

# An *Alu*-Based Phylogeny of Gibbons (Hylobatidae)

Thomas J. Meyer,<sup>1</sup> Adam T. McLain,<sup>1</sup> J. Michael Oldenburg,<sup>1</sup> Christopher Faulk,<sup>1,2</sup> Matthew G. Bourgeois,<sup>1</sup> Erin M. Conlin,<sup>1</sup> Alan R. Mootnick,<sup>3</sup> Pieter J. de Jong,<sup>4</sup> Christian Roos,<sup>5</sup> Lucia Carbone,<sup>6</sup> and Mark A. Batzer<sup>\*,1</sup>

<sup>1</sup>Department of Biological Sciences, Louisiana State University

<sup>2</sup>Department of Environmental Health Sciences, University of Michigan

<sup>3</sup>Gibbon Conservation Center, Santa Clarita, California

<sup>4</sup>Children's Hospital Oakland Research Institute, Oakland, California

<sup>5</sup>Gene Bank of Primates and Primate Genetics Laboratory, German Primate Center, Göttingen, Germany

<sup>6</sup>Department of Behavioral Neuroscience, Oregon Health & Science University

\*Corresponding author: E-mail: mbatzer@lsu.edu.

Associate editor: Norihiro Okada

DNA sequences reported in this article have been deposited in GenBank (JQ611584-JQ611670, JQ861884-JQ861906, and JX034756-JX034761).

## Abstract

Gibbons (Hylobatidae) are small, arboreal apes indigenous to Southeast Asia that diverged from other apes ~15–18 Ma. Extant lineages radiated rapidly 6–10 Ma and are organized into four genera (*Hylobates*, *Hoolock*, *Symphalangus*, and *Nomascus*) consisting of 12–19 species. The use of short interspersed elements (SINEs) as phylogenetic markers has seen recent popularity due to several desirable characteristics: the ancestral state of a locus is known to be the absence of an element, rare potentially homoplasious events are relatively easy to resolve, and samples can be quickly and inexpensively genotyped. During radiation of primates, one particular family of SINEs, the *Alu* family, has proliferated in primate genomes. *Nomascus leucogenys* (northern white-cheeked gibbon) sequences were analyzed for repetitive content with RepeatMasker using a custom library. The sequences containing *Alu* elements identified as members of a gibbon-specific subfamily were then compared with orthologous positions in other primate genomes. A primate phylogenetic panel consisting of 18 primate species, including 13 gibbon species representing all four extant genera, was assayed for all loci, and a total of 125 gibbon-specific *Alu* insertions were identified. The resulting amplification patterns were used to generate a phylogenetic tree. We demonstrate significant support for *Symphalangus* as the most basal lineage within the family. Our findings also place *Nomascus* as a derived lineage, sister to *Hoolock*, with the *Nomascus*–*Hoolock* clade sister to *Hylobates*. Further, our analysis groups *N. leucogenys* and *Nomascus siki* as sister taxa to the exclusion of the other *Nomascus* species assayed. This study represents the first use of SINEs to determine the genus level phylogenetic relationships within the family Hylobatidae. These relationships have been resolved with robust support at most internal nodes, demonstrating the utility of SINE-based phylogenetic analysis. We postulate that hybridization and rapid radiation may have contributed to the complex and contradictory findings of the previous studies. Our findings will aid in the conservation of these threatened primates and inform future studies of the biogeographical history and distribution of modern gibbon species.

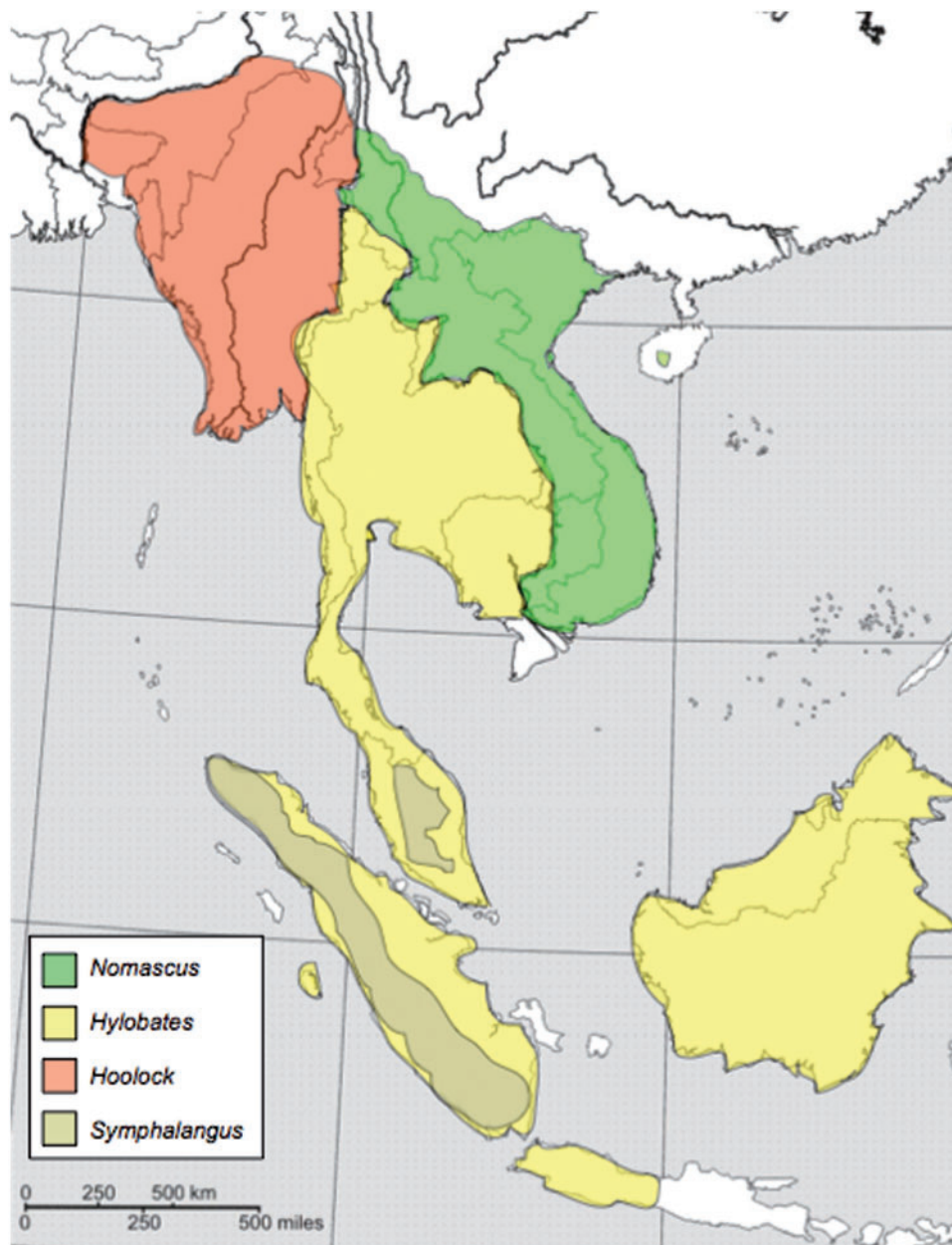
**Key words:** primate, ape, phylogeny, retrotransposon, SINE, mobile elements.

