Evolution of Alu Subfamily Structure in the *Saimiri* Lineage of New World Monkeys

Jasmine N. Baker[1], Jerilyn A. Walker[1], John A. Vanchiere[2], Kacie R. Phillippe[1], Corey P. St. Romain[1], Paulina Gonzalez-Quiroga[1], Michael W. Denham[1], Jackson R. Mierl[1], Miriam K. Konkel[1,3], and Mark A. Batzer[1,*]

[1]Department of Biological Sciences, Louisiana State University, Baton Rouge

[2]Department of Microbiology and Immunology, Louisiana State University Health Sciences Center, Shreveport

[3]Department of Biological Sciences, Clemson University, South Carolina

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

Accepted: August 31, 2017

Abstract

Squirrel monkeys, *Saimiri*, are commonly found in zoological parks and used in biomedical research. S. *boliviensis* is the most common species for research; however, there is little information about genome evolution within this primate lineage. Here, we reconstruct the Alu element sequence amplification and evolution in the genus *Saimiri* at the time of divergence within the family Cebidae lineage. Alu elements are the most successful SINE (Short Interspersed Element) in primates. Here, we report 46 *Saimiri* lineage specific Alu subfamilies. Retrotransposition activity involved subfamilies related to AluS, AluTa10, and AluTa15. Many subfamilies are simultaneously active within the *Saimiri* lineage, a finding which supports the stealth model of Alu amplification. We also report a high resolution analysis of Alu subfamilies within the S. *boliviensis* genome [saiBol1].

Key words: retrotransposon, Alu, Saimiri, subfamilies, evolution.

Introduction

Alu elements are ~300 base pair (bp) primate specific retrotransposons. (Batzer and Deininger 2002; Konkel et al. 2010; Deininger 2011). They are composed of two G–C rich 7SL RNA derived subunits connected via an a-rich linker region and ending in an a-rich tail (Jurka and Zuckerkandl 1991; Deininger 2011). Over time, Alu elements accrue diagnostic mutations that can be used for subfamily classification (Shedlock et al. 2004; Ray 2007) along with random mutations indicative of the age of the elements (Xing et al. 2004). Lineages where particular Alu subfamilies have been active show distinct insertion patterns within the genome. Some insertions are shared with closely related taxa, whereas, others will be more taxon specific. The presence of the same element in multiple taxa provides evidence of an integration event that took place in a common ancestor. Confirmation of phylogenetically significant integrations is possible because Alu elements are nonautonomous and must use the machinery of L1 elements to retrotranspose via Target Primed Reverse Transcription (TPRT) (Luan and Eickbush 1995; Dewannieux et al. 2003; Deininger 2011). By using TPRT, Alu integrations can be easily determined by the presence of element specific Target Site Duplications (TSDs) that are formed during their integration (Luan et al. 1993; Feng et al. 1996). Alu elements are nearly homoplasy free and unidirectional genetic characters with a known ancestral state, hence making them distinct phylogenetic markers (Batzer et al. 1994; Ray 2007).

The New World Monkey (NWM) lineage is one of the most studied and debated primate groups over the last 40 years and Alu elements have been helpful with understanding some phylogenetic relationships between species (Baba et al. 1979; Schneider 2000; Schrago and Russo 2003; Singer et al. 2003; Steiper and Ruvolo 2003; Ray et al. 2005; Bond et al. 2015; Kay 2015). Due to relatively poor fossil records of New World Monkeys it is hard to determine specific divergence times and precisely decipher speciation events. The traditional classification of New World Monkeys had two families: Callitrichidae and Cebidae, using morphology based taxonomy. In 2009, Osterholz (Osterholz et al. 2009) confirmed the monophyly of three families Cebidae, Atelidae, and Pitheciidae using Alu elements as cladistic markers.

Classification of the NWM phylogeny has since expanded to the acceptance of three families Cebidae (small bodies with claws), Atelidae (large fruit and leaf eating monkeys with

prehensile tails) and Pitheciidae (specialized seed predators) (Schneider and Sampaio 2015). There have been various studies conducted to determine the classification and divergence times of NWM, specifically the Cebidae family (Goodman et al. 1998; Schneider 2000; Steiper and Ruvolo 2003; Schrago 2007; Perez et al. 2013; Kay 2015). All of these studies have provided informative trees to show the development of species relationships. A fairly recent study on the first primate fossil found on a North American landmass estimates the minimum age of a split between two of the Cebidae subfamilies, Callitrichinae (marmosets and tamarins), and Cebinae to be about 20.77–21.90 Ma. The estimated divergence of Cebidae from Atelidae is 21.84–24.93 Ma. Despite some differences, it can be agreed upon there was a quick radiation of the ancestors of *Aotus*, *Saimiri*, Cebus and modern Callitrichine (marmosets and tamarins). Even though there seems to be a general consensus on the families of the New World Monkeys there is still some disagreement on subfamily structure within the NWM lineage. Results from studies varied based on the type of molecular test or marker. However, since the family Cebidae contains two species with annotated genomes, common marmoset (*Callithrix jacchus*) and squirrel monkey (*Saimiri bolivienis*) and two genera with scaffold genomes, Cebus (*Cebus capucinus*) and owl monkey (*Aotus nancymaae*) analyzing Alu insertion polymorphisms may be a promising route to resolve these lineages based on results of previous studies using Alu elements as markers (Shedlock et al. 2004; Ray et al. 2005; Konkel et al. 2010).

Various studies using Alu elements as genetic markers have tried to elucidate the NWM phylogeny by focusing on younger polymorphic elements. AluJ elements are the oldest Alu subfamily and found in all primate genomes, AluS are mainly found in anthropoid primates and AluTa are only found in Platyrrhine primates (NWMs) (Ray 2007; Konkel et al. 2010; Schmitz et al. 2016). Singer 2003 (Singer et al. 2003) provided evidence supporting platyrrhine monophyly and consecutive branching events in the callitrichine phylogeny. In 2005 Ray and Batzer reported (Ray et al. 2005), New World Monkey specific subfamilies. Those subfamilies were AluTa7, AluTa10, and AluTa15. AluTa elements are thought to have derived ~15 Ma from a gene conversion event of two ancestral AluS subfamilies (Alu Sc- and Alu Sp-). In 2008, Osterholz (Osterholz et al. 2008) identified a pattern to determine geographic origin with Alu insertions between two species of squirrel monkeys, *S. sciureus* and *S. boliviensis*. With the recent release of the common marmoset [calJac3] and Bolivian squirrel monkey [saiBol1] genomes it is now possible to take a more in depth look at these neotropical primates.

Squirrel monkeys, genus *Saimiri*, are one of the best known neotropical primates and the second most commonly used laboratory monkey (Kinzey 1997). The squirrel monkey is thought to have diverged 1.5 Ma (Chiou et al. 2011) in the NWM lineage. The two most well-known species of squirrel monkey were first named by Hershokovitz in 1984 (Hershkovitz 1984). He grouped the genus into the Roman type which contained *Saimiri boliviensis* and the Gothic type which contained *Saimiri sciureus*. *Saimiri boliviensis* is located in the upper Amazonian and *Saimiri sciureus* is distributed across tropical South America (the south bank Amazonia Rios Purus and Xingu, Pacific coastal area near Cost Rica and Panama—much of this geographical area overlaps *Saimiri boliviensis*) (Hershkovitz 1984; Coe and Rosenblum 1985; Alfaro et al. 2015). The current squirrel monkey genome [saiBol1] is roman type/the Bolivian squirrel monkey. Currently, the subfamily evolution of Alu elements of *Saimiri* is unknown and there are few Alu genetic markers used to determine inter species relationships of *Saimiri*. The common marmoset genome [calJac3] was the first NWM draft assembly that was analyzed and Alu subfamily evolution was reconstructed leading to the common marmoset (Worley et al. 2014). The marmoset genome contained ~1.1 million Alu elements and of those elements ~660,000 were full length. 87 new Alu element subfamilies were identified. The youngest Alu subfamilies were derived from derivatives of AluTa15 (Worley et al. 2014).

