Discovery of a new repeat family in the Callithrix jacchus genome

Miriam Konkel
*Louisiana State University and Agricultural and Mechanical College, konkel@lsu.edu*

Mark Batzer
*Louisiana State University and Agricultural and Mechanical College, mbatzer@lsu.edu*

Brygg Ullmer
*Louisiana State University and Agricultural and Mechanical College, ullmer@lsu.edu*

Erika L. Arceneaux

Sreeja Sanampudi

*See next page for additional authors*

Follow this and additional works at: [https://digitalcommons.lsu.edu/biosci_pubs](https://digitalcommons.lsu.edu/biosci_pubs)

**Recommended Citation**

Konkel, Miriam; Batzer, Mark; Ullmer, Brygg; Arceneaux, Erika L.; Sanampudi, Sreeja; Brantley, Sarah A.; Hubley, Robert; and Smit, Arian F.A., "Discovery of a new repeat family in the Callithrix jacchus genome" (2016). *Faculty Publications*. 6.

[https://digitalcommons.lsu.edu/biosci_pubs/6](https://digitalcommons.lsu.edu/biosci_pubs/6)

This Article is brought to you for free and open access by the Department of Biological Sciences at LSU Digital Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of LSU Digital Commons. For more information, please contact gcoste1@lsu.edu.
Discovery of a new repeat family in the
*Callithrix jacchus* genome

Miriam K. Konkel,1,4 Brygg Ullmer,2,4 Erika L. Arceneaux,1 Sreeja Sanampudi,1
Sarah A. Brantley,1 Robert Hubley,3 Arian F.A. Smit,3 and Mark A. Batzer1

1Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA; 2School of Electrical Engineering and Computer Science, Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803, USA; 3Institute for Systems Biology, Seattle, Washington 98109-5263, USA

We identified a novel repeat family, termed Platy-1, in the *Callithrix jacchus* (common marmoset) genome that arose around the time of the divergence of platyrrhines and catarrhines and established itself as a repeat family in New World monkeys (NWMs). A full-length Platy-1 element is ∼100 bp in length, making it the shortest known short interspersed element (SINE) in primates, and harbors features characteristic of non-LTR retrotransposons. We identified 2268 full-length Platy-1 elements across 62 subfamilies in the common marmoset genome. Our subfamily reconstruction and phylogenetic analyses support Platy-1 propagation throughout the evolution of NWMs in the lineage leading to *C. jacchus*. Platy-1 appears to have reached its amplification peak in the common ancestor of current day marmosets and has since moderately declined. However, identification of more than 200 Platy-1 elements identical to their respective consensus sequence, and the presence of polymorphic elements within common marmoset populations, suggests ongoing retrotransposition activity. Platy-1, a SINE, appears to have originated from an *Alu* element, and hence is likely derived from 7SL RNA. Our analyses illustrate the birth of a new repeat family and its propagation dynamics in the lineage leading to the common marmoset over the last 40 million years.

[Supplemental material is available for this article.]

The common marmoset (*Callithrix jacchus*), also known as the white-tufted-ear marmoset, is a platyrrhine native to the Atlantic coastal forest of northeastern Brazil. Platyrrhines, commonly referred to as New World monkeys (NWMs), are primates indigenous to South America. They represent a diverse group of animals that diverged from catarrhines about 35–47 million years ago (Schrago and Russo 2003; Perelman et al. 2011; The Marmoset Genome Sequencing and Analysis Consortium 2014). The overall repeat content of the *C. jacchus* genome, which represents the first sequenced NWM genome, is similar to other previously analyzed primate genomes, with transposable elements making up at least half the genome mass (International Human Genome Sequencing Consortium 2001; The Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; Locke et al. 2011; Carbone et al. 2014; The Marmoset Genome Sequencing and Analysis Consortium 2014).

The major drivers of repeat-driven genome expansion in primate genomes are retrotransposons. They move in a “copy-and-paste” fashion throughout the genome using an RNA intermediate (Ostertag and Kazazian 2001; Belancio et al. 2008; Cordaux et al. 2009) and can be subdivided into elements with long terminal repeats (LTRs), i.e., endogenous retroviruses, and those that lack LTRs. The latter are commonly referred to as non-LTR retrotransposons. In primates, the two major non-LTR elements and the largest contributors to genome expansion are long interspersed element 1, L1 (LINE1) and the primate-specific *Alu* element, a short interspersed element (SINE) (International Human Genome Sequencing Consortium 2001; Cordaux et al. 2009; Konkel et al. 2010). A full-length *Alu* element is ∼300 bp in length and terminates in an adenosine-rich tail (A-tail), a typical characteristic of non-LTR retrotransposons (Batzinger Deininger 2002). L1 predates the origin of primates, has been active throughout the radiation of primates, and represents the only known currently propagating autonomous non-LTR retrotransposon in primate genomes (Smit 1999; Ostertag and Kazazian 2001; Cordaux et al. 2009; Burns and Boeke 2012; Huang et al. 2012). LINEs and SINEs propagate in genomes via a process termed target-primed reverse transcription (TPRT) (Luan et al. 1993; Dewannieux et al. 2003) by utilizing the reverse transcriptase and endonuclease encoded by open reading frame 2 (ORF2) (Mathias et al. 1991; Feng et al. 1996; Cordaux et al. 2009). As a consequence of TPRT, classical retrotransposon insertions share hallmarks such as target site duplications (TSDs) (International Human Genome Sequencing Consortium 2001; Szak et al. 2002; Cordaux et al. 2009).

