE element that contains...
FIG. 3.—The sequence of THE p2, the inserted Alu element, and flanking DNA. The THE p2 and Alu sequence are uppercase and the flanking sequence is lowercase. The 5-base target-site repeat formed by the insertion of THE p2 is underlined (ccttg), as is the 7-base repeat formed by the insertion of the Alu element; i.e., all the sequence between the two TCAGGTA repeats is Alu sequence. Note that the strand of Alu sequence is the opposite of the one normally shown, in order to present the same strand of THE p2 as is shown for THE 1-A by Paulson et al. (1985). The 5' end of the Alu, as it is usually represented, is at position 782.
an insertion of Alu in precisely the same position as in THE p2. The sequence of the ADA THE is also shown in figure 4; it is 77% identical to THE p2. A probe from the region flanking THE p2 was used to check for RFLPs in five different human DNAs treated with 17 different restriction enzymes. No evidence of RFLPs involving either the THE or the Alu element was revealed (data not shown).

---

FIG. 4.—Sequence comparison of THE p2 to other sequenced THEs. The p2 is THE p2, ADA is the THE found in sequencing the human ADA gene (Wiginton et al. 1986), 1-A L and 1-A R are the left and right terminal repeats, respectively, of THE 1-A (Paulson et al. 1985), and 04 and 05 are the solo terminal repeats sequenced by Sun et al. (1984). The entire sequence of THE p2 is shown. Positions where the base present in ADA, 1-A L, 1-A R, 04, or 05 match p2 are left blank; at positions where a different base is present, that base is identified. A dash represents a gap introduced to improve the alignment. The Alu target sites in THE p2 and THE ADA are underlined, but there is no Alu sequence in this figure. The ADA Alu target-site repeat is imperfect; the third residue is an A in the left-hand repeat and a G in the right-hand repeat. The region of THE p2 and THE ADA that shows little similarity to any of the other THEs is between positions 206 and 226 of p2.

---
The Alu sequence has inserted into THE p2 in a position 42 bp from the right end of the p2 LTR. It is 90% identical to an Alu consensus sequence presented by Schmid and Shen (1985) and 82% identical to the ADA Alu. The ADA and p2 Alu elements are surrounded by 7-bp target-site duplications (fig. 4). Since the p2 and ADA Alu elements are inserted into exactly the same position in respective THEs, their target-site duplications are closely related. The p2 Alu is surrounded by perfect direct repeats (TCAGGTA). One of the ADA Alu repeats matches this sequence; the other differs in that the third base is replaced by a G.

In general, there is no strong consensus target-site sequence into which Alu elements insert. However, Daniels and Deininger (1985) have shown that, in humans, the direct repeats surrounding Alu elements tend to be A-rich at the 5′ ends. Also, the 10 bp of DNA flanking these integration sites to the 5′ side of the Alu tend to be A-T rich (72% on average). Though the p2/ADA Alu target site is not A-rich, the 10 bp of THE sequence 5′ to these duplications is, respectively, 80%/60% A-T.

We were interested in determining the relative abundance of THE-related sequences in various primates. Equal quantities of a number of primate DNAs were spotted on a nylon membrane and hybridized to another sequenced THE probe (THE S1) containing the left half of a THE LTR and 80 bp of flanking sequence (fig. 5). The flanking sequence was compared to those in the GenBank library and showed no similarity to sequences therein; this ensures that other identified repetitive families such as L1 and Alu are not contained on the probe and will not affect hybridization results. Since hybridization of divergent copies in distant species might be difficult, the blots were washed at low stringency (2 × SSC at 50 C). Interestingly, hybridization