The goal of this study was to identify patterns of Alu subfamily amplification and evolution in the genus *Saimiri* after their divergence from the rest of Cebidae. Understanding Alu subfamily evolution in *Saimiri* would help resolve years of questions about New World Monkey relationships. We aimed to use Alu insertion polymorphisms from the squirrel monkey genome to determine the historical relationship of the squirrel monkey lineage.

Materials and Methods

A summary of the methods can be found in figure 1. A data set of full length Alu elements from the *Saimiri* genome [saiBol1] was generated by using the Blat Table Browser (Donna Karolchik et al. 2004). Full length elements are described as beginning within 4 bp of its prospective consensus sequence and being ≥267 bp. All full length Alu elements were extracted from the genome using the table browser with ~600 bp of 5′ and 3′ flanking unique DNA sequences. A total of 739,636 full length Alu elements were extracted from the genome. Data were analyzed using wet bench and computational techniques.

Species Comparison

The data were compared against Human [hg38], Marmoset [calJac3] and *Aotus* [Anan_1.0] genomes using the Blast Like Alignment Tool (BLAT) (Kent 2002). The human and marmoset genome were obtained from the UCSC Genome Browser and the *Aotus* genome was obtained from NCBI in scaffold format (BCM-HGSC 2015). Duplicate elements were removed from the Blat output file leaving the result per locus with the highest Smith–Waterman score. The Blat output from each run was further analyzed using a custom python script to

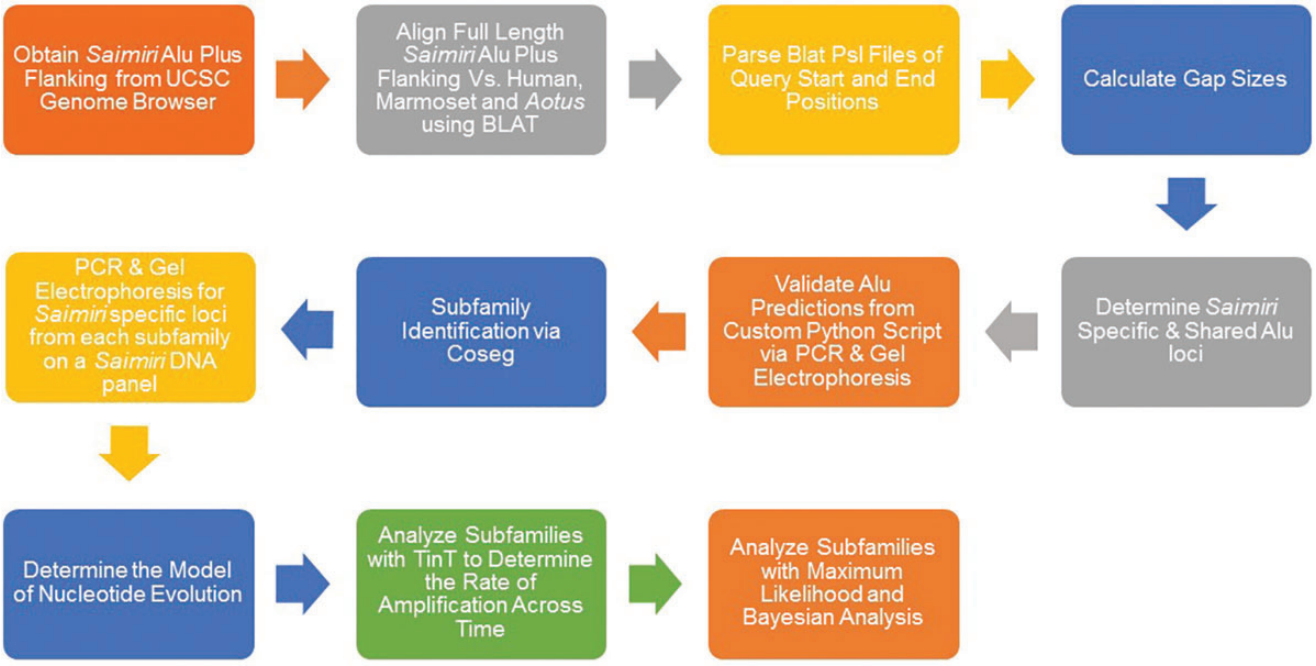

Fig. 1.—Methods Flowchart. The flow chart entails the computational and wet bench methods used to analyze Alu elements from the *Saimiri* reference genome.

determine if insertions were shared or unique among species in the following manner. The python script reads the Blat psl file. The script checks for the starting locations of query and target gap starts and sizes. If all query and target genome gap sizes were smaller than 150 and 250 bp, respectively, then the insertion was shared among the analyzed species. If query gap size of interest was within 30 bp of 300 bp and all target gap sizes $<$250 bp, then the locus was identified as squirrel monkey specific. Last, a custom python script was used to determine how many specific sequences were high quality. High quality is defined here as having a full length Alu element with at least 150 bp of flanking sequence on both the 5′ and 3′ side of the element. These parameters were selected to ensure the Alu element of interest was being identified and that there would be enough shared regions between the target and query to design unique sequence oligonucleotide primers for polymerase chain reaction (PCR) amplification of the locus.

PCR Analysis

PCR amplification was performed in 25 μl reactions that contained 25–50 ng of template DNA, 200 nM of each primer, 1.5 mM $MgCl_2$, 10$\times$ PCR buffer, 0.2 mM deoxyribonucleotide triphosphates and 1 unit of *Taq* DNA polymerase. The PCR protocol is as follows: 95 °C for 1 min, 32 cycles of denaturation at 94 °C for 30 s, 30 s at the respective annealing temperature, and extension at 72 °C for 30 s, followed by a final extension step at 72 °C for 2 min. Gel electrophoresis was performed on a 2% agarose gel containing 0.2 μg/ml ethidium bromide for 60 min at 175 V. UV fluorescence was used to visualize the DNA fragments using a BioRad ChemiDoc XRS imaging system (Hercules, CA).

