Analytical Biochemistry 593 (2020) 113516

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Sensitivity of the polyDetect computational pipeline for phylogenetic analyses

Jessica M. Storer, Jerilyn A. Walker, Vallmer E. Jordan, Mark A. Batzer*

Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA, 70803, USA

ARTICLE INFO

Keywords:
Phylogeny
Cebidae
Alu
Computational
polyDetect
NWM

ABSTRACT

Alu elements are powerful phylogenetic markers. The combination of a recently-developed computational pipeline, polyDetect, with high copy number *Alu* insertions has previously been utilized to help resolve the *Papio* baboon phylogeny with high statistical support. Here, the polyDetect method was applied to the highly contentious Cebidae phylogeny within New World monkeys (NWM). The polyDetect method relies on conserved homology/identity of short read sequence data among the species being compared to accurately map predicted shared *Alu* insertions to each unique flanking sequence. The results of this comprehensive assessment indicate that there were insufficient sequence homology/identity stretches in non-repeated DNA sequences among the four Cebidae genera analyzed in this study to make this strategy phylogenetically viable. The ~20 million years of evolutionary divergence of the Cebidae genera has resulted in random sequence decay within the short read data, obscuring potentially orthologous elements in the species tested. These analyses suggest that the polyDetect pipeline is best suited to resolving phylogenies of more recently diverged lineages when high-quality assembled genomes are not available for the taxa of interest.

1. Introduction

New World monkeys (NWM), or Platyrrhines, diverged from Old World monkeys (OWM), or Catarrhines, roughly ~35 million years ago and inhabit the tropical forests of Mexico and Central and South America [1]. The harsh conditions in South America, and in particular the Amazonian region, do not foster ideal conditions for fossil formation. This, in combination with lack of behavioral knowledge compared to OWM, creates a circumstance where the phylogeny of NWM has been highly contentious. An initial analysis of NWM cranial morphology and presumed important speciation characteristics in 1977 by Hershkovitz [2] led to a phylogeny with three families: 1) Callitrichidae, containing four genera, which included tamarins and marmosets; 2) the monotypic Callimiconidae, containing the genus *Callimico*, and 3) Cebidae, containing eleven genera, which included all other NWM types, including capuchin monkeys, owl monkeys, squirrel monkeys, and monkeys that now belong to different NWM families, such as those belonging to the *Pithecia* and *Brachyteles* genera. While this classification was undisputed for a long period of time, it ignored the more difficult placements that include the genera *Cebus, Aotus,* and *Saimiri* [2]. It was not until molecular markers were utilized, instead of morphology that a consensus was reached on the three NWM families: Cebidae (containing the *Cebus, Aotus, Saimiri* genera and the Callitrichidae), Atelidae and Pitheciidae [3]. However, there is not a consensus as to the inter-generic relationships within each family. For example, the position of the *Aotus* genus within Cebidae remains controversial. Although there are shared *Alu* mobile element genetic markers that place *Aotus* squarely within the Cebidae family [4,5], there is conflicting information as to the position of *Aotus* in the Cebidae lineage [4]. This ambiguity is due to the rapid speciation event(s) occurring over 1-2 my leading to the separation of the Cebidae families that could produce incomplete lineage sorting events. Incomplete lineage sorting (ILS) gives rise to homoplasic events that may occur if at the time of speciation there is a polymorphic insertion in the population [6–8] that later becomes fixed or lost at random in subsequent lineages. ILS creates situations where the gene trees are different from the species relationships. In addition, the NWM rapid speciation occurred 19-20 mya [9,10]. This contrasts with other species with rapid, yet recent, radiation events such as the baboon, genus *Papio* [11]. There is current admixture and incomplete lineage sorting that occurs within the *Papio* lineage [11]. However, this change is on-going and can be parsed out by using larger sample sizes of modern-day baboons [12]. This is not the case with NWM. Therefore, careful consideration and thorough analysis must supplement any phylogenetic tree produced by any means for NWM.

A recent study examined the amplification of a NWM specific retrotransposable element, Platy-1 [13]. However, due to the low number

* Corresponding author.
E-mail address: mbatzer@lsu.edu (M.A. Batzer).

https://doi.org/10.1016/j.ab.2019.113516
Received 15 October 2019; Received in revised form 19 November 2019; Accepted 25 November 2019
Available online 30 November 2019
0003-2697/

of Platy-1 elements in the owl monkey, capuchin monkey and squirrel monkey genomes, as well as the high percentage of Platy-1 insertions found at orthologous positions in all analyzed genomes, it was determined that Platy-1 elements would not be informative enough to resolve the position of the *Aotus* genus within the Cebidae lineage. Additionally, with so few insertions, the Platy-1 elements would not likely overcome the potentially high levels of ILS in order to resolve this portion of the NWM phylogeny.

Alu mobile elements are examples of non-autonomous short interspersed elements (SINE) specific to primates [14]. These elements comprise at least 17% of the human genome [15] and are the most successful transposable element in terms of copy number with 1.1 million copies in most primate genomes [15,16]. *Alu* elements are short, ~300 bp, with a dimeric structure [17–20]. The entire element is surrounded on both the 5′ and 3' ends by target site duplications (TSDs) as a product of movement via TPRT [19–22].

Alu elements can be broadly categorized into three groups. The oldest subfamily, *AluJ*, can be found in all primate lineages, suggesting early mobilization prior to speciation roughly 65 million years ago (mya) [23]. The second oldest subfamily, *AluS*, was active after the separation of Strepsirrhines and Tarsiformes from Platyrrhines and Catarrhines [20,24,25]. The youngest subfamily, *AluY*, is found only in Catarrhines [20,26–28]. Parallel activity of *Alu* subfamilies can occur in different primate species with the rise of new *Alu* subfamilies in any given lineage. Due to this parallel evolution, each lineage of the primate order contains its own unique set of *Alu* mobile elements. New *Alu* subfamilies occur via the stepwise accumulation of diagnostic mutations [24,29].

Alu elements are a unique phylogenetic marker. *Alu* elements are nearly homoplasy-free markers with the known ancestral state being the absence of the insertion [8,27,28,30,31]. Near-parallel and precise-parallel insertions are rare and are readily distinguished by sequence analysis [8]. In addition, these short elements are easily analyzed via polymerase chain reaction (PCR) and subsequent gel electrophoresis analysis [31]. In addition, other types of molecular markers such as single nucleotide polymorphisms (SNPs) are identical by state, not necessarily by descent. While older *Alu* elements that are fixed in a population are phylogenetically informative, younger insertions are informative for population genetics [32–36]. *Alu* insertions have proven instrumental in the resolution of many phylogenies in primates [5,12,37–42].

Recently, the *Papio* baboon phylogeny was resolved using SRA (sequence read archive) data with the polyDetect program pipeline [12]. It was thought that the same pipeline could be applied to *Alu* insertions from NWM SRA data to overcome previous difficulties that have arisen in attempts to resolve the Cebidae NWM phylogeny. The combination of the copy number and characteristics of *Alu* element insertions as well as the high throughput polyDetect program is a powerful method to analyze the position of the *Aotus* genus within Cebidae. Additionally, polyDetect makes use of a common reference sequence. The subsequent output that uses the same coordinates for analyzing all SRA data is helpful in determining shared *Alu* insertions. This study attempts to elucidate one small portion of the NWM phylogeny; the position of the *Aotus* genus among the Cebidae family [3,10,43] using the polyDetect pipeline and available SRA NWM data.