Following the divergence of two primate lineages, each lineage evolves uniquely and independently. Over time, this results in distinctive changes specific to each lineage; mobile elements are no different in this regard. Consequently, each lineage accumulates lineage-specific mobile element insertions and mobile element-mediated rearrangements (Cordaux et al. 2009; Konkel and Batzer 2010; Konkel et al. 2010). Moreover, mobile element subfamilies evolve uniquely in each lineage. In this respect, the lineage leading to the common marmoset had at least 35 million years (Perelman et al. 2011) of platyrrhine-specific evolution.
of an endonuclease cleavage site (Fig. 1D) and termination in an A-tail of varying length without a polyadenylation signal.

The majority of Platy-1 elements have an A-tail between 11 and 20 nucleotides, with 8.2% (57/427) having a pristine A-tail larger than 20 bp (Fig. 1C). About one-third (37.6%) of the 427 analyzed putative full-length Platy-1 elements have A-tails harboring nucleotides other than adenosine. The two longest A-tails of putative Platy-1 elements contained tetranucleotide microsatellites of 13 and 15 repeat units, respectively. Approximately 55% (41 of 75) of Platy-1 loci with an A-tail larger than 20 bp and at most one substitution harbored a Pol III termination signal within 100 bp of the downstream flanking sequence and 21 loci within 25 bp. Some of these loci may be source elements and, thus, may have contributed to the expansion of the Platy-1 repeat family.

Platy-1 subfamily reconstruction and evolution

Our Platy-1 subfamily structure analysis of all 2183 full-length Platy-1 elements, for which we used a majority rule approach, retrieved 62 Platy-1 subfamilies and 2275 elements (Supplemental Methods S1/Material S6; Supplemental Fig. S2). The higher number of Platy-1 elements in our final data set is a direct result of the subfamily reconstruction, as elements too diverged to be initially identified as a Platy-1 element were discovered with a more defined subfamily library. Of the 209 Platy-1 elements with an asterisk in our RepeatMasker output, indicating a second, higher scoring match with partial overlap (Smit et al. 2013–2015), we determined that the vast majority (95%) represent true Platy-1 insertions (Supplemental Methods S1). The remaining seven loci appear to be derived from a different repeat family. Thus, we estimate that the common marmoset genome contains 2268 full-length Platy-1 elements (Supplemental Table S3) translating to a density of approximately 0.77 Platy-1 elements per megabase.

As expected, the Platy-1 subfamily sizes vary (Table 1), with the largest subfamily (Platy-1-13) including 160 members and the smallest subfamily (Platy-1-16c) encompassing seven elements. More than two-thirds (43/62) of the subfamilies contained Platy-1 elements perfectly matching their respective consensus
sequence. Intriguingly, 26 subfamilies had three or more identical members, with one subfamily having 44 perfect Platy-1 elements. Altogether, almost 10% (224/2268) of Platy-1 elements are pristine, and more than a quarter (628/2268) have a divergence of 1.5% or less (Fig. 2). Our analyses show that both the average percent divergence and the divergence spectrum (i.e., element with lowest/highest substitution rate) varied considerably between subfamilies, indicating that some subfamilies were active for a longer time period than others (Table 1). Alternately, subfamilies with a wider spectrum may harbor more than one subfamily that cannot be distinguished due to high divergence and/or low number of elements.

Since non-LTR retrotransposons mutate at a neutral rate (Cordaux et al. 2006), the divergence from the consensus sequence can be used to approximate the age of a subfamily if the mutation rate is known and the molecular clock is constant. However, the molecular clock appears heterogeneous across primates (Li and Tanimura 1987; Hwang and Green 2004; Steiper et al. 2004; Kim et al. 2006; Steiper and Young 2006; Perelman et al. 2011), making time estimates more complicated. For mammals, an average neutral substitution rate of $2.2 \times 10^{-9}$ per base per year has been suggested (Kumar and Subramanian 2002); (age1) assumed mutation rate of $0.55 \times 10^{-9}$ (Lipson et al. 2015), integrating 29 yr/generation human-referenced estimate; (age2) assumed mutation rate of $7.53 \times 10^{-10}$ (Perez et al. 2013); (age3) assumed mutation rate of $0.55 \times 10^{-9}$ (Lipson et al. 2015), 6 yr generation time (Perez et al. 2013). These numbers should be regarded as rough estimates.