---

**FIG. 5.**—Dot-blot hybridization of THE S1 to various primate genomic DNAs. All primate dots are in duplicate and are 0.5 μg of DNA. The blot was hybridized in 1 M NaCl at 50 C and washed in 2 × SSC at 50 C. The primate abbreviations are as follows: HL = HeLa DNA; Hs = Homo sapiens; Pt = Pan troglodytes (chimpanzee); Gg = Gorilla gorilla; Pp = Pongo pygmaeus (orangutan); Hl = Hylobates lar (white-handed gibbon); Ss = Saimiri sciurea (squirrel monkey); Tb = Tarsius bancanus (Harsfield’s tarsier); Lc = Lemur catta (ring-tailed lemur); and Gga = Galago garnetti (bush baby). Tg = Tupaiia glis (tree shrew). High-molecular-weight HeLa DNA and G. garnetti DNA were prepared from cultured cells. All other genomic DNAs are sonicated. Dots were made using a BRL dot-blot apparatus.
to THE S1 appears strong in all the monkeys and apes examined. There is a distinct
drop in hybridization to prosimian (tarsier, lemur, and galago) and tree shrew DNA,
although galago DNA showed hybridization even under quite stringent conditions (57
C, 0.1 × SSC) (not shown). On long exposure, even tree shrew shows some hybridization
to THE S1 (not shown), but the significance of this is unclear. The evidence supports
the view that the THE family is present in prosimians—but in a form quite divergent
from and/or in a copy number lower than that in humans.

Discussion

The φ-screen procedure can be used to identify differentially hybridizing repetitive
sequences in any desired species. To do a φ-screen one must have a library of genomic
DNA from the species of interest, as well as total DNA isolated from a few related
species of increasing evolutionary distance. Some trial and error may be necessary to
determine the evolutionary distance between species that is necessary to detect differ-
ences, but this is a minor problem because the original blots can be reused with ad-
ditional probes. A consideration of the rates of nucleotide substitution in various
taxonomic groups could influence such decisions. We were not able to find any dif-
ferences between human and African green monkey in looking at 1/3,000 of the
genome, whereas Wichman et al. (1985) were able to easily detect differentially hy-
bridizing sequences in closely related rodents, in which the substitution rate is reported
to be higher than that in primates (Goodman 1985; Wu and Li 1985; Britten 1986).

In our screen of 500 human clones, compared to Galago crassicaudatus (bush-
baby) repetitive sequences, 20 differentially hybridizing clones were detected, repre-
senting three different categories: 5' regions of L1 sequences (which may not include
the putative coding region conserved between mouse and man [Martin et al. 1984;
Singer and Skowronski 1985]), alphoid satellite-like sequences, and THE-like se-
quences. These 500 10-kb clones represent ~5,000 kb of human DNA or 1/600 of the
haploid human genome. Assuming that our hybridization sensitivity allows us to detect
elements present ≥1,000 times per haploid genome (Wichman et al. 1985), we likely
would have identified any differentially hybridizing sequence of that or greater copy
number.

The p2 element sequenced here appears to be a 364-bp solo LTR of the THE
family recently described (Paulson et al. 1985), though one containing an inserted
Alu repeat. It is of particular interest that the position of the inserted Alu element in
THE p2 is precisely the same as that for a THE-Alu composite element sequenced in
a study of the human ADA gene (Wiginton et al. 1986). This poses the question of
whether (1) Alu elements transposed into THE elements on two separate occasions
or (2) a single initial composite element proliferated. The likelihood that two such
exact transposition events should occur is exceedingly small, unless this particular
position is a preferred site. As discussed above in Results, there is some evidence for
target-site preference for Alu elements in that A-T-rich regions are often found 5' to
the inserted Alu (Daniels and Deininger 1985). However, although the region 5' to
the Alu in THE p2 is A-T rich, there are several other equally or more A-T-rich regions
also present in THE p2.

The alternative explanation—that there was a single Alu insertion event—is sup-
ported by the following evidence. First, the ADA THE and the p2 THE are 77%
identical, showing more overall similarity to each other than to any other sequenced
THE. Furthermore, between THE p2 positions 206 and 226, THE p2 and THE ADA are 81% identical, whereas THE p2 shows only 33% identity and THE ADA 39% identity to any of the other THEs. Similarly, between THE p2 positions 308 and 315, both THE p2 and THE ADA contain an insertion of 8 bp not found in any of the other THEs. Also, although the staggered cuts made when Alu elements insert can be separated by various numbers of bases, thereby yielding target-site duplications of different lengths, both THE p2 and THE ADA contain target-site duplications that are the same size, as well as in the same position. It is possible that a significant fraction of THE LTRs in the human genome contain Alu insertions, since two of the six THE LTRs sequenced so far are composite elements.