With the advent of faster and cheaper sequencing technologies comes a massive amount of data to sort through. Not only are model organisms sequenced but also non-model organisms allowing for a greater breadth and depth of research to be conducted. At its current pace the amount of genomic data doubles every seven months [44]. It is predicted that this will mean 2.5 million genomes sequenced by the year 2025 [44], not including the number of individual human genomes that may be sequenced for personal or medical reasons. The question then is not about data availability, but how to accurately process and draw conclusions from the potentially heterogeneous data obtained from different tissues and species, as well as the type of sequencing performed, read depth coverage, assembly method etc., [44,45]. Keeping this potential data heterogeneity in mind, the pipeline was fully assessed for accuracy when applied in the current study. The application of the polyDetect pipeline to this portion of the NWM phylogeny highlights the limitations of this program when applied to highly diverged genera.

2. Materials and Methods

2.1. Shared Alu elements

To analyze *Alu* elements found at orthologous loci in NWM (i.e., shared among NWM), the polyDetect pipeline was used as previously described [12]. Briefly, short read data (referred to as SRA data) from the common marmoset (*Callithrix jacchus*; caljac3), capuchin monkey (*Cebus imitator*; Cebus_imitator-1.0), squirrel monkey (*Saimiri boliviensis*; saiBol1) and owl monkey (*Aotus nancymaae*; Anan_2.0) were downloaded from NCBI. Two sets of SRA data were utilized for each NWM. SRA data set 1 (DS1) included SRA files containing similar amounts of data to attempt to ensure even coverage for all organisms used (Coverage average ± standard deviation: 30.69 ± 2.96) with the same platform utilized for the common marmoset, squirrel monkey, capuchin monkey and owl monkey (Table A.1). SRA data set 2 (DS2) contained SRA data for the same NWM as DS1, but the sequencing platforms and coverage vary (Table A.2; Average coverage ± standard deviation: 261.31 ± 196.43). These two datasets are mutually exclusive. The pipeline then maps these reads to a common reference *Alu* consensus sequence via BWA mem [46]. The various reference *Alu* sequences utilized in this study were: *Alu*S, *Alu*Ta7, *Alu*Ta10 and *Alu*Ta15 [5]. The *AluS* consensus sequence was obtained from RepBase [47] while the *Alu*T consensus sequences were obtained from Ray et al. (2005) [40]. The *Alu* portion of the split-read was then cleaved and the remaining sequence mapped to a reference genome assembly using bowtie2 [48]. The different NWM reference genomes used to map the remaining portion of the reads were: two NWM belonging to the Cebidae family (the common marmoset and the squirrel monkey), a member of the Pitheciidae family of NWM (the white-faced saki -*Pithecia pithecia*; PitPit_v1_BIUU) and a member of the Atelidae family of NWM (the black-handed spider monkey *Ateles geoffroyi*; AteGeo_v1_BIUU). The genomes for the white-faced saki and black-handed spider monkey were also obtained from NCBI. It should be noted that there was no similar SRA data available for the white-faced saki or the black-handed spider monkey. The resulting polyDetect output indicates the chromosomal location where an *Alu* insertion is found. All organisms whose SRA indicated that an *Alu* insertion was in that organism were listed in the polyDetect output, allowing for the identification of shared *Alu* loci. At least two separate reads were required to validate the presence of an *Alu* insertion. All possible combinations of organisms sharing an *Alu* insertions are listed in Table 1. The polyDetect genotypes were used to generate a nexus file for PAUP analysis.

2.2. Phylogenetic analysis

A heuristic search was performed using PAUP* 4.0b10 [49]. Because it is assumed that the absence of an *Alu* insertion is the ancestral state of each locus, Dollo's law of irreversibility was used in this analysis. All loci were set to Dollo.up in the PAUP* analysis. From the computationally derived genotype data obtained from the polyDetect program using DS1, a nexus file was generated using a custom python script. The presence of an insertion was scored as “1” for a filled site and “0” for an empty site. Ten thousand bootstrap replicates were performed with the maximum tree space set to all possible trees.

2.3. Lineage-specific Alu elements

The capuchin monkey, owl monkey, human (*Homo sapiens*;

Table 1
Possible polyDetect output categories. The first column shows all the possible polyDetect output combinations. 'C', 'M', 'O' and 'S' indicate that an *Alu* insertion is present in the capuchin monkey, marmoset, owl monkey, or squirrel monkey SRA data, respectively. An 'x' in a row indicates the organisms in a category that would share an *Alu* insertion, while the exclusion of an organism from a category is indicated by a gray box in that row.

	Capuchin monkey	Marmoset	Owl monkey	Squirrel monkey
CMOS	x	x	x	x
CMO	x	x	x	
COS	x		x	x
CMS	x	x		x
MOS		x	x	x
CM	x	x		
CO	x		x	
CS	x			x
MO		x	x	
MS		x		x
OS			x	x
C	x			
M		x		
O			x	
S				x

GRCh38.p13), common marmoset and squirrel monkey genomes were obtained from NCBI. The capuchin monkey and owl monkey genomes were analyzed for full-length *Alu* elements using RepeatMasker [50]. Full-length *Alu* elements are defined as possessing a start position no less than 4 bp and an end position not shorter than 267 bp. Full-length elements were extracted from the RepeatMasker output using a custom python script. These elements, along with 600 bp 5′ and 3' flanking sequence, were then compared to the remaining genomes in a sequential BLAT [51] in the following order: 1) human; 2) common marmoset; 3) capuchin monkey or owl monkey; and 4) squirrel monkey. A sequential BLAT included determining lineage specificity after each BLAT to a genome by using a custom python script that analyzes the BLAT output for gaps of a specific length when comparing the query and target sequences that would indicate an *Alu* element is present in the genome of interest, but not the target genome. Such lineage-specific *Alu* elements would then be assessed for lineage specificity in the next genome comparison. This process was completed until all aforementioned genomes were compared to the *Alu* elements ascertained from either the capuchin monkey or owl monkey genome. For both the capuchin monkey and owl monkey lineage specific *Alu* insertions, 100 randomly selected insertions were chosen for the design of oligonucleotide primers and PCR analysis.

2.4. Oligonucleotide primer design

The loci determined to contain *Alu* elements unique in the owl monkey or capuchin monkey genome were put into individual files containing the orthologous sequences from marmoset, squirrel monkey, owl monkey and capuchin monkey genomes. These sequences were aligned using CLUSTALW [52] and/or MUSCLE [53]. Forward and reverse oligonucleotide primers for PCR were designed using Primer3 (v.0.4.0) and checked in BioEdit to ensure minimal mismatches to allow for the amplification of a PCR product in all genomes specified. In silico PCR [51] was used to confirm the oligonucleotide primers would amplify only one product in multiple species (Table A.3; Table A.4).

2.5. DNA samples

DNA samples are described in Tables A.5-A.7. Briefly, there were three panels utilized for this study: a NWM panel, an owl monkey panel, and a capuchin monkey panel. The NWM panel contained three Old World monkeys (OWM) and sixteen NWM species representing the three NWM families. This DNA panel was used to screen elements for lineage-specificity. The owl monkey panel included DNA samples from 23 individuals of the genus *Aotus* representing five species, and the capuchin monkey panel included DNA from 14 different capuchin monkeys, 8 *Cebus apella,* now considered genus *Sapajus apella* [54], and 6 individuals from genus *Cebus* including the *Cebus imitator* sample used as the reference genome.