### Table 1. Platy-1 subfamily sizes and age estimates

<table>
<thead>
<tr>
<th>Subfamily</th>
<th>Length</th>
<th>#subfamily</th>
<th>min % div</th>
<th>max % div</th>
<th>avg % div</th>
<th>age1 (my)</th>
<th>age2 (my)</th>
<th>age3 (my)</th>
<th>age4 (my)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>104</td>
<td>55</td>
<td>4.9</td>
<td>14.7</td>
<td>66.4</td>
<td>24.5</td>
<td>267.8</td>
<td>73.2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>104</td>
<td>27</td>
<td>3.8</td>
<td>12.5</td>
<td>55.4</td>
<td>20.4</td>
<td>223.5</td>
<td>61.1</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>104</td>
<td>51</td>
<td>3.0</td>
<td>16.5</td>
<td>47.2</td>
<td>17.4</td>
<td>190.4</td>
<td>62.0</td>
<td></td>
</tr>
<tr>
<td>2b</td>
<td>104</td>
<td>35</td>
<td>7.8</td>
<td>13.7</td>
<td>61.5</td>
<td>22.7</td>
<td>248.4</td>
<td>67.9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>104</td>
<td>20</td>
<td>4.9</td>
<td>9.8</td>
<td>44.5</td>
<td>16.4</td>
<td>179.8</td>
<td>49.1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>104</td>
<td>19</td>
<td>1.9</td>
<td>9.4</td>
<td>22.2</td>
<td>8.2</td>
<td>89.5</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>4a</td>
<td>104</td>
<td>19</td>
<td>1.9</td>
<td>9.4</td>
<td>22.2</td>
<td>8.2</td>
<td>89.5</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>108</td>
<td>8</td>
<td>2.8</td>
<td>13.3</td>
<td>35.5</td>
<td>13.1</td>
<td>143.1</td>
<td>39.1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>104</td>
<td>15</td>
<td>1.2</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>6a</td>
<td>104</td>
<td>74</td>
<td>2.0</td>
<td>21.2</td>
<td>21.0</td>
<td>7.7</td>
<td>64.7</td>
<td>23.2</td>
<td></td>
</tr>
<tr>
<td>6b</td>
<td>104</td>
<td>80</td>
<td>1.9</td>
<td>9.4</td>
<td>22.2</td>
<td>8.2</td>
<td>89.5</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>6c</td>
<td>129</td>
<td>1</td>
<td>13.7</td>
<td>46.1</td>
<td>20.8</td>
<td>7.7</td>
<td>83.8</td>
<td>22.9</td>
<td></td>
</tr>
<tr>
<td>6d</td>
<td>104</td>
<td>12</td>
<td>1.9</td>
<td>9.4</td>
<td>22.2</td>
<td>8.2</td>
<td>89.5</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>6e</td>
<td>104</td>
<td>14</td>
<td>1.9</td>
<td>9.4</td>
<td>22.2</td>
<td>8.2</td>
<td>89.5</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>6f</td>
<td>27</td>
<td>0</td>
<td>10.9</td>
<td>5.8</td>
<td>26.1</td>
<td>9.6</td>
<td>105.5</td>
<td>28.8</td>
<td></td>
</tr>
<tr>
<td>6g</td>
<td>104</td>
<td>32</td>
<td>1.2</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>6h</td>
<td>104</td>
<td>1</td>
<td>10.8</td>
<td>4.4</td>
<td>20.0</td>
<td>7.4</td>
<td>80.7</td>
<td>22.1</td>
<td></td>
</tr>
<tr>
<td>6i</td>
<td>104</td>
<td>32</td>
<td>1.2</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>6n</td>
<td>104</td>
<td>61</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>6x</td>
<td>104</td>
<td>32</td>
<td>1.2</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>104</td>
<td>66</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>7a</td>
<td>104</td>
<td>17</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>104</td>
<td>2</td>
<td>10.9</td>
<td>5.8</td>
<td>26.1</td>
<td>9.6</td>
<td>105.5</td>
<td>28.8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>104</td>
<td>37</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>9a</td>
<td>104</td>
<td>30</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>9b</td>
<td>104</td>
<td>97</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>9c</td>
<td>104</td>
<td>45</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>9d</td>
<td>104</td>
<td>33</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>9e</td>
<td>104</td>
<td>13</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>104</td>
<td>43</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>10a</td>
<td>104</td>
<td>18</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>104</td>
<td>35</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>11a</td>
<td>104</td>
<td>16</td>
<td>0.4</td>
<td>25.6</td>
<td>26.4</td>
<td>9.7</td>
<td>106.7</td>
<td>29.2</td>
<td></td>
</tr>
</tbody>
</table>