Subfamily Identification

Alu insertions determined to be *Saimiri* specific were aligned via Crossmatch (www.phrap.org/phredphrapconsed.html#block_phrap; last accessed July 2016) with the following settings: -gap_init -25, gap_ext -5, -min score 200, -min-match 6 –alignments –bandwidth 50 then analyzed via COSEG (www.repeatmasker.org/COSEGDownload.html; last accessed July 2016) to determine subfamily structure. The *Saimiri* specific data set was aligned against the AluS consensus sequence (Jurka and Zuckerkandl 1991). COSEG was then used to group Alu subfamilies. The middle A-rich region of the AluS consensus sequence was excluded from analysis when determining subfamilies, whereas tri and di segregating mutations were considered. Using these criteria, a group of ten or more identical sequences was considered a separate Alu subfamily. A network analysis of identified Alu subfamilies was created by inputting an excel file (supplementary file 4, Supplementary Material online) containing the source, target and weight into Gephi 0.9.1 (Bastian et al. 2009). Source refers to the parent nodes and target refers to nodes branching from a source. Weight refers to the differences between a source and target node and is reflected in the thickness of branches. The excel file was uploaded under "edges" and "create missing nodes" was selected. Once the data was imported, label names were created for each node and sizes were entered to reflect

the subfamily size indicated by COSEG. The layouts used were Fruchterman Reingold and Force Atlas2 (prevent overlap and scaling was changed to 50). Under "appearance", node size and edge/branch size was adjusted to reflect the numbers entered for size and weight.

Next, a custom RepeatMasker library was created containing the Alu consensus sequences from the *Saimiri* lineage specific subfamilies, as well as those identified in the Marmoset Consortium project and previously known Alu subfamilies was created to verify the presence of each subfamily in the data set. Then, using this custom library, we used RepeatMasker (Smit et al 2013-2015) to identify subfamilies unique to the full length Alu data set of squirrel monkey. The data set was repeat masked with the following settings: -s –nolow –lib libraryfilename –no_is dataset.fasta. Subfamilies with young Alu elements (2% or less diverged from the consensus sequence) were analyzed using polymerase chain reaction within a sample size of 32 squirrel monkeys (supplementary file 5, Supplementary Material online). The number of loci assessed from each subfamily ranged from 1 to 5 depending on how many loci were young.

Model Selection

All subfamilies identified were analyzed with jModelTest-2.17 (Darriba et al. 2012) to determine the best model of nucleotide evolution for the data set. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) models were a gamma distribution with an alpha value of 0.3770 and 0.3920, respectively. Since alpha was <1, it means the distribution has a highly skewed L-shape, and most sites have very low rates or are nearly "invariable", but there are some substitution hot spots with high rates of variation (Yang 2014). The AIC model selected was TrN + I+G(variable base frequencies, equal transversion rates, variable transition rates, unchanging sites, and gamma distributed rate variation among sites) and the BIC model selected was TrN + G(variable base frequencies, equal transversion rates, variable transition rates, and gamma distributed rate variation among sites). According to the BIC model selected, transitions rAG = 3.9349 and rCT = 8.3258 (rAG = rate of change from A to G and rCT = rate of change from C to T) which aligns with previous studies that CpG sites in Alu elements have six to ten times faster mutation rates than non CpG sites (Labuda and Striker 1989; Batzer et al. 1990; Xing et al. 2004).

Maximum Likelihood and Bayesian Analysis

An alignment and nexus file of 108 subfamilies identified in the *Saimiri* lineage was created with MEGA6 (Tamura et al. 2013). Garli-2.01 was used to infer a Maximum Likelihood tree for the data using the TrN + I + G model of DNA evolution and 10,000 bootstraps (Zwickl 2014). The resulting trees were subsequently analyzed using Sumtrees to produce a majority-rules consensus (Zwickl 2014; Sukumaran and Holder 2010; Holder 2015). BEAST software (Bayesian Evolutionary Analysis Sampling Trees) (Drummond et al 2012) was used for Bayesian analysis. The following settings were changed from the default settings: site heterogeneity = gamma, species tree prior = birth death process, and nucleotide model = TrN. The length of the chain was 30 million.

Transposition in Transposition Analysis

We estimated the chronological order of Alu element accumulation using the transposition in transposition (TinT) method (Kriegs et al. 2007; Churakov et al 2010) (http://www.compgen.unimuester.de/tools/tint/ last accessed July 13, 2017; default parameters for SINE elements) on all full length elements from the squirrel monkey genome. The resulting graph can be found in supplementary file 1, Supplementary Material online.

Results

We examined a total of 1,017,126 Alu elements in the squirrel monkey genome based on a Blat Table Browser search. 739,636 full length elements were identified.

Alu Prediction Validation

Approximately 80 of these loci were amplified by polymerase chain reaction to ensure accuracy of the detection program. The DNA panel used in these experiments and the results are shown in supplementary file 2, Supplementary Material online. Of the 739,636 full length Alu element insertions, 43,201 were determined to be specific to *Saimiri*. Of these *Saimiri* specific loci, a separate set of ~60 were randomly chosen to experimentally verify their *Saimiri* lineage specificity (fig. 2) along with the locus reported by (Osterholz et al. 2008) using the DNA panel described in supplementary file 2, Supplementary Material online.

Alu Subfamily Inference

Analysis with custom python scripts resulted in 41,782 high quality orthologous loci. Pairwise alignments were then completed using Crossmatch. 149 sequences were filtered due to poor quality alignments leaving 41,633 sequences to analyze in COSEG. Forty-six Alu subfamilies were identified (fig. 3). All 46 consensus sequences (supplementary file 8, Supplementary Material online) were aligned and analyzed to see if there were similarities to previously defined Alu subfamilies. Our results suggest, there were four major bursts of Alu element amplification in the squirrel monkey lineage.

Alu Subfamily Evolution

Major bursts, in approximate sequential order based on the network analysis (fig. 3), divergence calculations (fig. 5), phylogenetic analysis (figs. 6 and 7), and TinT (supplementary fig.

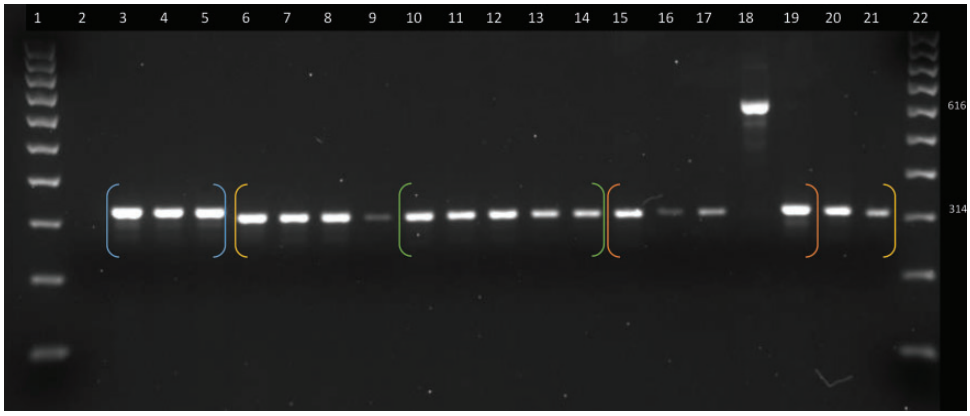

Fig. 2.—Squirrel Monkey Specific Alu Element. The presence of the Alu element (genomic location: JH378127: 2656603-2658120) is indicated by the ~ 616 bp band (lane 18) and the absence/empty site by the ~ 314 bp bands. Lanes: 1-100 bp ladder, 2-TLE (negative control), 3-Human (HeLa), 4-Chimpanzee, 5-African Green Monkey, 6-Wooly Monkey, 7-White Bellied Spider Monkey, 8-Black-handed Spider Monkey, 9-Bolivian Red Howler Monkey, 10-Common Marmoset, 11-Pygmy Marmoset, 12-Goeldi's Marmoset, 13-Red-chested Mustached Tamarin, 14-Geoffroys Saddle-Back Tamarin, 15-Capuchin Monkey, 16-Capuchin Monkey, 17-Capuchin Monkey, 18- Squirrel Monkey, 19-Owl Monkey, 20-Northern White-Faced Saki, 21-Bolivian Grey Titi, 22–100 bp ladder. Blue brackets= human and old world monkeys (3–5), yellow brackets = new world monkeys (6–21), Atelidae (6–9), green brackets = Callitrichidae (10–14), orange brackets= Cebidae (15–19), and Pitheciidae (20–21).