One might speculate that the presence of the foreign Alu elements inserted into THE p2 and THE ADA might inhibit gene conversion events that would otherwise serve to make individual THE elements more similar to each other (Schimenti and Duncan 1984). Indeed, THE p2 is only 69%–74% identical to the other THEs that have been sequenced and THE ADA is 66%–70% identical, whereas the average similarity of all the possible pairs of the other sequenced THEs is much higher (83%). In fact, of these four THE LTRs—O4, O5, THE 1-A right, and THE 1-A left—the LTR that is least similar to the others is THE 1-A left (78%, 82%, and 82% identical, respectively). The percent identity for other combinations is higher: O4 to O5, 88%; O4 to THE 1-A right, 84%; and THE 1-A right to O5, 85%. This is interesting because THE 1-A left contains, as does THE p2 and THE ADA, a sizable insertion of DNA foreign to THEs (22 bp at positions 159–180); the largest insertion in any of the other three LTRs is 6 bp, in O5. This evidence suggests that gene conversion plays a role in homogenizing THEs and that insertions have a negative effect on this process. However, one must consider that Sawada et al. (1985) have shown that gene conversion events have been rare in seven orthologous human and chimpanzee Alu elements, at least since the divergence of the two primates.

Dot-blot hybridization using a THE S1 probe shows that THE is indeed present in other primates, although prosimian (lemur, tarsier, and galago) DNAs appear to contain a lower copy number and/or more divergent form of THE than the monkeys and apes. These data are consistent with a homogenization of the anthropoid THE family that could be due to concerted evolution.

Other repetitive sequences, such as Alu elements, did not show the differential hybridization to HeLa versus galago probes that THE elements did. However, there are in fact two Alu subfamilies in galago, one of which contains some sequence not amplified in the human genome (Daniels and Deininger 1983; Daniels et al. 1983). Similarly, it is possible that the differential hybridization of THEs in arthropoids versus prosimians is the result of greater amplification of a particular subfamily of THE elements in higher versus lower primates.

Application of the $-$Screen to Evolutionary and Molecular Questions

An application of the $-$screen technique to the attempt to determine the evolutionary relatedness of organisms can be imagined. Closely related organisms A and C contain a repetitive sequence not found in organism B, as determined by $-$screen. The repetitive sequence probably entered the genome of a common ancestor to C and A after they branched from B. A less likely, though plausible, explanation would be that the element had entered the genomes of the two species on separate occasions.
Clearly, though, in conjunction with other data, the $\phi$-screen could be a powerful tool for confirming and determining branching of evolutionary trees.

Interestingly, in examining numerous clones we detected only three nonsatellite repeat families in the human genome with copy number $>1,000$. Alu is by far the most abundant family, followed by L1 and then by THE. Two of the three families contain sequence that differentially hybridizes between humans and prosimians, the $S'$ end of L1 and THE. It is rather surprising that human and prosimian repeat sequences are so similar, considering that the divergence time for prosimians is 50–75 Myr ago (Britten 1986). However, this could be due, at least in part, to a slowdown in the rate of nucleotide substitutions in the primates (Goodman 1985; Wu and Li 1985; Britten 1986). We feel that the sensitivity of the $\phi$-screen assay could easily be improved to allow detection of sequences with lower copy numbers.

Acknowledgments

The primate DNAs on the dot blots were donated by J. Ahlquist and C. Sibley, except for Galago garnetti DNA, which was prepared from cultured cells provided by the Duke University Primate Center. We thank W. Cavanee for searching for THE RFLPs in a number of human DNAs. J.A.L. wishes to thank L. Lukens and his lab for sharing space, ideas, and lunch. This work was supported by NCI grant CA-00852 and NIH grant GM-30816 to S.S.P.

LITERATURE CITED


ROY J. BRITTEN, reviewing editor

Received April 23, 1986; revision received November 4, 1986.