## Introduction

Gibbons (Hylobatidae) are small, arboreal apes indigenous to Southeast Asia. They occupy a range stretching from north-eastern India, south to Sumatra, Borneo, and Java, and north into the southernmost parts of China, with ranges among genera being largely separated by major rivers (fig. 1) (Geissmann 1995). All are threatened or endangered due to human activities, some critically (Mootnick 2006). Gibbons are by far the most speciose group of apes (Hominoidea) (Groves 2001), with most of the radiation of extant lineages occurring rapidly between 6 and 10 Ma (Matsudaira and Ishida 2010). They are organized into four genera (*Hylobates*, *Hoolock*, *Symphalangus*, and *Nomascus*) primarily based on diploid chromosome number, and there are

between 12 and 19 extant species, according to various classification schemes (Roos and Geissmann 2001; Chatterjee 2006; Mootnick 2006; Carbone et al. 2009; Thinh et al. 2010a, 2010b). The Hylobatidae represents a valuable perspective on the phylogenetic link between the great apes and the Old World monkeys (Cercopithecoidea) as well as a useful position just before the radiation of great apes. The Old World monkeys diverged from the apes ~25 Ma (Gibbs et al. 2007; Han et al. 2007), whereas the first divergence within the great apes that of the orangutan lineage occurred ~14 Ma (Locke et al. 2011). The current estimates show that gibbons diverged from the other apes ~15–18 Ma (Carbone et al. 2009).

Although clearly delineated from other apes, questions about the complex systematic relationships between



**Fig. 1.** Geographic distribution of the four gibbon genera. The Mekong River divides the *Nomascus* and mainland *Hylobates* groups. The *Hoolock* and mainland *Hylobates* groups are divided by the Salween River. Sympatry exists between *Symphalangus* and some members of the *Hylobates* group with conspecific breeding likely avoided due to significant size differences between these taxa. Map after Geissmann (1995) (<http://www.gibbons.de/main/system/system.html>).

gibbon genera and species remain contentious. Numerous phylogenies based on the behavior, morphology, geographic distribution, karyotype, and genetic analyses have been constructed, often leading to more questions than answers (Garza and Woodruff 1992; Geissmann 1995; Roos and Geissmann 2001; Muller, Hollatz, and Wienberg 2003; Takacs et al. 2005; Chatterjee 2006; Matsudaira and Ishida 2010; Thinh et al. 2010a; Kim et al. 2011). Especially debatable are the divergence dates of individual genera and the determination of which lineage is most basal.

The studies of mitochondrial markers have recently begun to achieve some statistical support for various phylogenetic

hypotheses. A study by Takacs et al. (2005) of sequence from the mitochondrial ND3–ND4 region was able to resolve species relationships within genera, but it was unable to resolve relationships between genera. Another more recent mitochondrial study of complete *cytochrome b* gene sequences also failed to robustly resolve the branching patterns among the four genera (Thinh et al. 2010a). However, analysis of complete mitochondrial sequences found *Nomascus* supported as the most basal group within Hylobatidae (Chan et al. 2010; Matsudaira and Ishida 2010). In addition, a surprising frequency of chromosomal rearrangement within Hylobatidae, resulting in a diverse range of diploid

chromosome numbers (*Nomascus*,  $2n = 52$ ; *Symphalangus*,  $2n = 50$ ; *Hoolock*,  $2n = 38$ ; and *Hylobates*,  $2n = 44$ ), has contributed to the confusion over the relationships between gibbons (Muller et al. 2003; Carbone et al. 2009).

The use of retrotransposon insertions as phylogenetic and population genetic markers has seen an increasing popularity in recent years (Okada 1991; Minghetti and Dugaiczky 1993; Novick et al. 1993; Batzer et al. 1994; Stoneking et al. 1997; Hamdi et al. 1999; Nasidze et al. 2001; Romualdi et al. 2002; Salem et al. 2003; Watkins et al. 2003; Roos et al. 2004; Shedlock et al. 2004; Ray et al. 2005; Schmitz et al. 2005; Nikaido et al. 2006; Witherspoon et al. 2006; Herke et al. 2007; Xing et al. 2007a; Osterholz et al. 2008; Li et al. 2009; Osterholz et al. 2009; Nilsson et al. 2010; Roos et al. 2011; Suh et al. 2011). Short interspersed elements (SINEs) are <500 bp in length and generate new copies of themselves through reverse transcription of an RNA intermediate into a new position in the genome (Luan et al. 1993; Luan and Eickbush 1995; Cost et al. 2002). This pattern of mobilization results in an increase in the copy number of most SINEs within their host genomes. The copies of these elements produce many important impacts on the structure and function of the genome, especially when their insertion disrupts coding or regulatory regions (Callinan and Batzer 2006; Belancio et al. 2009; Cordaux and Batzer 2009; Belancio et al. 2010; Konkel and Batzer 2010; Konkel et al. 2010; Beck et al. 2011). However, coding and regulatory regions comprise only ~5% of the genome (Lander et al. 2001), so that most new SINE insertions are likely to cause no disruption during their insertion (Cordaux et al. 2006; Konkel and Batzer 2010). This subset of insertions into neutrally evolving regions of the genome has a number of characteristics that make them desirable as phylogenetic markers. Genetic markers are said to be homoplasious if the phylogenetic tree inferred from them is not the true tree. SINEs are nearly homoplasious-free markers because the ancestral state of any locus is known to be the absence of the element (Batzer and Deininger 1991; Murata et al. 1993; Batzer et al. 1994; Shedlock and Okada 2000; Okada et al. 2004; Ray et al. 2006; Ray 2007; Xing et al. 2007b). The presence of a SINE insertion at a specific location within the genomes of two or more lineages is evidence that the insertion occurred at a point in time when the lineages shared a common ancestor. Potentially homoplasious events, such as precise deletion and parallel independent insertion, have been shown to be exceptionally rare, with most suspected cases being easily resolved by sequencing the suspicious loci (Ray et al. 2006; Xing et al. 2007a; Li et al. 2009). However, most loci are quickly and easily genotyped using a combination of polymerase chain reaction (PCR) and gel electrophoresis.

The expansion of one particular family of SINEs, the *Alu* family, began ~65 Ma and has dominated within the radiation of primates ever since (Batzer and Deininger 2002; Deininger and Batzer 2002; Kriegs et al. 2007; Konkel et al. 2010; Perelman et al. 2011). *Alu* elements are the most successful lineage of mobile elements in primate genomes, being present in ~1.2 million copies in the human genome and having been found in every primate sequenced thus far

(Lander et al. 2001; Chimpanzee Sequencing and Analysis Consortium 2005; Gibbs et al. 2007; Locke et al. 2011). An *Alu* element is ~300 bp long and is nonautonomous; hence, it does not encode the enzymatic machinery necessary for its own reverse transcription. Instead, *Alu* elements hijack the enzymatic machinery of a long interspersed element, L1 (Schmid 2003). The evolution of *Alu* elements within the primate radiation has resulted in a large number of subfamilies of elements, identifiable by the diagnostic mutations that they share with their progenitor copies (reviewed in Cordaux and Batzer 2009; Konkel et al. 2010). The mobilization activity of these subfamilies varies with time, allowing researchers to tailor their assays to specific subfamilies active only in the lineages and during the time periods in which they are interested (Churakov et al. 2010). Because of this, phylogenetic studies of primates using *Alu* elements as markers have been increasingly popular, helping to elucidate the relationships in many primate taxa, including those within the Hominae (Salem et al. 2003), Catarrhini (Xing et al. 2005, 2007a), Platyrrhini (Ray et al. 2005; Osterholz et al. 2009), and Strepsirrhini (Roos et al. 2004) clades, as well as the more detailed relationships within the genus *Macaca* (Li et al. 2009) and the Colobinae (Osterholz et al. 2008; Roos et al. 2011) and the affiliation of the genus *Tarsius* to other primates (Zietkiewicz et al. 1999; Schmitz et al. 2001).