2.6. PCR amplification

PCR amplification was performed in 25 μL reactions containing 25 μg of template DNA, 200 nM of each primer, 1.5 mM $MgCl_2$, 10x PCR buffer (1x: 50 mM KCl; 10 mM TrisHCl, pH 8.4), 0.2 mM dNTPs, and 1 unit of *Taq* DNA polymerase. The PCR reaction protocol is as follows: 94 °C for 1 min, 32 cycles of denaturation at 94 °C for 30 s, 30 s at the appropriate annealing temperature (typically 57 °C), extension at 72 °C for 30 s, followed by a final 72 °C extension step for 2 min. Gel electrophoresis was performed on a 2% agarose gel containing 0.2 μg/mL ethidium bromide for 60 min at 180 V. UV fluorescence was used to visualize the DNA fragments using a BioRad ChemiDoc XRS imaging system (Hercules, CA). If PCR results were weak or unresolved, the PCR reaction was repeated using hot-start with the JumpStart *Taq* DNA polymerase kit (Sigma Aldrich). Genotypes were recorded in a Microsoft Excel worksheet as (0,0) homozygous

Table 2
Different polyDetect program runs using the DS1 data set.

polyDetct Run #	Alu consensus	Reference assembly
1	*Alu*S	Common marmoset
2	*Alu*Ta7	Common marmoset
3	*Alu*Ta10	Common marmoset
4	*Alu*Ta15	Common marmoset
5	*Alu*Ta15	White-faced saki
6	*Alu*Ta15	Squirrel monkey
7	*Alu*Ta15	Black-handed spider monkey

absent, (1,1) homozygous present or (1,0) for heterozygous (Table A.8; Table A.9).

3. Results and discussion

3.1. Computational NWM phylogeny assessing shared Alu insertions

Using DS1 SRA data, seven different polyDetect program runs were completed (Table 2). Runs one through four were completed with different *Alu* consensus sequences but the same reference genome, the common marmoset, to determine if the *Alu* consensus sequence used would influence the program output. *Alu*T subfamilies are specific to NWM, with *Alu*Ta7 and *Alu*Ta10 common to all NWM, and *Alu*Ta15 specific to Cebidae [5]. The results of the analysis completed with runs one through four are in Fig. 1. The results of the polyDetect analyses for these four program runs are strikingly similar. This result is surprising, as the different subfamilies arose during different time points in primate evolution. These results indicate that this pipeline may not be able to distinguish between different *Alu* subfamilies. This is due in part to the age of each of the subfamilies with *Alu*S belonging to the oldest *Alu* family used in this study followed by *Alu*Ta7, then *Alu*Ta10, and ending with *Alu*Ta15 as the youngest subfamily. In addition, *Alu*Ta15 has been shown to be specific to the Cebidae lineage [5]. As such, it was hypothesized that most of the computationally derived insertions would belong to the *Alu*Ta15 subfamily. This potential disparity might have to do with the alignment of the reads to the *Alu* consensus sequence. The resulting position of the *Alu* insertion is based upon the 5′ end of the *Alu* sequence. The 5′ end of the *Alu* sequence contains the left monomer and the A and B boxes necessary for transcription by RNA polymerase III. Therefore, the 5' end of the *Alu* sequence would need to be highly conserved between subfamilies in order to preserve the A and B boxes, and consequently the first step in retrotransposon mobilization.

No lineage-specific insertions were observed in the polyDetect output for any of the organisms used. This contrasts with the initial use of this program to investigate the *Papio* lineage, where many lineage-specific insertions were found for all individuals analyzed [12]. In addition, the smallest category within every run was that of *Alu* insertions shared among all four NWM individuals (Category CMOS, Table 1). This is in contrast with previous analyses using *Alu* elements as phylogenetic markers [4,5,40] as well as the results using a different retrotransposable element, Platy-1, in which most of the elements found were shared by all NWM individuals analyzed [13].

The underlying principle of the polyDetect pipeline is homology of the SRA data to a selected reference genome and the *Alu* element consensus sequence. To test the hypothesis of low homology/identity to the marmoset genome, three additional polyDetect program runs using DS1 with the *Alu*Ta15 as the reference *Alu* consensus sequence and three different reference NWM genome assemblies were completed (Fig. 2). A similarly low percentage value of the total polyDetect output was observed in the CMOS with all the different reference genomes utilized. It was also noted that similar trends were observed for the two non-Cebidae reference genomes (Ateles & Pithecia), while the values for certain categories changed depending upon the Cebidae reference genome used. For example, the insertion percentage in the CM category (insertions shared between capuchin monkey and the common marmoset to the exclusion of owl monkey and squirrel monkey) was much higher when the common marmoset was used as the reference genome assembly than when the squirrel monkey was used as the reference assembly (18% and 11%, respectively). A trend was observed where if there was a category in which the reference genome was also a member, then there was an increase in that category relative to to the other Cebidae reference genome and a corresponding decrease in the categories where the reference genome was not a member, highlighting the influence of the chosen reference genomes on the polyDetect output. The categories CMO, CM, CS, MS, MO and OS were particularly sensitive to the phenomenon (Fig. 2). However, overall similar trends were observed using different reference genomes with an emphasis on the low percentage of insertions shared by all NWM in this study and the absence of lineage-specific insertions identified.

An additional possibility is that DS1 did not contain adequate coverage of the different individuals. Therefore, a second data set, DS2, with larger SRA files and corresponding higher coverage was utilized. This data set used the *Alu*Ta15 consensus sequence and the marmoset genome as the reference genome assembly. It was observed that there was a percentage increase in the categories where squirrel monkey was

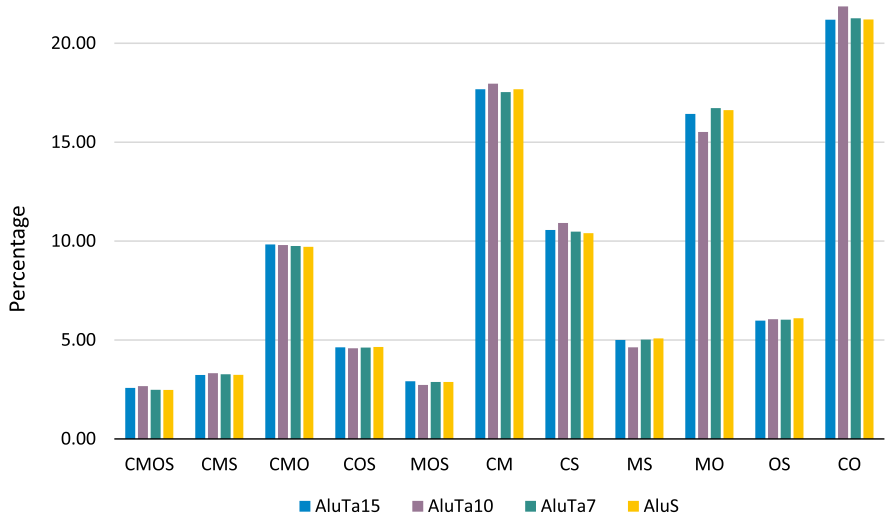

Fig. 1. Output of the polyDetect program pipeline using different *Alu* consensus sequences. The output of the polyDetect pipeline using the *Alu*Ta15 (blue), *Alu*Ta10 (purple), *Alu*Ta7 (green) and *Alu*S (yellow) consensus sequences. Each of these consensus sequences is associated with a different run of the polyDetect program using DS1 data set with the common marmoset as the reference genome assembly. The X-axis indicates pre-defined shared categories (Table 1). Percentage on the Y-axis indicates the percent of the total number of predicted shared insertions as seen in polyDetect program output that belong to that category.