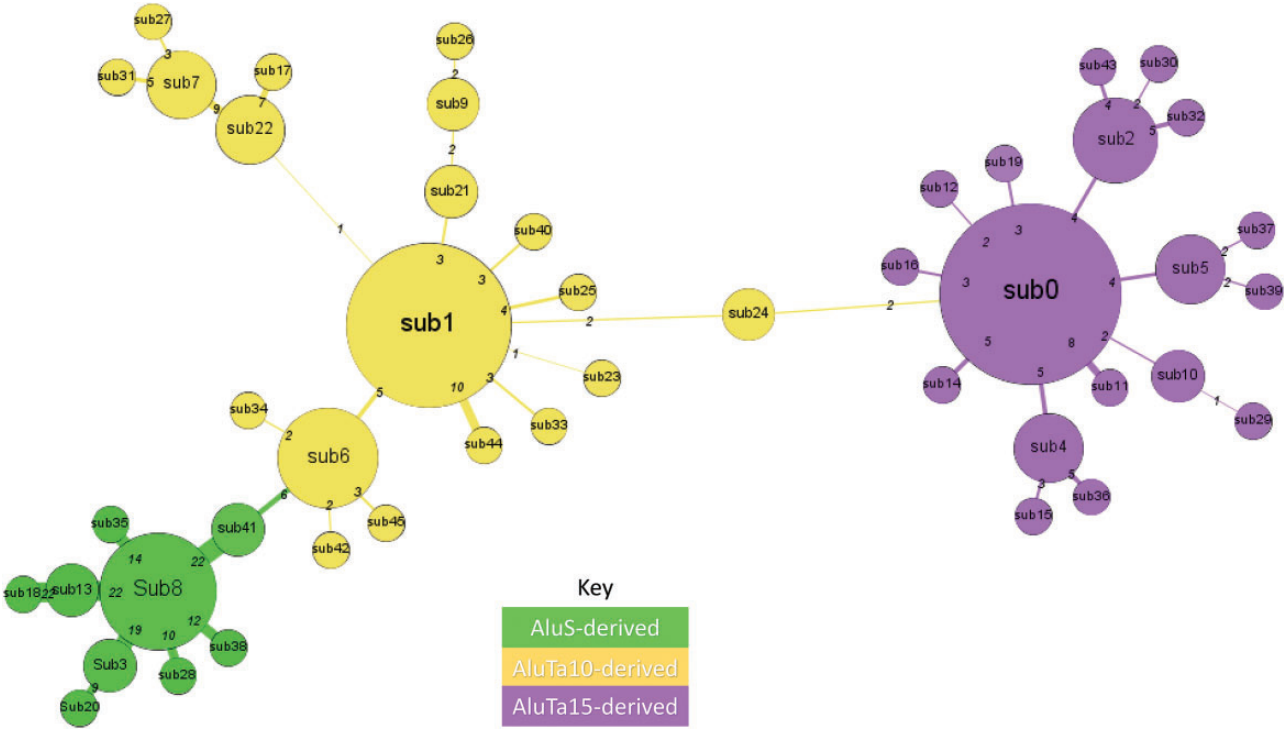

Fig. 3.—Network Analysis of Lineage specific Alu Subfamilies in the genome. This is a network schematic (diagram) of the 46 lineage specific Alu subfamilies identified in [saiBol1] via COSEG and generated in Gephi. The size of each node represents the number of individual Alu elements in each subfamily (larger circles have more members). The thickness of the branch lines correlates to the number of nucleotide substitutions between the source node and target node, also printed at the base of each branch. Green nodes are AluS derived, yellow nodes are AluTa10 derived and purple nodes are AluTa15 derived.

1, Supplementary Material online) were AluS derived subfamilies (i.e., sub8), AluTa10 derived elements (i.e., sub6 and sub1), and AluTa15 derived subfamilies (i.e., sub0). The root and eldest families prior to the first transposition burst are AluS derived (figs. 3 and 4). The first and second burst of Alu elements appear to be AluTa10 derived and the most recent burst appear to be AluTa15 derived. The consensus sequence for each of the 46 lineage specific Alu subfamilies

Fig. 4.—Alignment of Alu Subfamilies prior to the first burst of AluTa10 related subfamilies. The top sequence in the oldest subfamily identified in the squirrel monkey specific data set. The next three sequences are more recent subfamilies and are derivatives of AluS prior to the first burst of AluTa10, illustrating that AluS derived subfamilies played a role in the subsequent Alu amplification in the *Saimiri* lineage. The dots represent sequence identities and the dashes represent deletions. Mutations are denoted with the appropriate nucleotide.

is shown in FASTA format in supplementary file 4, Supplementary Material online. To further investigate subfamily structure of the *Saimiri* lineage, a custom RepeatMasker library was created and the data set was RepeatMasked against the subfamilies identified by COSEG, the subfamilies identified in the Marmoset Consortium project and the Alu Consensus sequences from the RepeatMasker library. This was done to ensure sequences were as closely (matched) identified to its prospective consensus sequence based on the Alu consensus sequences known to be in New World Monkeys.

The majority, 96% (40,198 of 41,633 elements), of the squirrel monkey specific Alu elements were properly identified during the RepeatMasker analysis. All 46 of the newly identified *Saimiri* subfamilies (denoted as Sub in fig. 3) were identified in the data set along with another 46 separate Alu subfamilies identified from the Marmoset Consortium project (denoted as sf). Another 16 subfamilies were classified as AluJ and S subfamilies, resulting in a total of 108 Alu subfamilies identified in the genome [saiBol1]. The subfamily with the highest amplification throughout the squirrel monkey specific loci was sf63 (supplementary file 6, Supplementary Material online). This subfamily seemed to have the greatest retention (27,574 loci in the genome) as well as amplification after speciation accounting for 13,193 squirrel monkey specific loci (fig. 5). Of the loci used for verification, 33 subfamilies were represented (24 newly determined subfamilies, 8 Marmoset Consortium subfamilies, and 1 previously known New World Monkey specific subfamily—AluTa10). The locus from Osterholz was also RepeatMasked and identified as sf63—a subfamily identified during the Marmoset Consortium Project that had high amplification within the squirrel monkey specific loci (fig. 5). This locus is thought to be a possible subspecies molecular marker (Osterholz et al. 2008).