In this study, we resolve relationships within gibbons using *Alu* elements as phylogenetic markers. We computationally screened the available genomic sequences of *Nomascus leucogenys* that have been generated as a part of the ongoing gibbon genome project. PCR assays in a panel of 18 primate species, including 13 gibbon species, produced 125 gibbon-specific *Alu* insertions that were used to generate a phylogeny of the Hylobatidae.

## Materials and Methods

### Computational Analysis of Candidate Loci

Genomic sequence generated by the gibbon sequencing consortium for *N. leucogenys* (northern white-cheeked gibbon) in the form of a large number of sequenced bacterial artificial chromosome (BAC) clones was obtained from the Ensembl database system (Hubbard et al. 2009). A local installation of RepeatMasker was then used to scan the sequences on the sensitive setting to classify all identifiable repetitive sequences (Smit et al. 1996–2010). Next, in-house Perl scripts were used to filter the RepeatMasker output, keeping only those sequences that contained *Alu* elements classified by RepeatMasker in the *AluYc3* subfamily. However, further inspection revealed these elements to have been misidentified. They were instead members of the *AluYd3a1\_gib* subfamily, which was not in the RepeatMasker RepBase library available at the time of analysis (Jurka 2000). This yielded a total of 9,701 candidate sequences that contained putatively gibbon-specific *Alu* insertions.

These computationally derived candidates were then queried against four available outgroup genomes, the human (hg19), chimpanzee (panTro2), orangutan (ponAbe2), and rhesus macaque (rheMac2), using BLAT searches (Kent



2002) via the UCSC Genome Bioinformatics website (<http://genome.ucsc.edu/>). Only those sequences whose putatively gibbon-specific *Alu* insertions were not present at the orthologous loci in the outgroups were retained for further analysis. In addition, the position of each orthologous locus in the human genome was noted for each candidate sequence that passed our BLAT analysis. Since at the time of analysis, the Ensembl data set represented an unassembled genome, many of the sequences were duplicate reads of the same region. The orthologous human positions of the gibbon-specific *Alu* insertions were used to identify all sequences in the data set that corresponded to a single locus. A total of 430 gibbon-specific candidate loci were identified using this strategy.

### PCR Amplification and DNA Sequencing of Candidate Loci

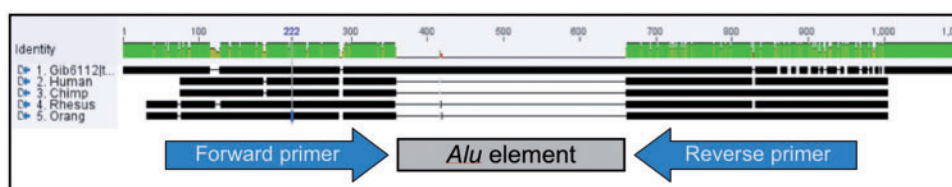
Flanking oligonucleotide primers for PCR amplification of each gibbon-specific *Alu* element were designed using the CLC Main Workbench v.5 software suite ([cited 2012 Jun 25], <http://www.clcbio.com/index.php?id=92>). This involved the creation of an alignment of each gibbon sequence to any orthologous sequences obtained from the outgroup genomes. The gibbon-specific *Alu* element and ~20 bp of flanking sequence upstream and downstream were designated as the region to amplify (fig. 2). Oligonucleotide primers were selected as close to the element as possible while still amplifying some orthologous flanking sequence. This is because increased amplicon sizes containing excessive flanking sequence into which near parallel independent insertions may occur can potentially confound analyses. This problem is greater for analyses of distantly related species because researchers may need to search further into flanking regions to find sequences sufficiently conserved to allow primer amplification across all species. However, with due diligence and sequencing of questionable loci, this can be overcome. Next, the suggested primers were screened computationally against the available human, chimp, orangutan, and rhesus macaque genomes using the In-Silico PCR function available at the UCSC Genome Bioinformatics website. This was done to reaffirm the established gibbon specificity of the locus to be amplified and to determine whether the primer pairs amplified single loci (i.e., they were not designed within other repeats). Of the 430 candidate loci, only 231 had sufficient

unique flanking sequence available to allow primers to be designed that passed in silico PCR analysis.

Because our supply of some gibbon species and individual DNA was limited, and to verify that our primers amplified unambiguous loci and absent sites in non-gibbon primates, we first PCR amplified each *Alu* insertion locus against a primate panel consisting of one of each of the following species: *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo pygmaeus*, *Hylobates lar* (PR00715), *Nomascus gabriellae* (PR00652), *Symphalangus syndactylus* (KB11539), and *Chlorocebus aethiops*. Of the 231 primer pairs tested, pairs that did not result in amplification in many species, amplified multiple paralogous fragments in each species, or produced smears were removed from further analysis. The remaining loci were amplified in a larger gibbon panel consisting of 31 individuals representing 18 species as follows: one *H. sapiens*, one *P. troglodytes*, one *G. gorilla*, one *P. pygmaeus*, four *N. leucogenys*, one *Nomascus siki*, one *Nomascus annamensis*, two *N. gabriellae*, three *Hylobates moloch*, one *Hylobates agilis*, one *Hylobates albibarbis*, one *Hylobates muelleri*, two *H. lar*, two *Hylobates pileatus*, one *Hylobates klossii*, three *Hoolock leuconedys*, three *S. syndactylus*, and one *C. aethiops*. A complete list of all DNA samples used in our analyses is available in the [supplementary data \(supplementary table S1, Supplementary Material online\)](#).

Due to the limited quantity of genomic DNA available for some individuals, some samples were subjected to whole genome amplification using the GenomiPhi genome amplification kit (Amersham, Sunnyvale, CA) following the manufacturer's instructions. PCR amplification was performed in the following conditions: 25  $\mu$ l total volume using 15 ng of template DNA, 200 nM of each primer, 200  $\mu$ M dNTPs in 50 mM KCl, 1.5 mM  $MgCl_2$ , 10 mM Tris-HCl (pH 8.4), and 2 units of *Taq* DNA polymerase. The conditions for PCR were as follows: an initial denaturation at 95°C for 1 min, followed by 32 cycles of denaturation at 95°C, annealing at 55°C, and extension at 72°C for 30 s each, followed by a final extension of 72°C for 1 min. Products were run out on 2% agarose stained with 0.25  $\mu$ g ethidium bromide and visualized with ultraviolet fluorescence. The complete list of loci and primer sequences are available in the [supplementary data \(supplementary table S2, Supplementary Material online\)](#).

