Alu elements and hominin phylogenetics

Randall K. Garber*, David J. Witherspoon§, Lynn B. Jorde‡, and Mark A. Batzer*‡

*Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803; †Department of Anatomy, Faculty of Medicine, Suez Canal University, Ismailia 4111, Egypt; and §Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, UT 84112

Edited by Morris Goodman, Wayne State University School of Medicine, Detroit, MI, and approved August 26, 2003 (received for review June 18, 2003)

Alu elements have inserted in primate genomes throughout the evolution of the order. One particular Alu lineage (Ye) began amplifying relatively early in hominoid evolution and continued propagating at a low level as many of its members are found in a variety of hominoid genomes. This study represents the first conclusive application of short interspersed elements, which are considered nearly homoplasy-free, to elucidate the phylogeny of hominids. Phylogenetic analysis of Alu Ye5 elements and elements from several other subfamilies reveals high levels of support for monophyly of Hominidae, tribe Hominini and subtribe Hominina. Here we present the strongest evidence reported to date for a sister relationship between humans and chimpanzees while clearly distinguishing the chimpanzee and human lineages.

A

lue elements are short (∼300 bp) interspersed elements that amplify in primate genomes through a process termed retroposition (1–3). The proliferation of these elements has had a significant impact on the architecture of primate genomes (1). They comprise >10% of the human genome by mass and are the most abundant short interspersed element (SINE) in primate genomes (2). The majority of Alu amplification occurred early in primate evolution, and the current rate of Alu retroposition is at least 100-fold slower than the peak of amplification that occurred 30–50 million years ago (mya) (2–5). Alu elements are therefore a rich source of inter- and intraspecies primate genomic variation.

We have previously characterized >2,500 recently integrated Alu elements from the human genome that fall into six distinct subfamilies based on their diagnostic mutations (6–18). In the course of these investigations, we and others have found that individual loci from these subfamilies are informative for the study of primate systematics (19) as well as human population genetics and can be useful tools for resolving the remaining questions of hominid (siamang, orangutan, gorilla, chimpanzee, and human) phylogeny (20, 21).

Resolving the relationships among human (H), chimpanzee (C), and gorilla (G) (i.e., the trichotomy problem) has been particularly difficult. Which of the four possible relationships, ((H,C)G), ((H,G)C), ((C,G)H), and (H,C,G), reflects the true phylogeny of the three species? The consensus approach identifies the chimpanzee as the nearest living relative of humans, but which of the four possible relationships is supported with weak morphological data, the Homo–Pan clade, although some studies prefer the trichotomy hypothesis.”

DNA hybridization data support sister species relationships between humans and chimpanzees (28). Mitochondrial (mt) DNA studies based on restriction endonuclease digestion of 12S rRNA-encoding gene seemed to support either an equidistant relationship among the three lineages or a sister relationship between the chimpanzees and the gorillas (29). Additional mtDNA sequencing supported either the ((H,C)G) or the ((H,G)C) relationship depending on the genes analyzed (30, 31), whereas complete mtDNA sequences supported the former relationship (32).

Some analyses of nuclear loci (e.g., the β-globin cluster) (21, 33) supported the ((H,C)G) relationship, and others (e.g., the involucrin gene) (34) supported the ((C,G)H) relationship. Satta et al. (24) studied sequences from 45 nuclear loci and found that 60% of the loci support the human–chimpanzee clade and the remaining 40% support the two alternatives equally. Eleven of 14 DNA sequence data sets analyzed by Ruvolo (35) supported a human–chimpanzee clade, two supported a chimpanzee–gorilla clade and one supported a human–gorilla clade. The chimpanzees and their sibling species, the bonobos, differ from humans by an average of 0.6% at nonsynonymous sites of 97 studied human genes (25) and 1.2% in overall genomic DNA sequences (26), and they are estimated to have shared a common ancestor with humans 4.0–6.0 mya (22). Gorillas differ from humans by an average of 1.6% in genomic DNA sequences and are estimated to have shared a common ancestor with humans, chimpanzees, and bonobos 6.2–8.4 mya (26). Thus based on nucleotide divergence, chimpanzees and bonobos are the species most closely related to humans.