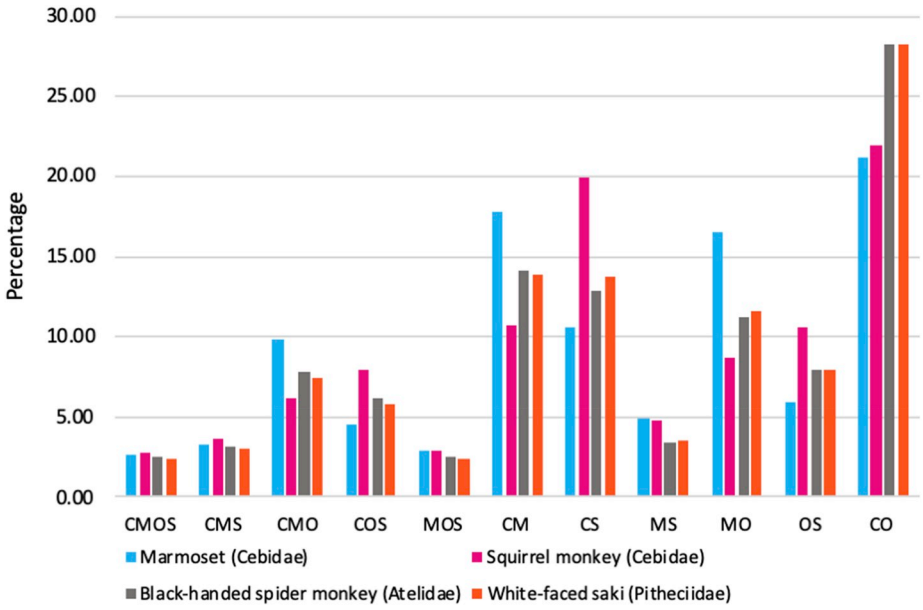

Fig. 2. Output of the polyDetect program pipeline using different reference genome assemblies. The output of the polyDetect program using the common marmoset (blue), squirrel monkey (pink), black-handed spider monkey (gray) and the white-faced saki (orange) as the reference genome assembly. The New World monkey family name is in parentheses following the common name. Each different genome assembly was a separate run on the polyDetect program pipeline. Each program was completed using the *Alu*Ta15 consensus sequence and the DS1 data set. The X-axis indicates pre-defined shared categories (Table 1). Percentage on the Y-axis indicates the percent of the total number of predicted shared insertions as seen in polyDetect program output that belong to that category.

a member (Fig. 3). The percent increase indicates that the DS1 data set had lower sequence coverage of squirrel monkey compared to the other individuals in the same data set. The lower number of squirrel monkey calls was verified by parsing through the polyDetect output (data not shown). Two categories showed a drastic change between data sets: CM and OS. From the smaller data set (DS1) to the larger data set (DS2) a large portion of the insertions belonging to the CM category were lost while insertions in the OS category were gained. It is possible that due to the increased overall sequence coverage in DS2 and increased coverage in the squirrel monkey individual in particular that there was a shift in the categories away from only being shared between capuchin monkey and marmoset to more genera. Therefore, higher coverage should lead to a large increase in the CMOS and CMS categories. However, only a slight increase is observed between data sets in the CMOS category, with no change seen in the CMS category. Another possibility could therefore be that the second data set has its own inherent level of coverage challenges with variable read quality for each individual. Even with higher sequence coverage there were no lineage-specific elements and a low number of elements shared by all four NWM found in this run using the DS2 dataset implying quantity does not always ensure quality.

3.2. PAUP analysis

The resulting phylogenetic trees produced by the pipeline output varied greatly depending on the different reference genomes utilized. The different topologies from the PAUP analysis using the different reference genomes show many possible phylogenetic scenarios (Fig. 4). When the squirrel monkey genome was used as a reference there was a close relationship seen with the capuchin monkey genome with 99.36% bootstrap support for a branch that indicated squirrel and capuchin monkeys were sister to the exclusion of owl monkey and marmoset. In addition, there was 100% bootstrap support for a branch grouping capuchin monkey, squirrel monkey and owl monkey to the exclusion of marmoset. However, when either the white-faced saki or the black-handed spider monkey were used as the reference genome, there was 100% bootstrap support for capuchin monkey and owl monkey as sister groups with 100% bootstrap support for marmoset as the outgroup to capuchin monkey and owl monkey. Identical trees were observed when either the white-faced saki or the black-handed spider monkey genome was used. This is most likely a result of both of these NWM belonging to a family outside of the Cebidae lineage. Both NWM would therefore have the same relationship to the Cebidae lineage. When the marmoset

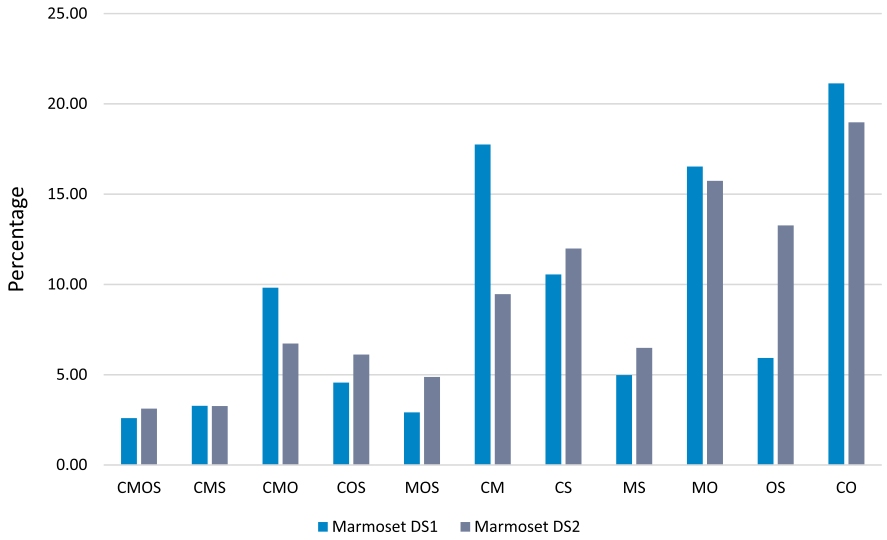

Fig. 3. Output of the polyDetect program pipeline comparing two different SRA datasets. The polyDetect output is the result of using DS1 (light blue) and DS2 (dark blue). Both data sets were completed using the common marmoset as the reference genome assembly and *Alu*Ta15 as the *Alu* consensus sequence. The X-axis indicates pre-defined shared categories (Table 1). Percentage on the Y-axis indicates the percent of the total number of predicted shared insertions as seen in polyDetect program output that belong to that category.

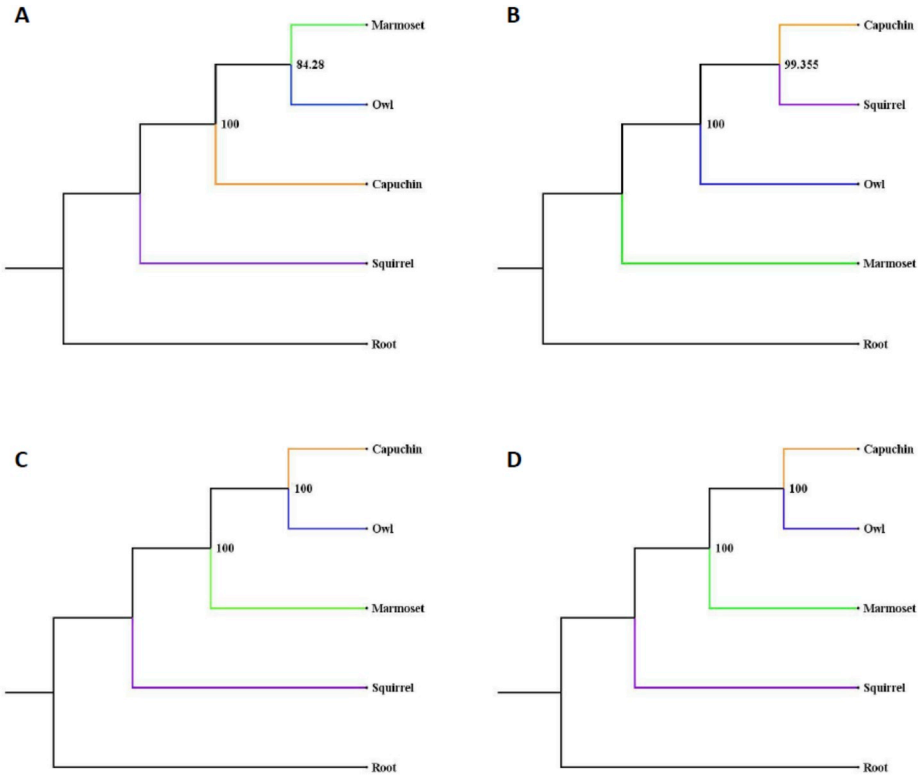

Fig. 4. Phylogenetic analysis of the polyDetect pipeline output. A tree was generated using PAUP with the polyDetect data generated with the following reference genomes using DS1 with the *Alu*Ta15 consensus sequence: A) common marmoset B) squirrel monkey C) white-faced saki D) black-handed spider monkey. On each phylogeny the colors correspond to the following organism: green (marmoset), orange (capuchin monkey; referred to as, 'capuchin'), blue (owl monkey; referred to as, 'owl') and purple (squirrel monkey; referred to as, 'squirrel').