Any loci that indicated contradictory relationships to those shown in the most parsimonious tree were also subjected to automated DNA sequence analysis to verify the



**Fig. 2.** An illustration of how multiple alignments of orthologous loci are used to design oligonucleotide primers for lineage-specific insertions. The multiple alignment was performed using Clustal X. The gibbon sequence is *N. leucogenys* and contains a gibbon-specific *Alu* insertion (indicated by the 300 bp region in the center that does not match the other sequences). Also included in this alignment are *H. sapiens*, *P. troglodytes*, *Pongo abelii*, and *Macaca mulatta*. The gray rectangle shows the location of the *Alu* insertion in the gibbon. The arrows indicate the orthologous flanking regions in which primers are designed that will match all five species in the alignment.

computationally derived *Alu* insertion sequences and to confirm the phylogenetic distribution of the insertions. In the case of these potentially homoplasious loci, representative filled and empty sites from the taxa involved were selected for sequencing analysis to determine the origin of the homoplasious signal. Ten such loci were apparent in our analysis and investigated through DNA sequencing. PCR products from selected individuals were directly sequenced as described previously (Ray et al. 2005). In addition, six critical diagnostic loci that inform the center of our tree (three grouping *Nomascus*, *Hoolock*, and *Hylobates*, and three grouping *Nomascus* and *Hoolock*) were subjected to DNA sequencing. All sequences were then aligned against orthologous sequences from other sequenced individuals and those obtained from BLAT analysis. These alignments are available in the [supplementary data](#) (Supplementary Material online), and all DNA sequences generated for this project have been deposited in GenBank under accession numbers (JQ611584-JQ611670, JQ861884-JQ861906, and JX034756-JX034761). In all but the one potentially homoplasious instance (locus m37; see Results and Discussion), sequencing allowed the resolution of the contradicting loci. All nine of the critical diagnostic loci reported had the insertions and associated target site duplications (TSDs) and flanking confirmed through sequencing, confirming significant support for these loci (see below).

### Phylogenetic Analysis

A total of 125 loci were amplified in a majority of the gibbons in the 31 individual panel as well as in at least 4 of the 5 outgroups and were therefore used in the phylogenetic analysis. Each locus for each individual was scored based on whether it amplified as a filled site (presence of an *Alu* insertion), an empty site (absence of an insertion), or showed no amplification (unknown insertion status) (fig. 3). Bands in individuals heterozygous for both filled and empty sites were sequenced, and if the filled site was found to be an identical insertion, they were scored as having a filled site in our matrix. Filled sites were scored as a “1,” empty sites were scored as a “0,” and unknown sites were scored as “?” Scores were entered using the Mesquite program ([cited 2012 Jun 25], <http://mesquiteproject.org/>), and all loci were set to Dollo.up for parsimony analysis. The matrix used in the analysis can be found in the [supplementary data](#) (supplementary table S3, Supplementary Material online).

A heuristic search in PAUP\* 4.0b10 (Swofford 2000) was performed as described by Xing (2005). Briefly, *C. aethiops* was set as the outgroup, and all loci were treated as individual insertions following the criteria required for Dollo parsimony analysis. Trees and statistics were generated by PAUP\*. Because our loci were harvested from the *N. leucogenys* genome, no resolution of the relationships within *Hylobates* was possible, resulting in a large polytomy for the members of this genus. Therefore, in the final tree, only *H. lar* and *H. agilis* were included as representatives of the genus. Tree space was limited to 100 trees, and 10,000 bootstrap replicates were run to provide percentage support for each branch under the

majority rule tree, which omits bootstrap values less than 50% per branch (126 steps; consistency index [CI] = 0.9921; homoplasy index [HI] = 0.0079; and retention index [RI] = 0.9935). The statistically significant number of SINE insertions required for acceptance is provided under the likelihood model described by Waddell et al. (2001) (fig. 4). The tree was visualized using the FigTree software ([cited 2012 Jun 25], <http://tree.bio.ed.ac.uk/software/figtree/>).

## Results and Discussion

### Computational Data Mining of the *N. leucogenys* Genome

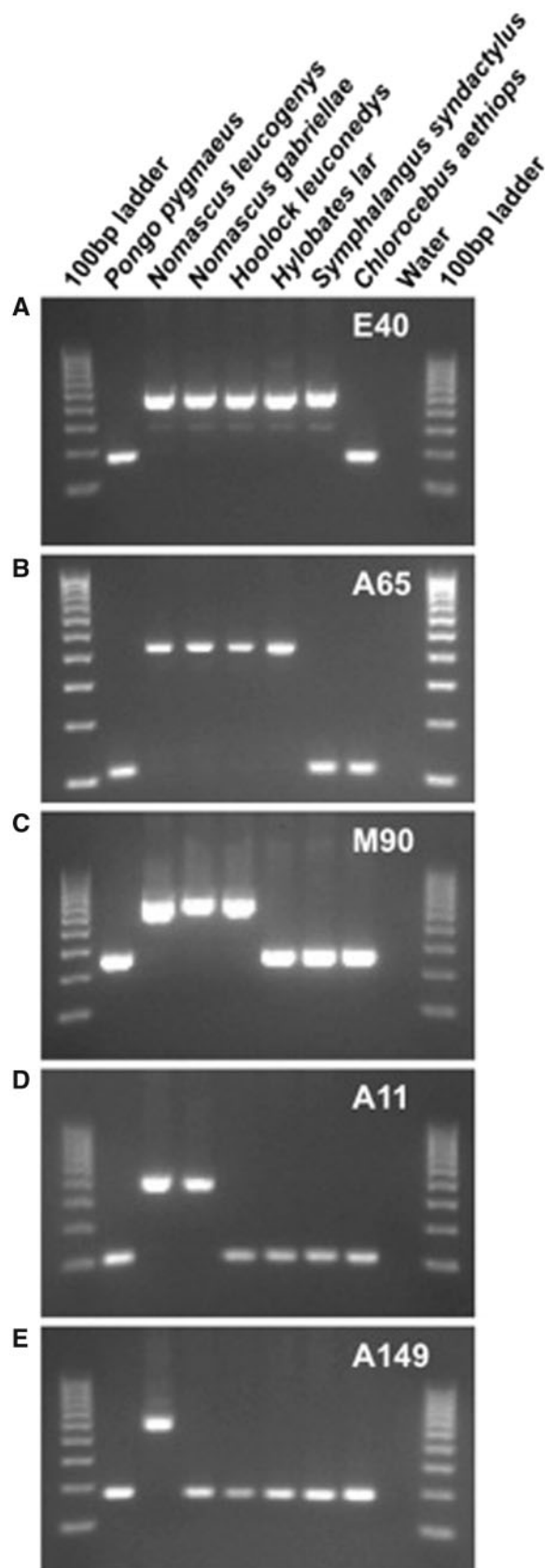
The unassembled sequence data available from the northern white-cheeked gibbon (*N. leucogenys*) genome sequencing project consists of 750,923 sequences generated via combined whole-genome shotgun plasmid, fosmid, and BAC end sequences. We analyzed the repetitive content of this data set using the RepeatMasker program. We identified 9,701 sequences of interest that contain *Alu* elements identified as belonging to the *AluYc3* subfamily via an in-house Perl script (available on request). These sequences were subjected to BLAT alignment to determine whether the *Alu* of interest was shared with outgroup species from the Hominidae (human, chimpanzee, and orangutan) or the Old World monkeys (rhesus macaque). We later determined that those *Alu* insertions that were gibbon specific were not, in fact, *AluYc3* elements, but were actually members of a previously reported gibbon-specific subfamily, *AluYd3a1\_gib*. Although this subfamily was present in the RepBase database, it was not included in the repeat library used in the RepeatMasker analysis, explaining the initial misidentification.