To compare subfamily composition in the lineage specific loci data set against the full length elements of the entire genome and to ensure accurate subfamily identification, all the full length elements of the squirrel monkey genome were RepeatMasked against a custom RepeatMasker library that included the newly identified 46 subfamilies. The original full length Alu element data set contained 739,636 elements and of those 614,652 (83%) were identified to match a prospective consensus sequence via RepeatMasker. The subfamily identification from this run was used in all further analyses. Supplementary file 6, Supplementary Material online displays the breakdown of individual subfamilies that were detected in the genome data set and may have had a decrease in representation in the squirrel monkey specific data set. Last, a TinT analysis (supplementary file 1, Supplementary Material online) was completed with the Alu elements in the reference

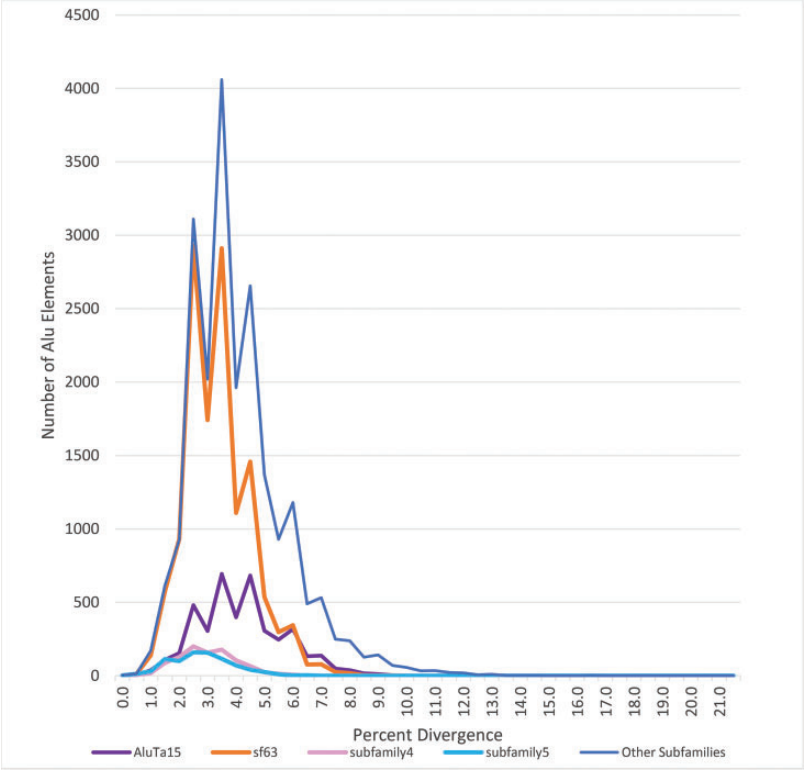

Fig. 5.—Percent divergence of Alu subfamilies with the highest retention in Squirrel Monkey. Percent divergence ranges from 0 to 21.5% diverged from its prospective consensus sequence. Each subfamily is identified by a different color. The figure shows AluTa15, sf63 and subfamily 4 and subfamily 5. The remaining subfamilies identified in the genome are combined into a single category, "Other subfamilies." A full list of subfamilies and their count can be found in supplementary files 6 and 7, Supplementary Material online.

genome to analyze the rate of amplification across time. The activity is similar to the marmoset lineage and other primate lineages bases on the previous TinT analysis of primates in Churakov et al. (2010).

Subfamily Divergence

Subfamily divergence was analyzed via excel (supplementary file 7, Supplementary Material online). A total of 4,174 Alu elements were classified as young (0.0–2.0% diverged). 51 of 108 subfamilies in the *Saimiri* genome contained loci that were 0–2% diverged from their respective consensus sequences (supplementary file 7, Supplementary Material online). 667 of 4,174 are from the 26 lineage specific subfamilies that contained young elements (table 1). Younger elements were overrepresented in AluTa15 derived subfamilies (fig. 5 and supplementary file 6, Supplementary Material online). Figure 5 depicts sf63, subfamily 4 and subfamily 5 and are most likely derived from a common ancestor of AluTa15. 129 young Alu elements from the 51 subfamilies were analyzed by PCR and 68 of the 129 loci were polymorphic (53%). Younger elements are more likely to be polymorphic; however, this rate has the possibility to vary with an increase in sample size.

The tree produced shows branching in relation to all 108 subfamilies identified in the *Saimiri* lineage. The Maximum Likelihood tree (fig. 6) was largely unresolved; however, the Bayesian analysis was able to resolve all branches (fig. 7). Bayesian analysis of the subfamily consensus sequences agreed in large part with the network analysis in figure 3. Based on the tree branching we were able to determine which families were more related to each other and their possible derivatives.

Discussion

The results of this study outline the evolution and amplification of lineage specific Alu subfamilies in the *Saimiri* NWM lineage. Of the 108 Alu subfamilies identified from the squirrel monkey reference genome [saiBol1], 46 appeared *Saimiri* lineage specific in that they evolved after the divergence from the marmoset lineage. Alu subfamilies shared with marmoset were generally less prolific in the *Saimiri* lineage in numbers compared with the squirrel monkey specific genome content (supplementary file 6, Supplementary Material online). Some of the subfamilies identified in the marmoset consortium project were not recovered in this study (supplementary file 6, Supplementary Material online). Since there is currently no