Table 3
PAUP input and parsimony statistics. C.I. (Consistency index); R.I. (Retention index); H.I. (Homoplasy index). Each row represents a separate polyDetect run with the reference genome and *Alu* sequence indicated.

DS1						
Reference genome	Reference Alu	Characters	Phylogenetically informative	C.I.	R.I.	H.I.
Ateles	AluTa15	7135	6955	0.549	0.531	0.451
Pithecia	AluTa15	7057	6890	0.545	0.527	0.455
Saimiri	AluTa15	8617	8374	0.541	0.512	0.459
Marmoset	AluTa15	8080	7871	0.549	0.529	0.451
Marmoset	AluSc7	8251	8046	0.548	0.528	0.452
Marmoset	AluTa7	8238	8033	0.549	0.53	0.451
Marmoset	AluTa10	8238	8033	0.549	0.53	0.451
DS2						
Marmoset	AluTa15	20871	20206	0.515	0.455	0.485

genome was used as the reference genome, there was 84% bootstrap support for a branching pattern indicating marmoset and owl monkey as sister groups to the exclusion of the other two genomes. There was 100% bootstrap support for capuchin monkey as an outgroup to this branching pattern. This indicates that the reference genome used greatly influences the output of this program most likely due to the homology/identity-based nature of this program. In each of the trees generated (Fig. 4), there were thousands of input loci and thousands of phylogenetically-informative loci for any combination of reference genome and consensus *Alu* sequence analyzed (Table 3). However, the consistency index (CI), retention index (RI) and homoplasy index (HI) indicate that a significant amount of homoplasy is present in each of the datasets. The high level of homoplasy could be the result of ILS. An HI closer to zero would indicate low homoplasy and a CI and RI closer to one would also indicate low homoplasy in the data set, whereas all indices are mid-range for every comparison reported here signaling potential inadequacies with this approach. It should be noted that there was no SRA data available for the white-faced saki or the black-handed spider monkey. However, use of their genome coordinates was used to generate nexus files for PAUP analysis and phylogenetic tree generation.

Interestingly, every tree generated from these comparisons showed that either marmoset or squirrel monkey was most basal while the capuchin monkey and owl monkey were never basal to all other NWM in this study. In addition, no differences in tree topology were observed with the use of lower (DS1) and higher (DS2) coverage data sets, although the branch indicating the capuchin monkey as the most basal of capuchin monkey, owl monkey and marmoset using the DS2 SRA data had lower bootstrap support (Fig. 5).

3.3. Lineage-specific insertions in the capuchin and owl monkey genomes

Previous studies have reported new *Alu* subfamilies and analyzed the number of lineage specific *Alu* element insertions in the common marmoset genome [55] and the squirrel monkey genome [16] but these analyses had not yet been conducted for capuchin or owl monkey. Due

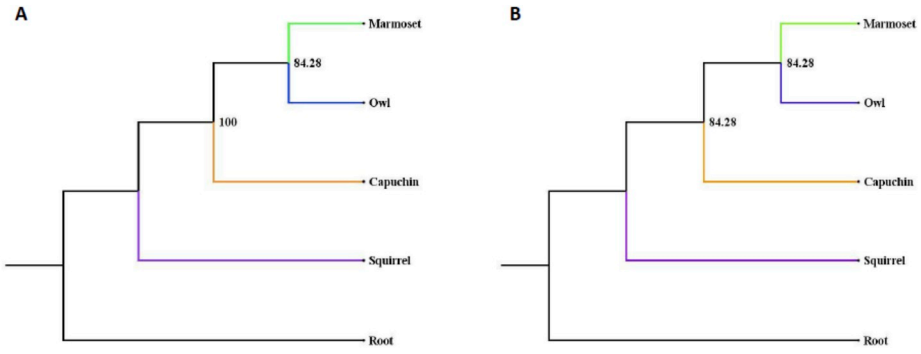

Fig. 5. PAUP comparison between two SRA datasets. A tree was generated using PAUP with polyDetect data generated with the marmoset genome as the reference, the *Alu*Ta15 consensus sequence with the following SRA datasets: A) DS1 B) DS2. On each phylogeny the colors correspond to the following organism: green (marmoset), orange (capuchin monkey; referred to as, 'capuchin'), blue (owl monkey; referred to as, 'owl') and purple (squirrel monkey; referred to as, 'squirrel').

to the low number of insertions in the CMOS category in all analyses using polyDetect, it is possible that the number of lineage-specific insertions in two of the genomes not assessed for lineage-specific insertions, owl monkey and/or capuchin monkey, have a higher number of lineage-specific *Alu* elements, leading to a low number of shared elements by all four of the NWM tested here.

To test this hypothesis RepeatMasker analyses of the capuchin and owl monkey genomes were performed, providing 617,132 and 658,009 full-length *Alu* elements, respectively. Following the sequential BLAT (see Materials and Methods), 9602, or 1.55% of the 617,132 full-length *Alu* elements in the capuchin genome were lineage specific. The same procedure yielded 12,225, or 1.86% of the 658,009 full-length *Alu* elements that were lineage specific to the owl monkey genome.

To verify that this filtering procedure produced only lineage specific insertions, 100 randomly selected loci from each genome were analyzed via PCR on a NWM panel (Fig. 6). Of the 100 random putative lineage specific *Alu* insertions ascertained from the capuchin monkey genome, 90 of these were determined to be specific to the capuchin monkey individuals on the DNA panel with 40 of these loci polymorphic for insertion presence/absence (Table A.8). 4 of the 100 loci repeatedly failed to amplify a PCR product and were eliminated. Of the 100 random putative lineage specific *Alu* insertions recovered from the owl monkey genome, 88 of these loci were specific to owl monkey individuals with 19 of these loci being polymorphic for the insertion among *Aotus* samples (Table A.9). 6 of the 100 loci from the owl monkey genome repeatedly failed to amplify during PCR and were discarded. These results verify that the method of filtering correctly identified lineage specific insertions.

Next, the full-length *Alu* insertions from both the capuchin and owl monkey genomes were compared against the human genome via BLAT to eliminate any *Alu* insertions that would be shared with primates before the rise of NWM. After this analysis, 58,952 and 77,564 insertions were remaining for the capuchin and owl monkey genomes, respectively. These numbers are far greater than the lineage specific insertions for each genome described above. Not including the linage specific insertions, there are 49,350 and 65,339 *Alu* insertions remaining in the capuchin and owl monkey genomes, respectively. This indicates that the lineage-specific *Alu* insertions belonging to both of these genomes compromises a small portion of insertions compared to those that may be shared with other NWM. Therefore, the low number of insertions found in the CMOS category after completing the polyDetect pipeline is not due to a high number of lineage specific insertions in these two genomes, but rather due to lack of sufficient detection.