The *N. leucogenys* genome will be a 6x coverage genome once assembly is complete. However, at the time of analysis, the sequence was unassembled, and as a result, many of the sequences analyzed were duplicate reads. Some loci were present in a single copy, but others were present in the data set more than 35 times. In total, 430 unique, computationally derived candidates containing a gibbon-specific *Alu* element were identified (fig. 2).

### PCR Assays of Gibbon-Specific *Alu* Insertions

Multiple alignments of gibbon and outgroup sequences and automated primer design were attempted for all 430 computational candidates. Some loci were removed from analysis due to a lack of sufficient or unique sequence flanking the *Alu* insertion in which to design primers. Primers were tested using in silico PCR to determine whether they would amplify single loci in most of the outgroups. A total of 231 loci passed this scrutiny and were then used in PCR verification.

Each locus was first amplified on a panel of primate DNA representing one individual from each of eight species, three of which were gibbons. Those primer pairs that amplified well on this panel and indicated gibbon specificity were then run against the second, larger primate panel. This panel included 31 individuals from 18 species; 26 of these individuals were gibbons from 13 gibbon species including species from each of the 4 gibbon genera.



**FIG. 3.** PCR amplification assay demonstrating how gibbon-specific *Alu* insertion polymorphisms can be used to distinguish lineages. The five gel electrophoresis images show the PCR amplification products from five primer sets in seven primate species, including five gibbons, one hominid, and one Old World monkey. The DNA template for each lane

The 125 loci that amplified unique fragments in a majority of the species on the panel were scored for use in the phylogenetic analysis. Figure 3 demonstrates how lineage-specific insertions can be used to infer the phylogenetic relationships between groups. Loci that produced smears or multiple nonspecific banding patterns were removed from further analysis. Several loci produced double-banding patterns in some species or individuals. These loci may represent a useful set of markers for population genetics studies of these species, as such patterns may be indicative of heterozygous individuals, indicating an insertion that is still polymorphic in the population (Cordaux et al. 2007; Ray et al. 2007). Other possible explanations for such loci include an *Alu* insertion that occurred after a duplication event generated two paralogous regions or introgression stemming from hybridization events (Xing et al. 2007a; Osterholz et al. 2008).

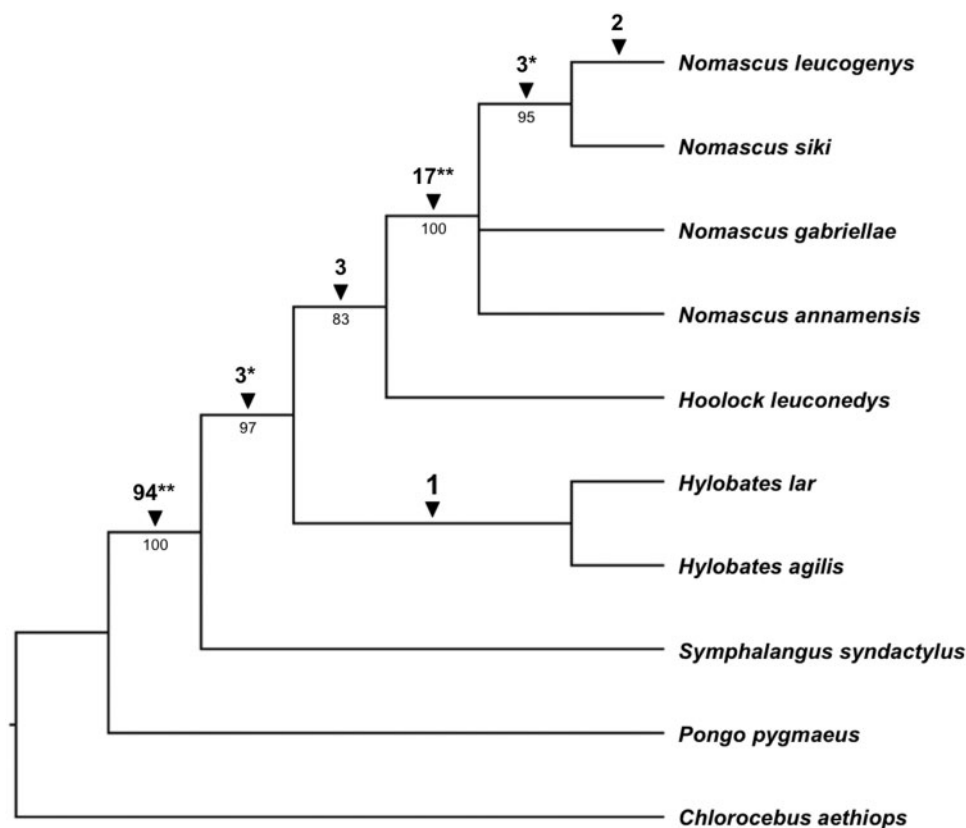
### Phylogenetic Relationships within the Hylobatidae

In total, we used 125 loci in the phylogenetic analysis. A heuristic search was performed using the Dollo parsimony method, with *C. aethiops* designated as the outgroup taxon. A total of 108 of the loci were found to be parsimony informative, resulting in a single most parsimonious tree (126 steps; CI = 0.9921; HI = 0.0079; RI = 0.9935) (fig. 4). In total, 10,000 bootstrap replicates were performed as well as a likelihood test based on the number of unambiguous *Alu* insertions supporting each node (Waddell et al. 2001). Bootstrap values, significance levels, and the number of supportive insertions for each node are indicated in figure 4. The likelihood tests of every genus level node except one within the phylogenetic tree were found to be significant.

The topology of our tree clearly defines relationships between the four extant genera within Hylobatidae: *Nomascus*, *Hoolock*, *Hylobates*, and *Symphalangus*. Previous attempts to reconstruct a genus level tree for Hylobatidae have proven problematic, resulting in many different topologies (Geissmann 1995; Roos and Geissmann 2001; Takacs et al. 2005; Chatterjee 2006; Chan et al. 2010; Matsudaira and Ishida 2010; Thinh et al. 2010a; Kim et al. 2011), often with insufficient support at the genus level nodes. This is likely due to the relatively recent and rapid radiation within the group (Matsudaira and Ishida 2010) leading to incomplete lineage sorting of characters used in phylogenetic analyses. Our finding for the most basal lineage is consistent with some of these previous studies. We demonstrate significant support for *Symphalangus* as the most basal lineage within the family. This was the finding in a previous study utilizing morphometric data (Creel and Preuschoft 1984) and one cytochrome *b*

is shown at the top, and the locus is indicated in the upper right of each image. Larger bands indicate “filled” sites in which *Alu* elements have inserted, whereas lower bands indicate “empty” sites containing no *Alu* insertion. (A) An *Alu* insertion present in all gibbon species from all four genera. (B) An *Alu* insertion present only in *Nomascus*, *Hoolock*, and *Hylobates*. (C) An *Alu* insertion restricted only to *Nomascus* and *Hoolock*. (D) An *Alu* insertion that is *Nomascus* specific. (E) An *Alu* insertion that is present only in the species *N. leucogenys*.