The age was calculated assuming two different mutation rates. For the calculation, the average divergence of each subfamily was used. Gray data bars indicate relative values with respect to size or divergence estimates. (#) number; (my) million years ago; (age1) assumed mutation rate of $2.2 \times 10^{-9}$ (Kumar and Subramanian 2002); (age2) assumed mutation rate of $7.53 \times 10^{-10}$ (Perez et al. 2013); (age3) assumed mutation rate of $0.55 \times 10^{-9}$ (Lipson et al. 2015), integrating 29 yr/generation human-referenced estimate; (age4) assumed mutation rate of $0.55 \times 10^{-9}$ (Lipson et al. 2015), 6 yr generation time (Perez et al. 2013). These numbers should be regarded as rough estimates.
within and across all chromosomes, excluding insertions on
Our analysis of Platy-1 elements shows a scattered distribution
Platy-1 genomic distribution

2010; Conrad et al. 2011; Kong et al. 2012; Scally and Durbin
Alternate approaches utilize a comparison of current-day
genomes and precisely dated ancient genomes with estimate rang-
es from 1.1 to 1.7 × 10⁻⁸ per base per generation (Fenner 2005; Fu
et al. 2014). A recent study utilizes the fine-scale human recombi-
nation map for mutation rate calibration, resulting in a mutation
rate of 1.6 × 10⁻⁸ per base per generation or 0.55 ± 0.05 × 10⁻⁹ per
base per year (Lipson et al. 2015).

Non-CpG SNPs have been linked to the generation time and
body size of species (Kim et al. 2006; Perez et al. 2013), which can
vary considerably within primates. While a convergent slowdown
in primate sequence evolution rates has been proposed (Steiper
and Seiffert 2012), this could not be confirmed for platyrhines
(Perez et al. 2013) in part due to varying body size and generation
time within NWMS. Given that the neutral mutation rate within
and between species continues to be a subject of debate, we pro-
vide average age estimates for Platy-1 subfamilies assuming differ-
ent mutation rates (Table 1). The divergence spectrum of Platy-1
elements (Fig. 2) suggests a slow propagation rate with only a
few subfamilies active in early Platy-1 evolution based on the average sequence
ance from the consensus sequence of all 2268 full-length sequences. The subfamilies are color-coded
es from 1.1 to 1.7 × 10⁻⁸ × 10⁻⁸ per base per generation
ments of insertions is not proportional to the size of the chromosome
Our BLAST (Altschul et al. 1990) query of the whole database
(excluding C. jacchus) with consensus sequences of the oldest
Platy-1 subfamilies identified mostly matches in primates, includ-
ing human, indicating that Platy-1 may be derived from 7SL RNA.
Our multiple sequence alignment of the Platy-1 consensus se-
quences against 7SL RNA, FRAM, FLAM, and a selection of Alu/S
consensus sequences further supports that Platy-1 is derived from
7SL RNA (Fig. 5; Supplemental Fig. S4). More specifically, Platy-1
may be derived from an Alu/S element based on sequence identity.

Amplification dynamics of Platy-1 repeat family

In order to determine when Platy-1 elements started amplifying in
primates, we compared full-length Platy-1 candidate loci with
orthologous sequences retrieved from the human genome
(hg19), which represents the genome with the highest quality of
all primate genomes. We did not find convincing evidence of
Platy-1 elements shared between common marmoset and human.
Given that some of the oldest elements may have accumulated
mutations to a degree that they are no longer recognized as
Platy-1-derived, we next screened the human genome for the pres-
ence of Platy-1 elements through a RepeatMasker analysis, reveal-
ing 27 full-length Platy-1 candidate loci (Methods; Supplemental
Table S4). We discarded 14 loci based on manual analysis because
of lack of TSDs and high sequence similarity upstream of or down-
stream from the putative Platy-1 sequence, with other Platy-1 can-
didate loci, an indication for nonproper Platy-1 insertions after
ruling out duplication events.

The remaining 13 putative Platy-1 loci appear to have been active early in Platy-1 evolution based on the average sequence
divergence and absence of additional Platy-1 insertions in the rhe-
sus macaque genome (rheMac3) (Supplemental Table S5). For 10
of the catarrhine-specific Platy-1 candidate loci, we were able to
identify TSDs, A-tails of varying length, and endonuclease cleav-
age sites (Supplemental Table S6), suggesting insertion through
retrotransposition. Two additional loci likely represent Platy-1 in-
sertions; however, these insertions occurred into adenosine-rich
sequences, obscuring accurate TSD identification. For one putative
Platy-1 locus (Chr 16: 74696055–74696158 [hg19]), we could not