SINE elements represent a powerful new tool for systematic biology that can be strategically integrated with other conventional phylogenetic characters, most notably morphology and DNA sequences (36–38). There is no known mechanism for the specific removal of SINE elements from the human genome (2), and only one partial deletion of an Alu element has ever been identified (39). Because their mode of evolution is unidirectional (i.e., they do not revert to their ancestral state), individual SINE elements are generally thought to be nearly homoplasy-free characters and are thus useful for resolving phylogenetic and population genetic questions (2, 36–38, 40–43). For example, Shimamura et al. (42) successfully used SINE elements to support the hypothesis that cetaceans (whales, dolphins, and porpoises) form a clade with Artiodactyla (even-toed ungulates, including cows, camels, and pigs). Takahashi et al. (43) also used SINE elements to elucidate the relationships among the cichlid fishes in Lake Malawi. In each of these studies, the presence of a SINE in any particular lineage unambiguously joined the members of that node with only one case of potential homoplasy being introduced either by lineage sorting or via interspecies hybridization.

This does not mean that Alu elements and other SINEs are without problems with regard to phylogenetic analysis. It is known that insertion homoplasy can occur across distantly related taxa as a function of evolutionary time and that variable retroposition rates among species can limit the application of
SINEs to recently diverged phylogenetic relationships (37, 38, 44). However, it is important to note that no instances of Alu insertion homoplasy in hominids have been recovered from the analysis of the recently integrated human Alu insertions (13, 17, 18, 40, 45–47). Random sorting of the ancestral allelic lineages, sequence convergence, and sequence exchanges between alleles or duplicated loci have also been identified as likely factors confounding the interpretation of the interrelationships among species.

Despite these potential problems, Alu elements are novel genetic characters that lack most of the shortcomings encountered when employing sequence or restriction site data. The relatively low rate of retroposition of the Alu Y derived lineages in hominid genomes within the primate order and the relatively late radiation of the hominids (22) make these Alu subfamilies nearly ideal phylogenetic markers to resolve branching order in hominoids. Until now, only a few Alu loci that were informative with regard to the trichotomy issue had been identified (18, 45). A previous study (41) was successful in identifying a limited number of lineage-specific Alu elements; however, because of the limited number of Alu elements examined, questions about hominid phylogeny remained. The additional loci presented in this study facilitate a comprehensive analysis of the phylogeny of the entire group.

Here, we have identified and characterized a total of 153 Alu Ye subfamily members from the draft sequence of the human genome. One hundred seventeen of these loci, as well as 16 loci from other Alu Y lineage subfamilies (17, 18, 45) shown in Fig. 1, were screened using PCR assays to ascertain their relative insertion points in hominid evolution (humans and greater and lesser apes).

**Materials and Methods**

Alu Ye5 subfamily members were identified in the draft sequence of the human genome (August 6, 2001, University of California, Santa Cruz, GoldenPath assembly) by using BLAST (48) queries of the draft sequence to identify exact complements to the oligonucleotide 5’-GAACCCCCGGGGGGGAGGCGTT-GCAG-3’, which is diagnostic for the Ye5 subfamily. We used this approach to identify 153 elements that comprise the Alu Ye5 subfamily. A complete list of all of the Alu elements identified in the searches can be found in Table 1, which is published as supporting information on the PNAS web site, www.pnas.org and at http://batzerlab.lsu.edu.

To aid in elucidating the relationship between the two chimpanzee species (Pan paniscus and Pan troglodytes) and to further clarify relationships among all of the great apes, we searched the available sequences from each of these species in GenBank by using the Alu Y subfamily consensus sequence as a BLAST query.

Sequences containing matches of 90% or greater were submitted to BLAT (http://genome.ucsc.edu; April 2003 assembly) in 20,000-bp batches to identify the homologous region of the human genome and examined computationally to determine whether the Alu insertion in Pan was also present in Homo. If the Alu appeared to be specific to chimpanzee, that locus was passed to the next step: primer design.

Oligonucleotide primers for PCR amplification of each of the Alu loci were designed as described (13, 15, 17). Using the primers and annealing temperatures for Alu Y lineage subfamily members shown in Table 2, which is published as supporting information on the PNAS web site and at http://batzerlab.lsu.edu, we amplified the loci from HeLa genomic DNA and from up to 12 individuals of each nonhuman primate species including Pan troglodytes (common chimpanzee), Pan paniscus (bonobo), Gorilla gorilla (lowland gorilla), Pongo pygmaeus (orangutan), Chlorocebus aethiops sabaenus (green monkey), and Aotus trivirgatus (three-striped owl monkey), which were available as a primate phylogenetic panel (PRP00001) from the Coriell Institute for Medical Research, and from Hylobates syndactylus (siamang). For the Ye5 subfamily, 117 elements identified in the draft human genomic sequence were amenable to PCR analysis. Fourteen loci obtained from the BLAST search of Pan spp. sequences were amenable to PCR analysis. Examination of the orthologous regions of the various nonhuman primate genomes displayed a series of different PCR bands pattern indicative of the time of insertion of each element into the primate genomes (Fig. 2). The distribution of all elements in the various primate genomes tested can be found in Table 3, which is published as supporting information on the PNAS web site and at http://batzerlab.lsu.edu.