**Fig. 4.** Cladogram of Hylobatidae phylogenetic relationships derived from 125 *Alu* insertion polymorphisms. Amplification patterns of insertion loci were used to construct a Dollo parsimony tree of genus level gibbon phylogenetic relationships using *C. aethiops* as an outgroup. Also included is *P. pygmaeus*, the orangutan, which is the hominid that diverged from other hominids relatively soon after gibbons diverged from other hominoids. The numbers below the branches are percentage bootstrap replicates (out of 10,000 iterations) producing trees that include that node. Numbers above the branches indicate the number of *Alu* insertions supporting the node. The significance level of each node supported by insertions as determined by likelihood testing is indicated by either \* $P < 0.05$  or \*\* $P < 0.01$ .

analysis (Garza and Woodruff 1992). However, this is inconsistent with other more recent cytochrome *b* analyses (Chatterjee 2006; Thinh et al. 2010a) as well as analyses of morphological and behavioral characters (Haimoff et al. 1982), chromosomal homology (Muller et al. 2003), and the mitochondrial control region (Roos and Geissmann 2001). Our findings also place *Nomascus* as a derived lineage, sister to *Hoolock*, with the *Nomascus*–*Hoolock* clade forming a sister to *Hylobates*. The insertions supporting these affiliations were subjected to DNA sequencing and analysis, and in every case, the shared insertion architecture was confirmed. Some previous studies placed *Nomascus* as the most basal clade (Roos and Geissmann 2001; Chan et al. 2010; Thinh et al. 2010a), representing a distinct difference in our reconstruction from most previous ones. Within the *Nomascus*, we found *N. leucogenys* and *N. siki* to be more closely related to one another than either was to the other two *Nomascus* species assayed. This is not surprising given that *N. siki* was previously classified as a subspecies of *N. leucogenys* (Garza and Woodruff 1992; Geissmann 1995), and this relationship is consistent with other recent reconstructions (Thinh et al. 2010a; Thinh et al. 2010c). Given the frequency of chromosomal rearrangements among gibbons and the rapid radiation of extant lineages, we favor the *Alu*-based phylogenetic reconstruction method over the alternatives. This method has allowed us

to pull a robust phylogenetic signal out of the data despite the difficulties that other systems have encountered with this group.

Only one locus in the analysis, m37, did not support the most parsimonious tree. This locus is the source of the small amount of homoplasy ( $HI = 0.0079$ ) observed. The insertion was found in all *Nomascus* species and all *Hylobates* species except *Hylobates pileatus*. However, this insertion was absent in *Hoolock* and *Symphalangus*. This locus was among the subset sequenced to verify the computational findings and analyze the sequence architecture of the insertion site in detail. It shows identical TSDs and identical *Alu* elements across all species sequenced. Possible explanations for these observations are that this locus represents an example of precise parallel insertion and/or precise deletion in the lineages affected or that this locus represents an incomplete lineage sorting event. We favor the latter explanation given the relatively rapid speciation time over which the extant gibbon genera are known to have diverged from one another (6–10 Ma) (Matsudaira and Ishida 2010). This locus affects the likelihood values supporting the *Nomascus*–*Hoolock* clade. With three loci supporting this clade, and one supporting a *Nomascus*–*Hylobates* grouping, the support for this branch is only  $P = 0.111$  (fig. 4).

The inconsistency with which the topology of the gibbon genus level tree has been resolved in the past, and our own difficulty in obtaining robust support for the *Nomascus*–*Hoolock* grouping, is presumably the result of the rapid radiation of extant genera. However, introgression may also play a large role. Gibbon species have been known to hybridize in the wild and generate viable offspring (Myers and Shafer 1979). Furthermore, most of the previous studies of gibbon phylogenetics involve samples taken from either zoo specimens or specimens from captive breeding programs. Gibbons are notoriously difficult to reliably identify to the species level, and hybridization in captivity is known to occur (Mootnick 2006). If such hybrids were misidentified as purebred individuals in previous studies, they could potentially confound the true phylogenetic signal (Churakov et al. 2009). In fact, the individual sampled need not be a direct hybrid, as such introgression events could have occurred several generations in the past and still adversely affect identification. However, given the support for our topology and the fact that only one homoplasious locus was identified in our data set, we believe it is unlikely that sufficient introduction of *Alu* insertion loci from other lineages has occurred via hybridization between members of different gibbon genera to affect our conclusions. In addition, this analysis represents a necessarily single-sided view of gibbon phylogeny, as it is dependent on the complete genome of *N. leucogenys*. In the future, we hope that complete genomes for members of the *Hylobates* and *Hoolock* genera become available to allow a comparison with our results from *Nomascus*. The *Alu* insertion markers developed by this, and future analyses will help conservation efforts by providing relatively inexpensive and quick assays to determine the species identity of individuals being selected for breeding programs.

## Conclusion

This study represents the first use of SINEs to investigate the phylogenetic relationships within the family Hylobatidae. By using a combination of computational and wet bench techniques, a total of 125 loci containing gibbon-specific *Alu* insertions were found. These were used to generate a phylogenetic hypothesis of gibbon phylogeny, with statistically significant support for all but one internal branch. The nearly homoplasmy-free nature of *Alu* insertions is ideally suited to the Dollo parsimony method for the reconstruction of well-supported primate phylogenies. We found *Symphalangus* to be the most basal lineage, followed by a divergence between the *Hylobates* and a clade consisting of *Hoolock* and *Nomascus*. The four *Nomascus* species analyzed, *N. leucogenys*, *N. siki*, *N. annamensis*, and *N. gabriellae*, were also well supported as a clade to the exclusion of *Hoolock*. Furthermore, *N. leucogenys* and *N. siki* were grouped as sister species within the *Nomascus*. The most striking difference between our tree and those produced in the previous analyses is the grouping of *Nomascus* with *Hoolock* as the most derived multigenera clade. A single locus in our data set contradicting this relationship may be the result of the hybridization and rapid radiation characteristic of the gibbon radiation. These factors may have contributed to the complex

and contradictory findings of previous studies. In addition, our findings may inform future studies of the biogeographical history and distribution of modern gibbon species. It is our hope that the markers developed in this study will provide a means by which individuals may be quickly and inexpensively genotyped by conservation workers, so that hybridization can be avoided in breeding programs.