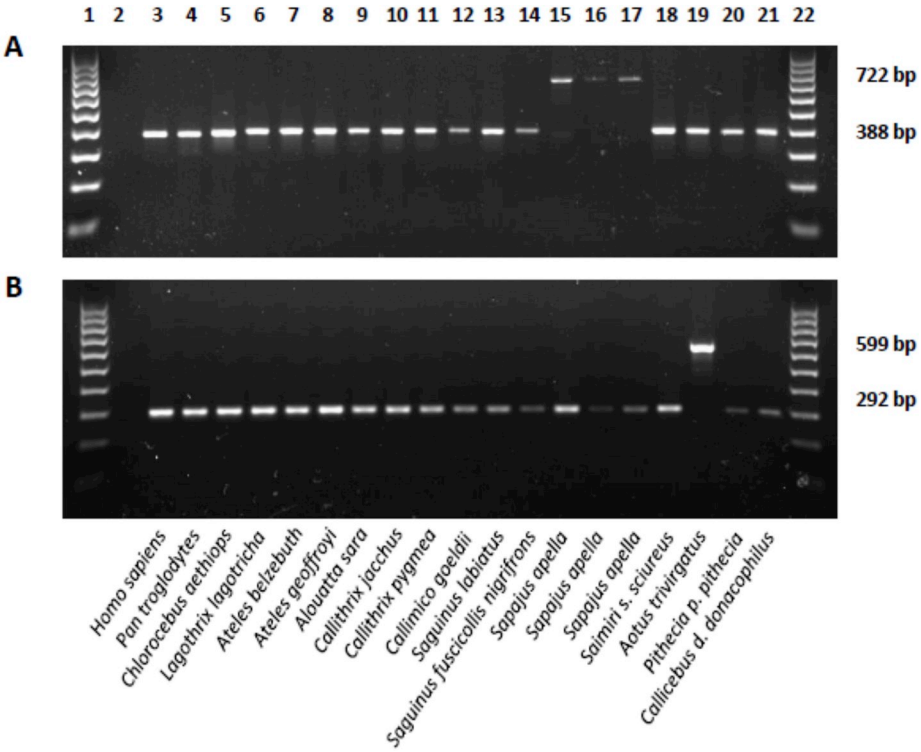

Fig. 6. Lineage-specific *Alu* elements. A) The presence of the Cebus_738_016107358_8091150 capuchin monkey specific *Alu* element is indicated by the higher of the two bands present (722 bp band), while the absence is indicated by the lower of the two bands present (388 bp band). B) The presence of the Aotus_3842_018509268_1336944 owl monkey specific *Alu* element is indicated by the higher of the two bands present (599 bp band), while the absence is indicated by the lower of the two bands present (292 bp band). 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; 15-17-Capuchin monkey; 18-Squirrel monkey; 19-Owl monkey; 20-Northern white-faced saki; 21-Bolivian gray titi; 22–100 bp ladder. Scientific names of the primates are indicated below the gel images.

Table 4
Output of bwa mem alignment of SRA reads to *Alu* consensus sequence for polyDetect run #4. PE (paired-end); SE (single-end).

Organism	Percent PE reads aligned	Total PE reads	Percent SE reads aligned	Total SE reads
Squirrel monkey	49.69	2,281,688	44.19	14,709,295
Owl monkey	42.59	2,804,336	40.78	15,906,904
Capuchin monkey	65.21	1,803,178	55.72	9,083,234
Marmoset	92.77	401,458	91.29	5,060,388

3.4. Alignment of polyDetect predicted shared Alu elements

The primary benefit of having genome assemblies available for the NWM in this computational study is the ability to compare the polyDetect pipeline output data to the assembled genome sequences. To determine if the pipeline was predicting shared *Alu* elements correctly, the coordinates from ten randomly selected loci from each shared category from the polyDetect output were selected. 600 bp of flanking sequence were added to the 5′ and 3' end of the predicted insertion breakpoint and the corresponding FASTA sequence extracted. The orthologous sequence from the capuchin monkey, owl monkey and squirrel monkey genomes were obtained via BLAT and aligned. All the predicted insertions were inaccurate and were in fact almost entirely shared by all four of the aforementioned NWM (Supplemental file 2). This indicates that the exclusion of an organism from the polyDetect output does not necessarily indicate its absence. This information explains the high bootstrap support for conflicting trees, as the information used to generate those trees was not accurate. There are several possible explanations for inaccuracies in the polyDetect output. The first potential explanation is an issue with overall read coverage for repeat regions of the genome. It is possible that using the unassembled read data failed to identify unique *Alu* insertions with such short reads. The second potential source of error is related to the comparisons being made. It is possible that the polyDetect program becomes less applicable as the species or genera of interest become farther diverged from one another.

3.5. Filtering split-reads based on unique 5' flanking sequence

One possibility of inaccurate pipeline predictions could arise from the inaccurate map placement of an *Alu* element due to lack of sufficient unique 5′ flanking sequence. A recent analysis using the polyDetect pipeline of *Alu* insertions shared between *Papio* baboons and a gelada monkey was completed [56]. Wet bench validation via PCR found that only 71% of the predicted calls by polyDetect were verified. After looking at the raw reads corresponding to a predicted *Alu* locus, it was found that longer flanking sequence corresponded with validated loci. The average TSD for retrotransposable elements that move via TPRT is 14 bp long [57,58]. This is typically an A/T rich region. Therefore, the unique 5′ flanking sequence past the *Alu* should reach past the TSD. Therefore, a 30 bp filter on split reads was imposed, where the minimum 5' flanking sequence in the read before the *Alu* insertion was 30 bp, which includes the 14 bp TSD and 16 bp additional flanking sequence. After this filter was imposed upon the read data obtained from DS2, the polyDetect program was completed using the marmoset genome as a reference and *Alu*Ta15 as the *Alu* consensus sequence, and the output analyzed. After extracting the predicted insertion loci from the output of the program, the sequences were extracted in FASTA format from the marmoset genome as before and the orthologous sequence obtained via BLAT from the capuchin monkey, owl monkey and squirrel monkey genomes. Unfortunately, this additional filter, while helpful to other studies, such as the aforementioned baboon study, was not beneficial in this application. All of the loci analyzed were indeed inaccurate and were instead shared among all four NWM genomes (data not shown). This indicates that this pipeline is not a viable option for the reconstruction of the NWM phylogeny. It is possible that not enough homology/identity exists for the short reads to map properly to the reference genome when multiple genera are being studied.

3.6. Alignment of SRA data to Alu consensus sequences

It is clear from the data presented thus far that the polyDetect pipeline is not applicable to this particular study. However, it is unknown at which point the error or problem occurs in the pipeline. There are two possibilities: either the SRA data is not aligning to the *Alu* sequence or after the alignment to the *Alu* consensus sequence there is not enough flanking to accurately map to the reference genome. Using the data from polyDetect run #4 (Table 2) as an example, the first alignment of the SRA data to the *Alu* consensus sequence using bwa mem was determined (Table 4).

These data indicate that alignment to the *Alu* consensus sequence was properly achieved as the percent of reads that aligned from the paired-end (PE) reads ranged from 42.59 to 92.77% and the single-end (SE) reads ranged from 40.78 to 91.29% (Table 4). This indicates that the bowtie2 alignment step of aligning *Alu* flanking sequence was inefficient, likely due to evolutionary decay of homologous sequence. In addition, the dataEval program, a part of the polyDetect pipeline, only analyzes split reads that match to the 5' end of the *Alu* consensus sequence. This could also result in a fewer number of shared *Alu* insertions. Therefore, the ~20 my divergence between the organisms used in this study could not be overcome with this pipeline.