Platy-1 genomic distribution

Our analysis of Platy-1 elements shows a scattered distribution
within and across all chromosomes, excluding insertions on “chromo-
some” Random (Fig. 4; Supplemental Fig. S3A). All chromo-
somes except Y contain Platy-1 insertions; this is not unexpected,
as the Y Chromosome is relatively small in size, and the genome
of a female marmoset contains only a small fraction (3.5 Mb) of Y
chromosomal DNA from her fraternal twin due to chimerism. The
intrachromosomal distribution appears to be random, supported
by a random insertion model (Supplemental Methods S3; Sup-
plemental Fig. S3B). In contrast, we rejected a random insertion
model for Platy-1 distribution across the genome (i.e., the number of
insertions is not proportional to the size of the chromosome)
based on the results of a χ² analysis (χ² = 51.35, df = 25, P =
0.000656). Hence, Platy-1 elements are not evenly distributed
across all chromosomes. Chromosome 4 exhibited a much lower density and
Chromosomes 18, 19, and 22 a higher density (Fig. 4). An uneven distribution
of non-LTR retrotransposons across the chromosomes has been observed previ-
ously (e.g., SVA or Alu elements) (International Human Genome Sequencing

Origin of the Platy-1 repeat family

Next, we investigated the origin of the Platy-1 repeat family. The internal pro-
moter regions of known SINEs seem to be derived from tRNA, 7SL RNA, or SS
RNA. Another alternative, though less likely, is the origin of a mobile element
through horizontal transfer (Gilbert et al. 2012). To confirm that Platy-1 was
not derived from tRNA, we performed a tRNA-Scan (Lowe and Eddy 1997;
Schattner et al. 2005) and confirmed

Figure 2. Platy-1 evolution in NWMs. The histogram shows the Platy-1 distribution based on the diver-
gence from the consensus sequence of all 2268 full-length sequences. The subfamilies are color-coded
based on subfamily affiliation and grouped together based on age. The divergence from the respective
consensus sequence was retrieved from RepeatMasker and is shown on the x-axis. The y-axis shows the
number of elements with the indicated divergence. The plot is generated with custom BioPython scripts
and the Vega + D3 Vincent wrapper/package (Bostock et al. 2011) (http://github.com/wrobstory/
vincent).
identify TSDs. This locus terminated in an A-tail immediately followed by a stretch of thymine nucleotides. The majority (seven of 13) of the putative Platy-1 loci were absent from the common marmoset genome. Four loci, including the locus without TSDs, were shared with common marmoset; none of these were present in our common marmoset data set. For two loci, we could not unambiguously determine lineage specificity due to (partial) absence of our common marmoset data set. For two loci, we could not unambiguously determine lineage specificity due to (partial) absence of our common marmoset data set. For two loci, we could not unambiguously determine lineage specificity due to (partial) absence of our common marmoset data set. For two loci, we could not unambiguously determine lineage specificity due to (partial) absence of our common marmoset data set.

Occasionally, we encountered amplicon sizes of unexpected length (i.e., a size different from the predicted filled or empty amplicon size) in at least one NWM species. Given the different amplicon size in one or more species, these loci were easily identified. Intriguingly, we identified one locus (on Chr 13, locus 157) for which the amplicon pattern, i.e., a PCR product of the same approximate size as the amplicon containing a Platy-1 element, suggested a different relationship than the majority of the tested loci, which were in agreement with previous phylogenies (Ray et al. 2005; Osterholz et al. 2009). More specifically, the insertion suggested a close relationship between marmosets and the family Atelidae. DNA sequencing of this locus revealed a truncated Alu insertion of similar length as the Platy-1 insertion (Supplemental Fig. S5). In addition, we identified 31 Callithrichidae- and one Cebidae-specific Platy-1 insertions. Our analyses of the common marmoset-specific insertions on the population panel of 24 common marmosets (Supplemental Table S1A) show that some (three loci) Platy-1 elements are polymorphic within C. jacchus (Supplemental Table S7).

Discussion
We identified a new SINE family, Platy-1, in the common marmoset genome that is most likely derived from Alu. A full-length Platy-1 element is only ~100 bp in length, making it the shortest SINE in primates. A few active Platy-1 subfamilies contain deletions, illustrating that even shorter SINEs can maintain retrotranspositional activity. Although the A box shows considerable variation during the evolution of Platy-1, the B box has primarily remained stable. This, taken together with the presence of TSDs, an endonuclease cleavage site, and termination in an A-tail without polyadenylation signal indicates transcription of Platy-1 by Pol III and insertion into the genome via TPRT by the enzymatic machinery of L1. The lack of a polyadenylation signal prior to the A-tail also indicates that the A-tail is derived from the source element as previously reported for Alu elements (Batzer et al. 1990; Deininger and Batzer 1993; Shaikh and Deininger 1996). The absence of TSDs for ~10% of Platy-1 elements is likely primarily caused by decay due to the age of some of the insertions. Alternate reasons include insertion in sequences with simple repeat characteristics and insertion of other mobile elements into the tail immediately upstream of or downstream from Platy-1. A fraction of these insertions may also have occurred through endonuclease-independent insertion mechanisms as described for Alu and L1 elements (Morrish et al. 2002; Gilbert et al. 2005; Sen et al. 2007; Srikanta et al. 2009).
Figure 4. Platy-1 genomic distribution. The expected (yellow) and actual (purple) Platy-1 distributions across all chromosomes (excluding Chr Random and Chr Y) are illustrated. In addition, the density per megabase is shown for each chromosome. Due to omission of putative Platy-1 loci on Chr Random, 2221 full-length elements were included in this analysis.