## Supplementary Material

Supplementary tables S1–S3 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

## Acknowledgments

The authors wish to express their deep sadness for the loss of their colleague and coauthor, Alan R. Mootnick, who died during the preparation of this manuscript. He worked selflessly, establishing the Gibbon Conservation Center in 1976. He oversaw a successful gibbon breeding program there that gave hope that some of these fascinating apes may yet survive their endangered status. In addition, he worked tirelessly with the academic community to promote research into the animals he so loved. He will be greatly missed. They thank the Parc Zoologique de Clères, the Gladys Porter Zoo, the Los Angeles Zoo, the Fort Wayne Children's Zoo, the Henry Doorly Zoo, and the Louisiana Purchase Zoo for providing some of the gibbon samples used in this study, as well as the Gibbon Genome Sequencing and Analysis Consortium for providing the *N. leucogenys* genome sequence. The whole genome shotgun sequencing of the Northern white-cheeked gibbon was performed at the Washington University Genome Sequencing Center and the Baylor College of Medicine Human Genome Sequencing Center. They also thank J.A. Walker and M.K. Konkel for their invaluable help throughout this project. This work was supported by National Institutes of Health grant RO1 GM59290 to M.A.B.

## References

- Batzler MA, Deininger PL. 1991. A human-specific subfamily of *Alu* sequences. *Genomics* 9:481–487.
- Batzler MA, Deininger PL. 2002. *Alu* repeats and human genomic diversity. *Nat Rev Genet.* 3:370–379.
- Batzler MA, Stoneking M, Alegria-Hartman M, et al. (11 co-authors). 1994. African origin of human-specific polymorphic *Alu* insertions. *Proc Natl Acad Sci U S A.* 91:12288–12292.
- Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet.* 12: 187–215.
- Belancio VP, Deininger PL, Roy-Engel AM. 2009. LINE dancing in the human genome: transposable elements and disease. *Genome Med.* 1:97.
- Belancio VP, Roy-Engel AM, Deininger PL. 2010. All y'all need to know 'bout retroelements in cancer. *Semin Cancer Biol.* 20:200–210.
- Callinan PA, Batzler MA. 2006. Retrotransposable elements and human disease. *Genome Dyn.* 1:104–115.
- Carbone L, Harris RA, Vessere GM, et al. (12 co-authors). 2009. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS Genet.* 5: e1000538.



- Chan YC, Roos C, Inoue-Murayama M, Inoue E, Shih CC, Pei KJ, Vigilant L. 2010. Mitochondrial genome sequences effectively reveal the phylogeny of *Hylobates* gibbons. *PLoS One* 5:e14419.
- Chatterjee HJ. 2006. Phylogeny and biogeography of gibbons: a dispersal-variance analysis. *Int J Primatol* 27:699–712.
- Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
- Churakov G, Grundmann N, Kuritzin A, Brosius J, Makalowski W, Schmitz J. 2010. A novel web-based TinT application and the chronology of the Primate *Alu* retroposon activity. *BMC Evol Biol* 10:376.
- Churakov G, Kriegs JO, Baertsch R, Zemmann A, Brosius J, Schmitz J. 2009. Mosaic retroposon insertion patterns in placental mammals. *Genome Res* 19:868–875.
- Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* 10:691–703.
- Cordaux R, Lee J, Dinoso L, Batzer MA. 2006. Recently integrated *Alu* retrotransposons are essentially neutral residents of the human genome. *Gene* 373:138–144.
- Cordaux R, Srikanta D, Lee J, Stoneking M, Batzer MA. 2007. In search of polymorphic *Alu* insertions with restricted geographic distributions. *Genomics* 90:154–158.
- Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. *Embo J* 21:5899–5910.
- Creel N, Preuschoft H. 1984. Systematics of the lesser apes: a quantitative taxonomic analysis of craniometric and other variables. In: Preuschoft H, Chivers DJ, Broekelman WY, Creel N, editors. *The lesser apes: evolutionary and behavioural biology*. Edinburgh (United Kingdom): Edinburgh University Press. p. 562–613.
- Deininger PL, Batzer MA. 2002. Mammalian retroelements. *Genome Res* 12:1455–1465.
- Garza JC, Woodruff DS. 1992. A phylogenetic study of the gibbons (*Hylobates*) using DNA obtained noninvasively from hair. *Mol Phylogenet Evol* 1:202–210.
- Geissmann T. 1995. Gibbon systematics and species identification. *Int Zoo News* 42:467–501.
- Gibbs RA, Rogers J, Katze MG, et al. (176 co-authors). 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.
- Groves C. 2001. *Primate taxonomy*. Washington (DC): Smithsonian Press.
- Haimoff EH, Chivers DJ, Gittins SP, Whitten T. 1982. A phylogeny of gibbons (*Hylobates* spp.) based on morphological and behavioural characters. *Folia Primatol (Basel)* 39:213–237.
- Hamdi H, Nishio H, Zielinski R, Dugaiczak A. 1999. Origin and phylogenetic distribution of *Alu* DNA repeats: irreversible events in the evolution of primates. *J Mol Biol* 289:861–871.
- Han K, Konkel MK, Xing J, et al. (15 co-authors). 2007. Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. *Science* 316:238–240.
- Herke SW, Xing J, Ray DA, Zimmerman JW, Cordaux R, Batzer MA. 2007. A SINE-based dichotomous key for primate identification. *Gene* 390:39–51.
- Hubbard TJ, Aken BL, Ayling S, et al. (58 co-authors). 2009. Ensembl 2009. *Nucleic Acids Res* 37:D690–697.
- Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16:418–420.
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* 12:656–664.
- Kim SK, Carbone L, Becquet C, Mootnick AR, Li DJ, de Jong PJ, Wall JD. 2011. Patterns of genetic variation within and between gibbon species. *Mol Biol Evol* 28:2211–2218.
- Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* 20:211–221.
- Konkel MK, Walker JA, Batzer MA. 2010. LINEs and SINEs of primate evolution. *Evol Anthropol* 19:236–249.
- Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J. 2007. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet* 23:158–161.
- Lander ES, Linton LM, Birren B, et al. (249 co-authors). 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
- Li J, Han K, Xing J, Kim HS, Rogers J, Ryder OA, Disotell T, Yue B, Batzer MA. 2009. Phylogeny of the macaques (Cercopithecidae: *Macaca*) based on *Alu* elements. *Gene* 448:242–249.
- Locke DP, Hillier LW, Warren WC, et al. (102 co-authors). 2011. Comparative and demographic analysis of orangutan genomes. *Nature* 469:529–533.
- Luan DD, Eickbush TH. 1995. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol Cell Biol* 15:3882–3891.
- Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72:595–605.
- Matsudaira K, Ishida T. 2010. Phylogenetic relationships and divergence dates of the whole mitochondrial genome sequences among three gibbon genera. *Mol Phylogenet Evol* 55:454–459.
- Minghetti PP, Dugaiczak A. 1993. The emergence of new DNA repeats and the divergence of primates. *Proc Natl Acad Sci U S A* 90:1872–1876.
- Mootnick AR. 2006. Gibbon (Hylobatidae) species identification recommended for rescue or breeding centers. *Primate Conservation* 21:103–138.
- Muller S, Hollatz M, Wienberg J. 2003. Chromosomal phylogeny and evolution of gibbons (Hylobatidae). *Hum Genet* 113:493–501.
- Murata S, Takasaki N, Saitoh M, Okada N. 1993. Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution. *Proc Natl Acad Sci U S A* 90:6995–6999.
- Myers RH, Shafer DA. 1979. Hybrid ape offspring of a mating of gibbon and siamang. *Science* 205:308–310.
- Nasidze I, Risch GM, Robichaux M, Sherry ST, Batzer MA, Stoneking M. 2001. *Alu* insertion polymorphisms and the genetic structure of human populations from the Caucasus. *Eur J Hum Genet* 9:267–272.
- Nikaido M, Hamilton H, Makino H, Sasaki T, Takahashi K, Goto M, Kanda N, Pastene LA, Okada N. 2006. Baleen whale phylogeny and a past extensive radiation event revealed by SINE insertion analysis. *Mol Biol Evol* 23:866–873.
- Nilsson MA, Churakov G, Sommer M, Tran NV, Zemmann A, Brosius J, Schmitz J. 2010. Tracking marsupial evolution using archaic genomic retroposon insertions. *PLoS Biol* 8:e1000436.
- Novick GE, Gonzalez T, Garrison J, Novick CC, Batzer MA, Deininger PL, Herrera RJ. 1993. The use of polymorphic *Alu* insertions in human DNA fingerprinting. *Exs* 67:283–291.
- Okada N. 1991. SINEs. *Curr Opin Genet Dev* 1:498–504.