4. Conclusions

The polyDetect pipeline was initially designed and used for resolving the controversial baboon phylogeny resulting from a rapid divergence in the recent past and ongoing admixture within genus *Papio* [11,12]. Those studies utilized the closely related rhesus macaque (*Macaca mulatta*) genome as the reference for the program producing highly reproducible and high confidence mapped data that was verified by comparing read data produced from 12 individuals representing 6 different species within the same genera. In contrast, the NWM used in this study belong to four different genera. In addition, the rapid 1-2 my divergence in the NWM lineage took place ~19-20 mya [9,10]. Because polyDetect is a homology/identity-based program, this long time period of divergence might obscure any homology/identity that could be detected in short read sequence data [59]. In addition, there are two homology/identity searches that occur sequentially. First, the reads are mapped to the *Alu* consensus sequence. The nucleotides in the short reads that mapped to the *Alu* were clipped and the second homology/identity search to map the flanking sequence to the reference genome sequence was performed. As seen in Table 4, at least 40% of the reads effectively mapped to the *Alu*. It is also worth noting that of those reads that mapped to the *Alu* consensus sequence a portion of them mapped multiple times (data not shown). In addition, the highest number of elements in any data set in this study was 20,871 obtained from using the DS2 dataset, far below the expected number of *Alu* elements based on the literature. There are ~1 million *Alu* elements in primate genomes, with 600,000–730,000 of those elements constituting full-length repeats [16–19,40]. Taken together, these results indicate that there was insufficient unique flanking sequence to map to the reference

genome to ensure accuracy in the program output for these highly divergent organisms.

These analyses indicate that the polyDetect pipeline is best suited to resolving phylogenies of closely related organisms, with an emphasis that those organisms belonging to the same genus, and when assembled genomes are not available for the organisms of interest.

Authors' contributions

JMS performed the analyses of the Cebidae genomes and wrote the manuscript. VEJ developed the initial polyDetect program and JMS revised the pipeline and performed all the analyses for this project. JAW and MAB contributed to experimental design and performed the final edits to the manuscript. All authors read and approved the final manuscript.

Declaration of competing interest

None.

Acknowledgments and funding

The authors would like to thank all the members of the Batzer Lab for their helpful suggestions. This work was supported by National Institutes of Health Grant R01 GM59290 (M.A.B.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ab.2019.113516.

References

[1] S.I. Perez, M.F. Tejedor, N.M. Novo, L. Aristide, Divergence times and the evolutionary radiation of new world monkeys (Platyrrhini, primates): an analysis of fossil and molecular data, PLoS One 8 (2013) e68029.
[2] P. Hershkovitz, Living New World Monkeys (Platyrrhini), Volume 1 with an Introduction to Primates, (1977).
[3] H. Schneider, I. Sampaio, The systematics and evolution of New World primates - a review, Mol. Phylogenetics Evol. 82 Pt B (2015) 348–357.
[4] M. Osterholz, L. Walter, C. Roos, Retropositional events consolidate the branching order among New World monkey genera, Mol. Phylogenetics Evol. 50 (2009) 507–513.
[5] D.A. Ray, M.A. Batzer, Tracking alu evolution in new world primates, BMC Evol. Biol. 5 (2005) 51.
[6] M. Nei, Molecular Evolutionary Genetics, Columbia University Press, New York, 1987.
[7] P. Pamilo, M. Nei, Relationships between gene trees and species trees, Mol. Biol. Evol. 5 (1988) 568–583.
[8] D.A. Ray, J. Xing, A.H. Salem, M.A. Batzer, SINEs of a nearly perfect character, Syst. Biol. 55 (2006) 928–935.
[9] P. Perelman, W.E. Johnson, C. Roos, H.N. Seuanez, J.E. Horvath, M.A. Moreira, B. Kessing, J. Pontius, M. Roelke, Y. Rumpler, M.P. Schneider, A. Silva, S.J. O'Brien, J. Pecon-Slattery, A molecular phylogeny of living primates, PLoS Genet. 7 (2011) e1001342.
[10] H. Schneider, The current status of the New World monkey phylogeny, An. Acad. Bras. Cienc. 72 (2000) 165–172.
[11] J. Rogers, M. Raveendran, R.A. Harris, T. Mailund, K. Leppala, G. Athanasiadis, M.H. Schierup, J. Cheng, K. Munch, J.A. Walker, M.K. Konkel, V. Jordan, C.J. Steely, T.O. Beckstrom, C. Bergey, A. Burrell, D. Schrempf, A. Noll, M. Kothe, G.H. Kopp, Y. Liu, S. Murali, K. Billis, F.J. Martin, M. Muffato, L. Cox, J. Else, T. Disotell, D.M. Muzny, J. Phillips-Conroy, B. Aken, E.E. Eichler, T. Marques-Bonet, C. Kosiol, M.A. Batzer, M.W. Hahn, J. Tung, D. Zinner, C. Roos, C.J. Jolly, R.A. Gibbs, K.C. Worley, The comparative genomics and complex population history of Papio baboons, Sci. Adv. 5 (2019) eaau6947.
[12] V.E. Jordan, J.A. Walker, T.O. Beckstrom, C.J. Steely, C.L. McDaniel, C.P. St Romain, K.C. Worley, J. Phillips-Conroy, C.J. Jolly, J. Rogers, M.K. Konkel, M.A. Batzer, A computational reconstruction of Papio phylogeny using Alu insertion polymorphisms, Mob. DNA 9 (2018) 13.
[13] J.M. Storer, J.R. Mierl, S.A. Brantley, B. Threeton, Y. Sukharutski, L.C. Rewerts, C.P. St Romain, M.M. Foreman, J.N. Baker, J.A. Walker, J.D. Orkin, A.D. Melin, K.A. Phillips, M.K. Konkel, M.A. Batzer, Amplification dynamics of Platy-1 retrotransposons in the Cebidae platyrrhine lineage, Genome Biol. Evol. 11 (2019) 1105–1116.
[14] C.M. Houck, F.P. Rinehart, C.W. Schmid, A ubiquitous family of repeated DNA sequences in the human genome, J. Mol. Biol. 132 (1979) 289–306.
[15] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y.J. Chen, J. Szustakowki, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[16] J.N. Baker, J.A. Walker, J.A. Vanchiere, K.R. Phillippe, C.P. St Romain, P. Gonzalez-Quiroga, M.W. Denham, J.R. Mierl, M.K. Konkel, M.A. Batzer, Evolution of alu subfamily structure in the Saimiri lineage of new world monkeys, Genome Biol. Evol. 9 (2017) 2365–2376.
[17] M.A. Batzer, P.L. Deininger, Alu repeats and human genomic diversity, Nat. Rev. Genet. 3 (2002) 370–379.
[18] P. Deininger, Alu elements: know the SINEs, Genome Biol. 12 (2011) 236.
[19] P.L. Deininger, M.A. Batzer, Mammalian retroelements, Genome Res. 12 (2002) 1455–1465.
[20] M.K. Konkel, J.A. Walker, M.A. Batzer, LINEs and SINEs of primate evolution, Evol. Anthropol. 19 (2010) 236–249.
[21] R. Cordaux, M.A. Batzer, The impact of retrotransposons on human genome evolution, Nat. Rev. Genet. 10 (2009) 691–703.
[22] D.D. Luan, M.H. Korman, J.L. Jakubczak, T.H. Eickbush, Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition, Cell 72 (1993) 595–605.
[23] P. Deininger, G. Daniels, The recent evoltuion of mammalian repetitive DNA elements, Trends Genet. 2 (1986) 76–80.
[24] J. Jurka, T. Smith, A fundamental division in the Alu family of repeated sequences, Proc. Natl. Acad. Sci. U.S.A. 85 (1988) 4775–4778.
[25] V. Kapitonov, J. Jurka, The age of Alu subfamilies, J. Mol. Evol. 42 (1996) 59–65.
[26] M.A. Batzer, P.L. Deininger, A human-specific subfamily of Alu sequences, Genomics 9 (1991) 481–487.
[27] M.A. Batzer, V.A. Gudi, J.C. Mena, D.W. Foltz, R.J. Herrera, P.L. Deininger, Amplification dynamics of human-specific (HS) Alu family members, Nucleic Acids Res. 19 (1991) 3619–3623.
[28] M.A. Batzer, M. Stoneking, M. Alegria-Hartman, H. Bazan, D.H. Kass, T.H. Shaikh, G.E. Novick, P.A. Ioannou, W.D. Scheer, R.J. Herrera, et al., African origin of human-specific polymorphic Alu insertions, Proc. Natl. Acad. Sci. U.S.A. 91 (1994) 12288–12292.
[29] V. Slagel, E. Flemington, V. Traina-Dorge, H. Bradshaw, P. Deininger, Clustering and subfamily relationships of the Alu family in the human genome, Mol. Biol. Evol. 4 (1987) 19–29.
[30] D.M. Hillis, SINEs of the perfect character, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 9979–9981.
[31] N.T. Perna, M.A. Batzer, P.L. Deininger, M. Stoneking, Alu insertion polymorphism: a new type of marker for human population studies, Hum. Biol. 64 (1992) 641–648.
[32] J.N. Baker, J.A. Walker, M.W. Denham, C.D. Loupe 3rd, M.A. Batzer, Recently integrated Alu insertions in the squirrel monkey (Saimiri) lineage and application for population analyses, Mob. DNA 9 (2018) 9.
[33] M.J. Bamshad, S. Wooding, W.S. Watkins, C.T. Ostler, M.A. Batzer, L.B. Jorde, Human population genetic structure and inference of group membership, Am. J. Hum. Genet. 72 (2003) 578–589.
[34] C.J. Steely, J.A. Walker, V.E. Jordan, T.O. Beckstrom, C.L. McDaniel, C.P. St Romain, E.C. Bennett, A. Robichaux, B.N. Clement, M. Raveendran, K.C. Worley,