**Fig. 1.** Alu consensus sequence alignment. Shown is the alignment of the consensus sequences of each Alu subfamily used in the study. The dots represent the same base present in the Y subfamily consensus sequence. Mutations are denoted by the appropriate base, and deletions are denoted by dashes.
3100 automated DNA sequencer. The sequences of the orthologous Alu insertion loci (bonobo, common chimpanzee, and gorilla) have been assigned accession numbers (AY318829–AY318854). Sequences of all of the Ye5 subfamily members were aligned by using MEGALIGN software (DNASTar version 3.1.7 for Windows 3.2). A multiple alignment of all of the Alu Ye5 subfamily members can be found in Table 1.

To estimate the age of each informative Alu insertion in the human, chimpanzee, and gorilla branch of the tree, mutations from the subfamily consensus sequence were divided into those that occur at CpG dinucleotides and those that occur at non-CpG positions without including small insertions or deletions and ages were estimated as described (13, 15, 17). The mutations are divided into these two different classes because the CpG base positions in repeated sequences mutate at a rate that is ~10 times higher than non-CpG positions (4, 12) as a result of the spontaneous deamination of 5-methylcytosine residues (52). Differences between the two age estimates based on CpG and non-CpG mutation density presumably reflect the heterogeneous level of methylation throughout the genome which influence the CpG mutation spectra of the individual Alu elements differently (53). Here we report age estimates based on CpG mutation density. The estimates were obtained by dividing the number of mutations from the consensus sequence of the subfamily by the neutral rate for CpG mutations (0.0146 per million years).

Results and Discussion

No Alu Ye5 insertions were observed in either green monkey or owl monkey. Thus, we can state that the Alu Ye5 subfamily began to expand after the divergence of hominids from the remainder of the primates (~25 mya) (30). Of the 123 variable loci, 101 were parsimony-informative. Ninety-five loci support nodes unambiguously. Ambiguous loci were defined as loci where nonamplification in a particular taxon made defining the boundaries of a clade uncertain. Parsimony analysis resulted in a single most parsimonious tree (Fig. 3; 124 steps, consistency index = 0.992, RI = 0.995, HI = 0.008). With the exception of one node, bootstrap support was 100%. The node joining the subtribe Hominina (Homo and Pan) was supported with a value of 99%.

The patterns observed clearly indicate a stepwise pattern of insertion reflecting the relative divergence of each group in the hominid lineage. The green monkey (Chlorocebus aethiops sabaeus) was unambiguously placed external to Hominidae (Hylobates, Pongo, Gorilla, Pan, and Homo) by 16 loci. Monophyly of Hominini (Pongo, Gorilla, Pan, and Homo) was supported by 18 unambiguous loci out of 19. Thirty-three of 35 loci unambiguously supported monophyly of the great apes (gorillas, chimpanzees, and humans).

At eight Alu Ye5 loci and two previously identified Alu Yi and Yd loci (18, 45), amplification of filled sites was obtained in human, bonobo, and common chimpanzee. Three of these loci did not amplify filled or empty sites in Gorilla and are thus only ambiguously informative with regard to the trichotomy issue. However, five Alu Ye5 insertions as well as the Yi and Yd insertions were present in the human, bonobo, and common chimpanzee genomes but clearly absent from the gorilla genome. Therefore, a total of seven Alu insertions unequivocally support the hypothesis that the chimpanzee is the nearest living relative of humans as suggested by previous studies (22–24, 27, 28, 31–33). Six of seven Ye5 loci are unambiguously human-specific. Fourteen loci contain Alu elements that are specific to one or both chimpanzees. Thirteen of these belonged to the Yc1 subfamily (17) and one (Ys2DRPα28.2) may belong to an as yet undescribed subfamily. Of these 14 Pan-specific insertions, six were present in both P. paniscus and P. troglodytes. We are therefore able to distinguish the Homo and Pan lineages from one another and support monophyly of the two chimpanzee species with a support value of 100%.