- Okada N, Shedlock AM, Nikaido M. 2004. Retroposon mapping in molecular systematics. *Methods Mol Biol.* 260:189–226.
- Osterholz M, Vermeer J, Walter L, Roos C. 2008. A PCR-based marker to simply identify *Saimiri sciureus* and *S. boliviensis boliviensis*. *Am J Primatol.* 70:1177–1180.
- Osterholz M, Walter L, Roos C. 2008. Phylogenetic position of the langur genera *Semnopithecus* and *Trachypithecus* among Asian colobines, and genus affiliations of their species groups. *BMC Evol Biol.* 8:58.
- Osterholz M, Walter L, Roos C. 2009. Retropositional events consolidate the branching order among New World monkey genera. *Mol Phylogenet Evol.* 50:507–513.
- Perelman P, Johnson WE, Roos C, et al. (14 co-authors). 2011. A molecular phylogeny of living primates. *PLoS Genet.* 7:e1001342.
- Ray DA. 2007. SINEs of progress: mobile element applications to molecular ecology. *Mol Ecol.* 16:19–33.
- Ray DA, Walker JA, Batzer MA. 2007. Mobile element-based forensic genomics. *Mutat Res.* 616:24–33.
- Ray DA, Xing J, Hedges DJ, et al. (13 co-authors). 2005. *Alu* insertion loci and platyrrhine primate phylogeny. *Mol Phylogenet Evol.* 35:117–126.
- Ray DA, Xing J, Salem AH, Batzer MA. 2006. SINEs of a nearly perfect character. *Syst Biol.* 55:928–935.
- Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G. 2002. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res.* 12:602–612.
- Roos C, Geissmann T. 2001. Molecular phylogeny of the major hylobatid divisions. *Mol Phylogenet Evol.* 19:486–494.
- Roos C, Schmitz J, Zischler H. 2004. Primate jumping genes elucidate strepsirrhine phylogeny. *Proc Natl Acad Sci U S A.* 101:10650–10654.
- Roos C, Zinner D, Kubatko LS, et al. (16 co-authors). 2011. Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. *BMC Evol Biol.* 11:77.
- Salem AH, Ray DA, Xing J, Callinan PA, Myers JS, Hedges DJ, Garber RK, Witherspoon DJ, Jorde LB, Batzer MA. 2003. *Alu* elements and hominid phylogenetics. *Proc Natl Acad Sci U S A.* 100:12787–12791.
- Schmid CW. 2003. *Alu*: a parasite's parasite? *Nat Genet.* 35:15–16.
- Schmitz J, Ohme M, Zischler H. 2001. SINE insertions in cladistic analyses and the phylogenetic affiliations of *Tarsius bancanus* to other primates. *Genetics* 157:777–784.
- Schmitz J, Roos C, Zischler H. 2005. Primate phylogeny: molecular evidence from retroposons. *Cytogenet Genome Res.* 108:26–37.
- Shedlock AM, Okada N. 2000. SINE insertions: powerful tools for molecular systematics. *Bioessays* 22:148–160.
- Shedlock AM, Takahashi K, Okada N. 2004. SINEs of speciation: tracking lineages with retroposons. *Trends Ecol Evol.* 19:545–553.
- Smit AFA, Hubley R, Green P. 1996–2010. RepeatMasker Open-3.0. [cited 2012 Jun 25]. Available from: <http://www.repeatmasker.org>
- Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA. 1997. *Alu* insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* 7:1061–1071.
- Suh A, Paus M, Kieffmann M, Churakov G, Franke FA, Brosius J, Kriegs JO, Schmitz J. 2011. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat Commun.* 2:443.
- Swofford DL. 2000. PAUP: phylogenetic analysis using parsimony. Sunderland (MA): Sinauer Associates Inc.
- Takacs Z, Morales JC, Geissmann T, Melnick DJ. 2005. A complete species-level phylogeny of the Hylobatidae based on mitochondrial ND3-ND4 gene sequences. *Mol Phylogenet Evol.* 36:456–467.
- Thinh VN, Mootnick AR, Geissmann T, Li M, Ziegler T, Agil M, Moisson P, Nadler T, Walter L, Roos C. 2010a. Mitochondrial evidence for multiple radiations in the evolutionary history of small apes. *BMC Evol Biol.* 10:74.
- Thinh VN, Mootnick AR, Thanh VN, Nadler T, Roos C. 2010b. A new species of crested gibbon, from the central Annamite mountain range. *Vietn J Primatol.* 1:1–12.
- Thinh VN, Rawson B, Hallam C, Kenyon M, Nadler T, Walter L, Roos C. 2010c. Phylogeny and distribution of crested gibbons (genus *Nomascus*) based on mitochondrial cytochrome b gene sequence data. *Am J Primatol.* 72:1047–1054.
- Waddell PJ, Kishino H, Ota R. 2001. A phylogenetic foundation for comparative mammalian genomics. *Genome Inform Ser Workshop Genome Inform.* 12:141–154.
- Watkins WS, Rogers AR, Ostler CT, et al. (14 co-authors). 2003. Genetic variation among world populations: inferences from 100 *Alu* insertion polymorphisms. *Genome Res.* 13:1607–1618.
- Witherspoon DJ, Marchani EE, Watkins WS, et al. (13 co-authors). 2006. Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and *Alu* insertions. *Hum Hered.* 62:30–46.
- Xing J, Wang H, Han K, Ray DA, Huang CH, Chernick LG, Stewart CB, Disotell TR, Ryder OA, Batzer MA. 2005. A mobile element based phylogeny of Old World monkeys. *Mol Phylogenet Evol.* 37:872–880.
- Xing J, Wang H, Zhang Y, Ray DA, Tosi AJ, Disotell TR, Batzer MA. 2007a. A mobile element-based evolutionary history of guenons (tribe Cercopithecini). *BMC Biol.* 5:5.
- Xing J, Witherspoon DJ, Ray DA, Batzer MA, Jorde LB. 2007b. Mobile DNA elements in primate and human evolution. *Am J Phys Anthropol Suppl.* 45:2–19.
- Zietkiewicz E, Richer C, Labuda D. 1999. Phylogenetic affinities of tarsier in the context of primate *Alu* repeats. *Mol Phylogenet Evol.* 11:77–83.