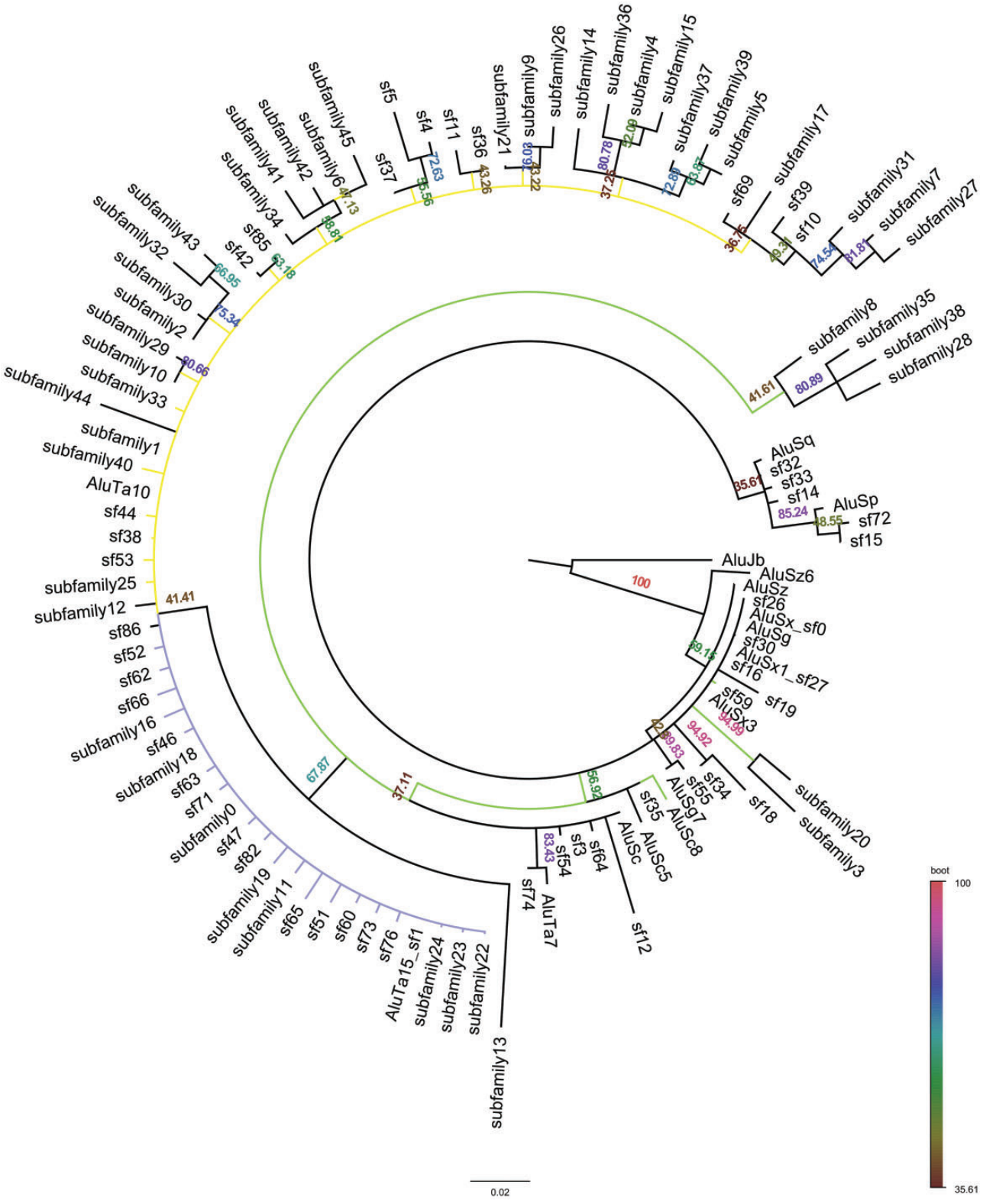

Fig. 6.—Maximum Likelihood Tree of Alu subfamilies identified in the *Saimiri* genome. This tree displays a consensus of subfamily evolution in *Saimiri*. Labels that begin with "sf" are the subfamilies identified during the Marmoset Genome Consortium project (Worley et al. 2014). Labels that begin with "subfamily" are the subfamilies identified during this study. Branches are color coded to match the subfamily bursts depicted in figure 3. Green denotes where AluS derived bursts occurred. Yellow denotes where AluTa10 bursts occurred. Purple denotes where AluTa15 derived bursts occurred. Bootstrap values are listed on each branch.

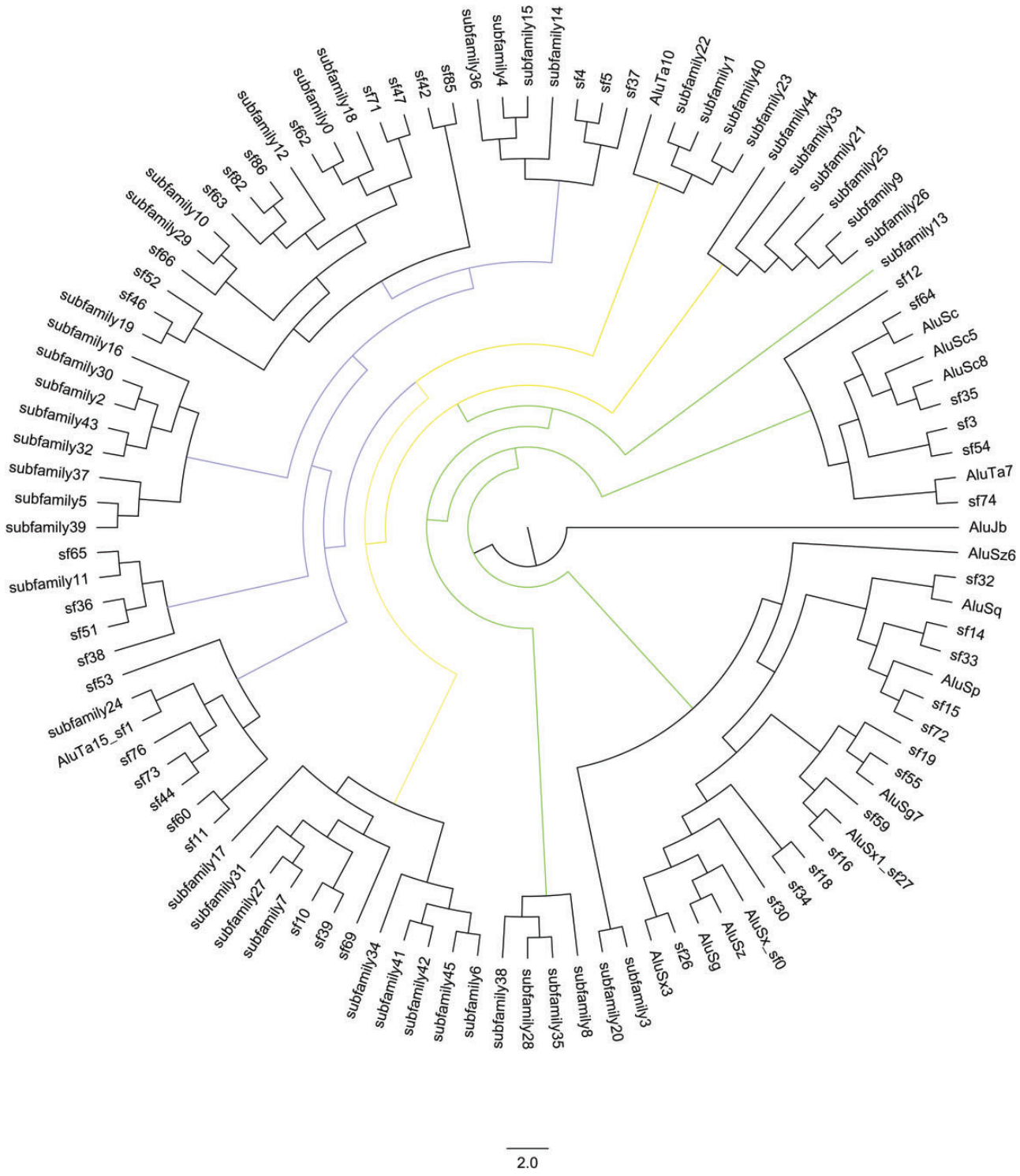

Fig. 7.—Bayesian Tree of Alu subfamilies identified in the *Saimiri* genome. This tree displays a consensus of subfamily evolution in *Saimiri*. Labels that begin with "sf" are the subfamilies identified during the Marmoset Genome Consortium project (Worley et al. 2014). Labels that begin with "subfamily" are the subfamilies identified during this study. Branches are color coded to match the subfamily bursts depicted in Figure 3. Green denotes where AluS derived bursts occurred. Yellow denotes where AluTa10 bursts occurred. Purple denotes where AluTa15 derived bursts occurred.