Figure 5. Alignment of Platy-1 with 7SL RNA and Alu elements. Shown is a multiple sequence alignment using MUSCLE (Edgar 2004) followed by manual curation of the oldest Platy-1 subfamilies with 7SL RNA, a selection of Alu consensus sequences, and FLAM. The alignment is visualized with AliView (Larsson 2014). Dashes indicate absence of the sequence. Also illustrated are the A box (consensus sequence: TRGYnnAnnG) and B box (consensus sequence: GWTCRAnnCc). The tail of the Platy-1 sequence aligned equally well to the regions prior to the middle A-rich region and the 3′ end of an Alu element. In the latter case, the deletion may have been caused by recombination between homologous sequences. This alignment assumes that the element terminates at the middle A-rich region. An alternate alignment is provided in Supplemental Figure S4.
activity of Alu (Roy-Engel et al. 2002; Bennett et al. 2008; Comeaux et al. 2009); this is likely true for Platy-1 as well. Based on this and the analysis of 427 elements, we estimate that ~8% of elements have a higher chance of being retrotranspositionally active (pristine A-tail and length >20 bp). Of these, more than half contained a Pol III termination signal downstream within 100 bp and 28% within 25 bp. The latter elements have the highest probability of being source elements, as a Pol III termination signal immediately downstream from an Alu element has been positively correlated with Alu retrotransposition efficiency in vitro (Comeaux et al. 2009). On the other hand, more than one-third of the inspected Platy-1 A-tails harbored nucleotide substitutions, with many of them containing microsatellites. This is in agreement with previous findings that A-tails of non-LTR retrotransposons, particularly Alu elements, represent seeds for microsatellite formation (Arcot et al. 1995; Nadir et al. 1996; Fungtammasan et al. 2012; Grandi and An 2013; Grandi et al. 2013) and indicates a role of Platy-1 in the birth of microsatellites, and consequently, to structural variation beyond insertional mutagenesis.

Figure 6. Phylogenetic distribution of Platy-1. (A) Four agarose gel chromatographs of our locus-specific phylogenetic analyses. An upper fragment indicates presence of a Platy-1 insertion; a lower fragment, absence. The vertical lines from left to right separate the outgroups from NWMs, Atelidae from Cebidae, and Cebidae from Pithecia. The gel chromatographs show (from left to right): (A) 100 bp ladder; (B) TLE; (C) human; (D) common chimpanzee; (E) African green monkey; (F) woolly monkey; (G) spider monkey; (H) red howler monkey; (I) common marmoset; (J) pygmy marmoset; (K) tamarin; (L) capuchin monkey; (M) squirrel monkey; (N) owl monkey; (O) titi; (P) saki. (For more detailed information regarding the species used, please see Supplemental Table S1A.) The top gel image shows a Platy-1 insertion shared across all NWMs. The gel chromatograph below shows an insertion specific to Callithricinae, which is followed by a marmoset-specific insertion. The gel chromatograph on the bottom shows a common marmoset-specific Platy-1 insertion. (B) The phylogenetic results for our informative loci are shown in a pie chart: (Marm) marmosets; (cM) common marmoset; (fp) false positive; (NWM) New World monkey; (Call) Callithricinae; (Ceb) Cebidae.

Lineage-specific repeat families containing Alu sequence have been identified in some primate lineages. For example, the hominoid-specific SVA element (Ostertag et al. 2003; Wang et al. 2005; Damert et al. 2009; Hancks and Kazazian 2010) and the gibbon-specific LAVA element (Carbone et al. 2012) harbor Alu sequence as a component of these composite mobile elements. In the Galago crassicaudatus, a lineage-specific SINE derived from an Alu element, into which a Type III element inserted, has been reported (Daniels and Deininger 1985; Roos et al. 2004). That said, Platy-1 not only represents the first lineage-specific SINE in NWMs, but also the first SINE in primates that is solely derived from an Alu element.

Platy-1 propagation

Overall, our results support the birth of Platy-1 around the time of divergence of catarrhines and platyrrhines, as both lineages show at least some evidence for Platy-1 mobilization. However, Platy-1 established itself as a repeat family only in NWMs. This raises the question why this may be the case, especially since the potential founding element for the Platy-1 repeat family, an element with a perfect A-tail followed immediately by a Pol III termination signal, appears to be present in both lineages. It may be that this sequence lost its propagation properties in the lineage leading to human prior to the rise of daughter elements with propagation capabilities. Alternatively, modifications in the upstream sequence (e.g., nucleotide substitutions or sequence rearrangements) may have played a role, as it is known that the upstream sequence of Alu elements is crucial for its retrotransposition capabilities (Roy et al. 2000b). However, the exact upstream sequence requirements remain elusive.