J. Phillips-Conroy, C.J. Jolly, J. Rogers, M.K. Konkel, M.A. Batzer, Alu insertion polymorphisms as evidence for population structure in baboons, Genome Biol. Evol. 9 (2017) 2418–2427.
[35] C. Stewart, D. Kural, M.P. Stromberg, J.A. Walker, M.K. Konkel, A.M. Stutz, A.E. Urban, F. Grubert, H.Y. Lam, W.P. Lee, M. Busby, A.R. Indap, E. Garrison, C. Huff, J. Xing, M.P. Snyder, L.B. Jorde, M.A. Batzer, J.O. Korbel, G.T. Marth, A comprehensive map of mobile element insertion polymorphisms in humans, PLoS Genet. 7 (2011) e1002236.
[36] W.S. Watkins, A.R. Rogers, C.T. Ostler, S. Wooding, M.J. Bamshad, A.M. Brassington, M.L. Carroll, S.V. Nguyen, J.A. Walker, B.V. Prasad, P.G. Reddy, P.K. Das, M.A. Batzer, L.B. Jorde, Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms, Genome Res. 13 (2003) 1607–1618.
[37] J. Li, K. Han, J. Xing, H.S. Kim, J. Rogers, O.A. Ryder, T. Disotell, B. Yue, M.A. Batzer, Phylogeny of the macaques (Cercopithecidae: Macaca) based on Alu elements, Gene 448 (2009) 242–249.
[38] A.T. McLain, T.J. Meyer, C. Faulk, S.W. Herke, J.M. Oldenburg, M.G. Bourgeois, C.F. Abshire, C. Roos, M.A. Batzer, An alu-based phylogeny of lemurs (infraorder: Lemuriformes), PLoS One 7 (2012) e44035.
[39] T.J. Meyer, A.T. McLain, J.M. Oldenburg, C. Faulk, M.G. Bourgeois, E.M. Conlin, A.R. Mootnick, P.J. de Jong, C. Roos, L. Carbone, M.A. Batzer, An Alu-based phylogeny of gibbons (hylobatidae), Mol. Biol. Evol. 29 (2012) 3441–3450.
[40] D.A. Ray, J. Xing, D.J. Hedges, M.A. Hall, M.E. Laborde, B.A. Anders, B.R. White, N. Stoilova, J.D. Fowlkes, K.E. Landry, L.G. Chemnick, O.A. Ryder, M.A. Batzer, Alu insertion loci and platyrrhine primate phylogeny, Mol. Phylogenetics Evol. 35 (2005) 117–126.
[41] C. Roos, J. Schmitz, H. Zischler, Primate jumping genes elucidate strepsirrhine phylogeny, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 10650–10654.
[42] S.S. Singer, J. Schmitz, C. Schwiegk, H. Zischler, Molecular cladistic markers in new world monkey phylogeny (Platyrrhini, primates), Mol. Phylogenetics Evol. 26 (2003) 490–501.
[43] H. Schneider, F.C. Canavez, I. Sampaio, M.A. Moreira, C.H. Tagliaro, H.N. Seuanez, Can molecular data place each neotropical monkey in its own branch? Chromosoma 109 (2001) 515–523.
[44] Z.D. Stephens, S.Y. Lee, F. Faghri, R.H. Campbell, C. Zhai, M.J. Efron, R. Iyer, M.C. Schatz, S. Sinha, G.E. Robinson, Big data: astronomical or genomical? PLoS Biol. 13 (2015) e1002195.
[45] J. Fan, F. Han, H. Liu, Challenges of big data analysis, Natl. Sci. Rev. 1 (2014) 293–314.
[46] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[47] J. Jurka, V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements, Cytogenet. Genome Res. 110 (2005) 462–467.
[48] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359.
[49] D.L. Swofford, PAUP*: Phylogenetic Analysis Using Parsimony, (2011) version 4.0b10.
[50] RepeatMasker-Open-4.0, http://www.repeatmasker.org.
[51] W.J. Kent, BLAT–the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.
[52] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[53] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinf. 5 (2004) 113.
[54] J.W. Alfaro, J.D. Silva Jr., A.B. Rylands, How different are robust and gracile capuchin monkeys? An argument for the use of sapajus and cebus, Am. J. Primatol. 74 (2012) 273–286.
[55] M.G.S.a.A. Consortium, The common marmoset genome provides insight into primate biology and evolution, Nat. Genet. 46 (2014) 850–857.
[56] J.A. Walker, V.E. Jordan, J.M. Storer, C.J. Steely, P. Gonzalez-Quiroga, T.O. Beckstrom, L.C. Rewerts, C.P. St Romain, C.E. Rockwell, J. Rogers, C.J. Jolly, M.K. Konkel, M.A. Batzer, Alu insertion polymorphisms shared by papio baboons and theropithecus gelada reveal an intertwined common ancestry, Mob. DNA 10 (2019) 46.
[57] M.K. Konkel, B. Ullmer, E.L. Arceneaux, S. Sanampudi, S.A. Brantley, R. Hubley, A.F. Smit, M.A. Batzer, Discovery of a new repeat family in the Callithrix jacchus genome, Genome Res. 26 (2016) 649–659.
[58] M.K. Konkel, J.A. Walker, A.B. Hotard, M.C. Ranck, C.C. Fontenot, J. Storer, C. Stewart, G.T. Marth, M.A. Batzer, Sequence analysis and characterization of active human alu subfamilies based on the 1000 genomes pilot project, Genome Biol. Evol. 7 (2015) 2608–2622.
[59] A.D. Ewing, Transposable element detection from whole genome sequence data, Mob. DNA 6 (2015) 24.