Table 1
Number of Young Alu Elements in Lineage Specific Subfamilies[a]

	0%	0.50%	1%	1.50%	2%
subfamily 0	0	0	1	3	7
subfamily11	0	0	1	2	5
subfamily12	0	3	2	13	4
subfamily13	0	0	0	1	1
subfamily14	0	0	0	2	5
subfamily15	0	0	0	3	4
subfamily17	0	0	0	0	1
subfamily18	0	0	0	0	1
subfamily2	0	0	0	2	1
subfamily21	0	0	0	1	0
subfamily26	0	1	1	8	7
subfamily27	0	0	0	0	1
subfamily29	0	0	1	2	4
subfamily30	0	0	0	0	1
subfamily32	1	3	4	5	9
subfamily33	0	0	0	1	0
subfamily36	0	0	5	7	9
subfamily37	0	0	1	6	1
subfamily39	0	1	0	2	3
subfamily4	0	6	18	87	128
subfamily40	0	0	0	0	1
subfamily43	0	0	0	6	4
subfamily45	1	0	0	1	1
subfamily5	2	12	30	113	100
subfamily7	0	0	0	0	1
subfamily9	1	0	1	3	4

aThis chart displays the 26 subfamilies discovered in the lineage specific subfamilies that contained young Alu elements. The first column is the subfamily name and following columns contain total number of elements for each % divergence.

known method of precise SINE excision, there may have been some host factors (Bogerd et al., 2006; Zamudio and Bourc'his, 2010; Rowe et al. 2010; Soifer et al., 2005) that limited further amplification in the [saiBol1] genome, whereas other subfamilies identified in the marmoset consortium project may have begun amplification after the squirrel monkey and marmoset split. Alternatively lineage sorting of polymorphic retrotransposition competent source loci after bottlenecks that occurred within the *Saimiri* lineage may have randomly impacted the amplification of certain active subfamilies within the lineage through loss of more active driver loci. There were also less lineage specific subfamilies than reported for the marmoset consortium project. This is important because both species are members of the Cebidae family of new world monkeys, and most likely split ~20 Ma (Perez et al. 2013) which is ~5 Ma before the amplification of New World Monkey specific Alu elements began (Ray 2007). However, as with previous genome studies, gradual refinement of smaller lineage specific subfamilies is expected as updated genome assemblies become available and additional wet bench validations are conducted.

Alu elements are known to propagate in a star like amplification pattern with multiple subfamilies concurrently active (Cordaux et al 2004). This pattern is seen in this data set (fig. 3). Each one of these bursts in amplification can be linked to different time periods of the primate phylogenetic tree. The order of subfamily amplifications were AluS related, AluTa10 and AluTa15 related, respectively. AluS subfamilies initially arose ~35 Ma and AluTa10 and AluTa15 subfamilies arose ~15 Ma (Ray 2007). AluTa10 and AluTa15 are NWM specific Alu subfamilies created via a gene conversion event of Alu Sc and -Sp elements. According to our data, AluS related elements were ancestral to the AluTa elements before the lineage specific bursts in agreement with previous studies (Ray and Batzer 2005; Ray 2007) This is confirmed by the subfamily sequence consensus tree (fig. 7) which shows the most likely relationships between the 108 subfamilies. The branching patterns in figure 7 show groupings consistent with the lineage specific burst patterns illustrated in the network analysis (fig. 3). 51 of those subfamilies were further tested to determine accumulation within the *Saimiri* lineage. 53% of young loci from various subfamilies tested displayed polymorphisms within a panel of 32 squirrel monkeys confirming that multiple subfamilies are still active and propagating within the *Saimiri* lineage.

Before the bursts of AluTa subfamilies, the most ancestral subfamilies consensus had high identity to Alu Sc and Alu Sp consensus sequences (fig. 4). However, there were some AluTa specific mutations. For example, there is a T at bp 57 that is a diagnostic mutation in 45 of 46 lineage specific subfamilies. Also, all AluTa related subfamilies and the first major burst node (subfamily 8) of the S related subfamilies have an adenosine at bp 78 (The B-box promoter); however, the A & B promoter boxes are pristine with perfect sequence for the first two parent nodes of the network analysis of squirrel monkey specific subfamilies. It seems like these particular AluS subfamilies remained active for extended periods of time and daughter subfamilies had a series of bursts in amplification. It is fair to say our data supports the stealth model of Alu mobilization. The stealth model of Alu mobilization states that driver elements maintain low retrotransposition activity over extended periods of time and create short-lived hyperactive copies that promote Alu mobilization in the genome (Han et al. 2005).

The data from this study are informative because we can target particular subfamily amplification roughly with different time periods in the primate lineage as well as to utilize polymorphisms to characterize individual geographic origins within the *Saimiri* lineage. For example, loci of various subfamilies were verified as polymorphic within individuals in the *Saimiri* lineage. Within those loci, the loci from Osterholz were reexamined on our squirrel monkey DNA panel (supplementary file 3, Supplementary Material online). According to Osterholz, this Alu insertion is primarily present in S. *boliviensis* and generally absent from S. *sciureus* and a heterozygous genotype means the squirrel monkey is likely a hybrid. However, with our DNA panel we were able to determine

that the locus was polymorphic in *S. sciureus* and *S. boliviensis* showing various patterns of homozygous and heterozygous individuals within both species. Therefore, this locus may not be as informative as a subspecies indicative molecular marker as originally reported. However, analysis of a larger panel including all squirrel monkey species with known ancestry would be beneficial to properly assay this locus. Considering *S. sciureus* and *S. boliviensis* are sister taxa and the insertion was 6% diverged from its consensus sequence it may not have been fixed within the population prior to *Saimiri* splitting into subspecies ~1.5 Ma (Chiou et al. 2011). With the recent release of the common marmoset [calJac3] and bolivian squirrel monkey [saiBol1] genomes it is now possible to take a more in depth look at these neotropical primates. Identifying more of these loci would be informative for studies involving subspecies of squirrel monkeys or species they may believe to be hybrids. It is important to keep in mind that these sequences were mined from *S. boliviensis*. *S. boliviensis* is sister taxa to *S. sciureus* and sequences would need to be tested between the squirrel monkey clade to determine intra species specificity. However, by the use of computational and wet bench techniques, we have been able to identify *Saimiri* lineage specific subfamilies and model the evolution of subfamily structure within the genome. With new sequencing technology constantly being developed we look forward to seeing how Alu composition can be further defined within the squirrel monkey lineage and other NWMs.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The authors thank all the members of the Batzer Lab, Arundhati Bakshi, Lorelei Patrick and Ana Arcanjo for their helpful suggestions and constructive criticism. The squirrel monkey genome assembly (*Saimiri boliviensis*) is provided with the following acknowledgements: We acknowledge the Broad Institute (Cambridge, MA) for the saiBol1 sequencing and assembly. We also acknowledge Hiram Clawson, Chin Li, Brian Raney, Pauline Fujita, Luvina Guruvadoo, Steve Heitner, Brooke Rhead, Greg Roe, and Donna Karolchik for the UCSC squirrel monkey genome browser/initial annotations. This research was supported by the National Institutes of Health R01 GM59290 (M.A.B). The authors also wish to thank the following people and institutions for their generous donation of samples: Dr. Frederick H. Sheldon, Curator, and Donna Dittmann of the Louisiana State University Museum of Natural Science Collection of Genetic Resources; Michale E. Keeling Center for Comparative Medicine and Research, The University of Texas MD Anderson Cancer Center, Bastrop, TX; San Diego Zoo Global Biomaterials Review Group, San Diego Zoo Institute for Conservation Research; Sharon Birks, Genetics Resources Collections Manager at the Burke Museum of Natural History and Culture, University of Washington; Kristof Zyskowski, Collection Manager at the Peabody Museum of Natural History, Yale University; Christopher C. Conroy, Curator, Mammals Collection at the Museum of Vertebrate Zoology, University of California—Berkeley, and Dr. John A. Vanchiere, Chief, Pediatric Infectious Diseases, Louisiana State University Health Sciences Center—Shreveport.