Platy-1 provided the unique opportunity to study its propagation dynamics from its birth up to the present—spanning more than 40 million years of evolution. The mammalian neutral substitution rate seems to better reconstruct the evolution of Platy-1 (Table 1; Supplemental Table S4) with respect to primate radiation estimates. This does not necessarily imply that the mammalian substitution rate is more accurate, as the relatively short length of the Platy-1 element does not allow for a fine-scale divergence resolution. Also, its high GC content, resulting in above-average Cpg sites, may contribute to the uncertainty of the average substitution rate and (consequently) age estimate of Platy-1 subfamilies. Moreover, the reconstruction of older subfamilies is less accurate due to mutation accumulation. Gene conversion represents another factor that would result in a higher divergence of Platy-1 sequences and as a result could impact age estimates of subfamilies. However, we believe that gene conversion plays a limited role for
Platy-1 elements given the lower number of these elements in the genome and the short length of Platy-1; both factors generally reduce the risk of gene conversion. Moreover, gene conversion events should be relatively easily detected, and these elements would have been grouped as a separate subfamily with hybrid sequence characteristics during manual curation, similar to those found in human Alu gene conversion events (Kass et al. 1995; Roy et al. 2000a).

Following an initially slow mobilization rate, Platy-1 propagation seems to have continuously increased until propagation reached a peak that coincided with the rise of the marmoset ancestor, suggesting a recent Platy-1 expansion with several retrotransposition–competent subfamilies propagating in parallel. This is in stark contrast to the early Platy-1 evolution, in which propagation was primarily linear (Fig. 3) and more closely resembled the typical subfamily evolution of LINES. This is likely a result of the low number of source elements early in Platy-1 evolution—initially, of a single subfamily. In contrast, the more recent history of Platy-1 follows a star-like pattern, typical for the propagation of Alu elements (Cordaux et al. 2006). Most recently, Platy-1 retrotransposition has somewhat slowed based on the relatively small number of polymorphic Platy-1 insertions within common marmoset populations (Supplemental Table S7). However, Platy-1 has been active very recently and is likely still propagating in common marmosets.

The timing of the recent Platy-1 expansion differs from the Alu propagation dynamics in the lineage leading to the common marmoset (The Marmoset Genome Sequencing and Analysis Consortium 2014). The history of Alu elements shows an exceptional retrotransposition peak about 40 million years ago (International Human Genome Sequencing Consortium 2001; The Marmoset Genome Sequencing and Analysis Consortium 2014), coinciding with the estimated divergence of the NWM lineage from the ancestral anthropoid lineage about 35–47 million years ago. This was followed by a decline of the Alu propagation rate both in Platyrhini (The Marmoset Genome Sequencing and Analysis Consortium 2014) and Catarrhini (International Human Genome Sequencing Consortium 2001). A second smaller Alu retrotransposition peak occurred in the NWM lineage, leading to the common marmoset in more recent history, and has subsequently declined (The Marmoset Genome Sequencing and Analysis Consortium 2014). Based on the divergence estimates, Platy-1 propagation peaked more recently than Alu retrotransposition in the NWM lineage and coincided with the rise of the common ancestor of marmosets. The more recent amplification success of Platy-1 may be the result of different factors, including a higher number of source elements in conjunction with escape from the host response against Platy-1 elements. Although speculative, the decline of the Alu mobilization rate could at least partially be caused by the higher number of retrotransposition–competent Platy-1 elements, as both Platy-1 and Alu likely compete for the enzymatic machinery of L1. In general, the interplay of different factors, including number of retrotransposition–competent elements, preference of the L1 enzymatic machinery, and host factors, all affect the amplification dynamics of non-LTR retrotransposons and result in varying retrotransposition rates over time.

Taken together, we identified and characterized a new SINE in the NWM lineage leading to the common marmoset. The birth of Platy-1 coincided with the rise of NWMs, allowing for the investigation of the Platy-1 propagation dynamics throughout the radiation of NWMs. The current study was performed through the lens of the common marmoset. It remains to future studies to determine the fate of Platy-1 in other NWM species.

Methods

Analysis of the original Platy-1 element

We retrieved orthologous nucleotide sequences for the original multispecies sequence alignment on Chromosome 3 using BLAT (Kent 2002) from the common marmoset (callJac3.2), human (hg19), chimpanzee (panTro2/panTro4), orangutan (ponAbe2), and rhesus macaque (rheMac2/rheMac3) genome assemblies. Assisted by the ClustalW function of BioEdit (Hall 1999), we performed a multiple species alignment. Next, we queried the original Platy-1 sequence against the C. jacchus genome (callJac3.2) using BLAT, as well as against all other aforementioned genome assemblies and the squirrel monkey draft genome assembly (saiBol1). We also queried the original sequence against the whole database using BLASTN (Altschul et al. 1990) to determine if the sequence was identified in species other than the common marmoset. To assess if the sequence was derived from a known repeat, we checked the sequence with RepeatMasker (Smit et al. 2013–2015).