We identified one Alu element (Ye5AH137) that was present in the genomes of human and gorilla but was absent from bonobo and common chimpanzee genomes. This locus, which supports a
Homo–Gorilla clade, is interesting and may be explained by insertion homoplasy at first glance. However, sequence analysis of the locus indicated that it was the result of incomplete lineage sorting of an ancestral polymorphic locus (Fig. 4). Age estimates of this and the other seven Alu insertions at the Homo–Pan node also support this explanation. We calculated the relative ages of each Alu insertion by using CpG mutation densities. Estimates of insertion times for each of the seven loci supporting the Homo–Pan clade range from 2.74 to 5.48 mya (mean = 4.11 ± 1.12 mya). The Ye5AH137 insertion appears to be older than the others with an estimated age of ~6.85 million years. Previously reported sequence data suggests that gorillas shared a common ancestor with humans, chimpanzees, and bonobos 6.2–8.4 mya (26), around the time of the Ye5AH137 insertion. Thus, if the ancestral population were polymorphic for the presence of an Alu insertion at this locus during this period, then fixation of one or the other character state would have occurred randomly in the daughter lineages. In this case, it appears that the insertion was fixed in both human and gorilla lineages but not in chimpanzee. Furthermore, these data lend support to the hypothesis that humans, chimpanzees, and gorillas diverged during a relatively narrow evolutionary time frame (25). It should be noted that

**Fig. 3.** Primate relationships reconstructed by using Dollo parsimony analysis of Alu elements. Primate relationships were derived from analysis of 133 Alu loci by using maximum parsimony criteria. The number of insertions observed along each branch of the tree is indicated, and bootstrap support values are placed above each node.

**Fig. 4.** Lineage sorting model proposed for the Alu Ya5AH137 locus. In this model, Alu Ye5AH137 is polymorphic (indicated as +/-) in the ancestral population that gave rise to humans, chimpanzees, and gorillas. Random fixation of the alternative alleles in each of the three subsequent lineages results in a pattern supporting the Homo–Gorilla association instead of Homo–Pan.
Takahashi et al. (43), in their study of cichlid phylogeny, also found a single SINE locus that may have been influenced by lineage sorting, although interspecies hybridization could also have been involved in that case.

In contrast, two loci that join Homo and Pan (Ye5AH31 and Ye5AH93) had an estimated age of 2.74 million years, substantially younger than the estimated time of the human chimpanzee divergence (∼4–6 mya). Although this is interesting in that these loci suggest very recent coancestry between Homo and Pan, it should be noted that age estimates were based on numbers of CpG mutations from the consensus sequence for the subfamily under examination. The stochastic nature of single nucleotide mutations and the short evolutionary times involved here produce large variances in age estimates for individual Alu elements (12). Therefore, age estimates based on the mean age of several Alu insertions are more accurate than estimates based on individual Alu repeats. Thus, the apparently young age estimates for these two insertions should not be taken as evidence for a more recent divergence between chimpanzees and humans when the mean age of all of the relevant insertions is ∼4.11 million years.

The utility of SINE insertions, and mobile elements in general, for phylogenetic analysis continues to be bolstered by studies such as this one. Here, we present the first application of SINES to fully elucidate the phylogeny of the hominid lineage and present the strongest evidence to date for phylogenetic relationships among the hominid lineages. Of the 133 Alu insertion loci, 95 were unambiguously informative for determining the relative divergence of each of the major lineages. In addition, seven Alu insertions informative with regard to the Homo–Pan–Gorilla trichotomy unambiguously support monophyly of humans and chimpanzees. Finally, six loci discovered by using searches of the currently available chimpanzee genomic sequences unambiguously join the two Pan species to the exclusion of Homo. Only a single insertion from the entire data set showed any potential evidence of insertion homoplaspy; however, the distribution of this element appears to be the result of lineage sorting of an ancestral polymorphism.

We thank Dr. Caro-Beth Stewart for providing Siamang DNA. This research was supported by Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05-05, 2000-05-01, and 2001-06-02) (to M.A.B.), National Institutes of Health Grant RO1 GM59290 (to M.A.B. and L.B.J.), and National Science Foundation Grants BCS-0218338 (to M.A.B.) and BCS-0218370 (to L.B.J.).