Literature Cited

Alfaro JW, et al. 2015. Biogeography of squirrel monkeys (genus Saimiri): south-central Amazon origin and rapid pan-Amazonian diversification of a lowland primate. Mol Phylogenet Evol. 82: 436–454.

Baba ML, Darga LL, Goodman M. 1979. Immunodiffusion systematics of the primates. Folia Primatol. 32(3): 207–238.

Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. In: *International AAAI Conference on Weblogs and Social Media*.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. Nat Rev Genet. 3(5): 370–379.

Batzer MA, et al. 1990. Structure and variability of recently inserted Alu family members. Nucleic Acids Res. 18(23): 6793–6798.

Batzer MA, et al. 1994. African origin of human-specific polymorphic Alu insertions. In: *African origin of human-specific polymorphic Alu insertions*.

BCM-HGSC. 2015. Baylor College of Medicine Human Genome Sequencing Center. *Owl Monkey Genome Project*. [Online] 2015. Available from: https://www.hgsc.bcm.edu/non-human-primates/owl-monkey-genome-project, last accessed December 2015.

Bogerd HP, et al. 2006. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proc Natl Acad Sci U S A. 103: 8780–8785.

Bond M, et al. 2015. Eocene primates of South America and the African origins of New World monkeys. Nature 520: 538–541.

Chiou KL, Pozzi L, Lynch Alfaro JW, Di Fiore A. 2011. Pleistocene diversification of living squirrel monkeys (Saimiri spp.) inferred from complete mitochondrial genome sequences. Mol Phylogenet Evol. 59(3): 736–745.

Churakov G, et al. 2010. A novel web based TinT application and the chronology of the primate Alu retroposon activity. BMC Evol Biol. 10: 376.

Coe CL, Rosenblum LA. 1985. Handbook of Squirrel Monkey Research. Boston, MA: Springer. 1–33.

Cordaux R, Hedges DJ, Batzer MA. 2004. Retrotransposition of Alu elements: how many sources?. Trends Genet. 20(10): 464–467.

COSEG. 2016. COSEG Download Page. Available from: www.repeatmasker.org/COSEGDownload.html, last accessed July 2016.

Crossmatch. 2016. Phred, Phrap, Consed Software Page, Available from: www.phrap.org/phredphrapconsed.html#block_phrap, last accessed July 2016.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9(8): 772–772.

Deininger P. 2011. Alu elements: know the SINEs. Genome Biol. 12(12): 236.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 35(1): 41–48.

Donna Karolchik ASH, et al. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32: 493–496.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and BEAST 1.7. Mol Biol Evol. 29(8): 1969–1973.
Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87(5): 905–916.
Goodman M, et al. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol. 9(3): 585–598.
Han K, et al. 2005. Under the genomic radar: the stealth model of Alu amplification. Genome Res. 15(5): 655–664.
Hershkovitz P. 1984. Taxonomy of squirrel monkeys genus Saimiri (Cebidae, Platyrrhini): a preliminary report with description of a hitherto unnamed form. American J Primatol. 7(2): 155–210.
Holder JSaM. 2015. In: SumTrees: Phylogentic Tree Summarization. Available from: https://github.com/jeetsukumaran/DendroPy, last accessed July 2016.
Jurka J, Zuckerkandl E. 1991. Free left arms as precursor molecules in the evolution of Alu sequences. J Mol Evol. 33(1): 49–56.
Kay RF. 2015. Anthropology. New World monkey origins. Science 347(6226): 1068–1069.
Kent WJ. 2002. BLAT: the BLAST-like alignment tool. Genome Rese. 12(4): 656–664.
Kinzey WG. 1997. *New World Primates: Ecology, Evolution, and Behavior*. New York: Transaction Publishers. 299–305.
Konkel MK, Walker JA, Batzer MA. 2010. LINEs and SINEs of primate evolution. Evol Anthropol. 19(6): 236–249.
Kriegs JO, et al. 2007. Waves of genomic hitchhikers shed light on the evolution of gamebirds(Aves: Galliformes). BCM Evol Biol. 7: 190.
Labuda D, Striker G. 1989. Sequence conservation in Alu evolution. Nucleic Acids Res. 17(7): 2477–2491.
Luan DD, Eickbush TH. 1995. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. Mol Cell Biol. 15(7): 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(4): 595–605.
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(12): 1177–1180.
Osterholz M, Walter L, Roos C. 2009. Retropositional events consolidate the branching order among New World monkey genera. Mol Phylogenet Evol. 50(3): 507–513.
Perez SI, Tejedor MF, Novo NM, Aristide L, Colgan DJ. 2013. Divergence times and the evolutionary radiation of New World monkeys (Platyrrhini, Primates): an analysis of fossil and molecular data. PLoS One 8(6): e68029.
Ray DA. 2007. SINEs of progress: mobile element applications to molecular ecology. Mol Ecol. 16(1): 19–33.
Ray DA, Batzer MA. 2005. Tracking Alu evolution in New World primates. BMC Evol Biol. 5: 51.
Ray DA, et al. 2005. Alu insertion loci and platyrrhine primate phylogeny. Mol Phylogenet Evol. 35(1): 117–126.
Rowe HM, et al. 2010. KAP1 controls endogenous retroviruses in embryonic stem cells. Nature 463(7278): 237–240.
Schmitz J, et al. 2016. Genome sequence of the basal haplorrhine primate *Tarsius syrichta* reveals unusual insertions. Nat Commun. 7: 12997.
Schneider H. 2000. The current status of the New World monkey phylogeny. An Acad Bras Cienc. 72(2): 165–172.
Schneider H, Sampaio I. 2015. The systematics and evolution of New World primates: a review. Mol Phylogenet Evol. 82: 348–357.
Schrago CG. 2007. On the time scale of New World primate diversification. Am J Phys Anthropol. 132(3): 344–354.
Schrago CG, Russo CAM. 2003. Timing the origin of New World monkeys. Mol Biol Evol. 20(10): 1620–1625.
Shedlock AM, Takahashi K, Okada N. 2004. SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol. 19(10): 545–553.
Singer SS, Schmitz J, Schwiegk C, Zischler H. 2003. Molecular cladistic markers in New World monkey phylogeny (Platyrrhini, Primates). Mol Phylogenet. Evol. 26(3): 490–501.
Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0.
Soifer HS, Zaragoza A, Peyvan M, Behlke MA, Rossi JJ. 2005. A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. Nucleic Acids Res. 33(3): 846–856.
Steiper ME, Ruvolo M. 2003. New World monkey phylogeny based on X-linked G6PD DNA sequences. Mol Phylogenet Evol. 27(1): 121–130.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26(12): 1569–1571.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar, S 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30(12): 2725–2729.
Worley KC, et al. 2014. The common marmoset genome provides insight into primate biology and evolution. Nat Genet. 46: 850–857.
Xing J, et al. 2004. Alu element mutation spectra: molecular clocks and the effect of DNA methylation. J Mol Biol. 344(3): 675–682.
Yang Z. 2014. Molecular evolution: a statistical approach. Oxford: Oxford Univeristy Press.
Zamudio N, Bourc'his D. 2010. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap[quest]. Heredity 105(1): 92–104.
Zwickl D. 2014. GARLI: genetic algorithm for rapid likelihood inference. 2006. Ph.D. dissertation, The University of Texas at Austin. Available from: http://garli. googlecode. com, last accessed July 2016.

Associate editor: Mar Alba