Subfamily reconstruction

We selected the 10 best matches from our C. jacchus BLAT query and reconstructed a preliminary query sequence using a majority rule approach. We included this sequence in the library of our local RepeatMasker analysis (Supplemental Methods S1) and retrieved all full-length Platy-1 hits, defined as start position <4 bp and end position not shorter than two nucleotides prior to the A-tail. Following multiple sequence alignments with ClustalW (Hall 1999) and/or MUSCLE (Edgar 2004), and together with manual curation, the Platy-1 elements were sorted based on presence/absence of SNPs and short indels (deletions/insertions of nucleotides) using a majority rule approach (for further information, please refer to Supplemental Methods S1). Following the creation of Platy-1 consensus sequences, we RepeatMasked the C. jacchus assembly and determined for each subfamily (1) if the minimum requirements for the presence of a subfamily was met (i.e., minimum number of elements); and (2) if we could identify additional subfamilies. This step was repeated until we identified and confirmed all subfamilies.

Phylogenetic subfamily tree construction

We reconstructed the subfamily evolution using DNAdist, neighbor, and drawgram from the Phylib v3.695 (Felsenstein 1989) analytics suite for generation of a neighbor-joining tree using the Kimura-2-parameter model. As input, we used a file containing all Platy-1 subfamilies. We manually moved subfamilies with deletions to the appropriate locations in the tree because these subfamilies evolved most likely from each other, as the independent recurrence of the same deletion is highly unlikely. In parallel, we performed a network analysis using Network version 4.612 (Supplemental Methods S2; Bandelt et al. 1999). To avoid misplacing subfamilies, we removed all deletions except for one nucleotide from the respective consensus sequences.

Platy-1 sequence feature analysis

For our Platy-1 sequence analyses, we scrutinized the elements as well as sequences immediately upstream and downstream for the presence of common sequence features. Specifically, A-tail length and composition, TSDs, and endonuclease cleavage sites were manually determined and recorded. For the determination of
TSDs, we allowed for the presence of up to four SNPs, because TSDs decay over time—likely at a neutral rate similar to non-LTR retrotransposons (Cordaux et al. 2009). We excluded loci without TSDs from our downstream analyses and used Weblogo (Crooks et al. 2004) to visualize the endonuclease cleavage site.

**Platy-1 genomic distribution**

We calculated the density of full-length Platy-1 elements across the *C. jacchus* genome by counting Platy-1 elements per chromosome divided by the number of megabases of nucleotide sequences for each chromosome (excluding ChrRand, which, reduced the data set from 2268 to 2221 putative Platy-1 loci) and performed a χ² analysis. We utilized the GATK software (McKenna et al. 2010) to calculate the GC content of the flanking sequence. We also determined the intrachromosomal Platy-1 distribution and performed simulations of a random intrachromosomal distribution model (Supplemental Methods S3).

**Determination of Platy-1 lineage specificity**

To determine when Platy-1 originated, we queried all full-length Platy-1 elements, including 500 bp of flanking sequence, against the human genome (hg19) using a local BLAT installation (Kent 2002) and retrieved the resulting sequences. We aligned the human sequence against the *C. jacchus* sequence using MUSCLE (Edgar 2004) for matches in human of <10,000 bp in length and manually determined the presence of Platy-1 in catarrhines. In addition, we RepeatMasked the human genome (hg38) with a custom library and retrieved all full-length Platy-1 elements, including 500 bp of flanking sequence. Next, we performed the same analysis as described above by querying against the common marmoset genome and checked all putative Platy-1 loci for the presence of typical non-LTR features. We also RepeatMasked the rhesus macaque (rheMac3) genome, retrieved all full-length Platy-1 insertions, and compared Platy-1 insertions from the rhesus macaque genome with hits from the human genome.

**Age estimate calculation**

Based on neutral mutation rate estimates of 8.5 × 10⁻¹⁰—6.06 × 10⁻¹⁰ per base per generation for crown Platyrrhini and of 9.07 × 10⁻¹⁰—6.47 × 10⁻¹⁰ for crown Cebidae (Perez et al. 2013), we assumed an average substitution rate of 7.53 × 10⁻¹⁰ per base per generation for our subfamily age estimates. We used 8 yr as an approximate generation time given that marmosets have a generation time of about 6 yr, and the generation likely was longer earlier in primate radiation. We also calculated the average age of the elements using an estimated neutral substitution rate for mammals of 2.2 × 10⁻⁹ per year per base (Kumar and Subramanian 2002) and one derived from human studies using 0.55 × 10⁻⁸ and varying generation times (Lipson et al. 2015).

**Oligonucleotide primer design and PCR analyses**

For primer design, we retrieved the Platy-1 element plus 600 bp of flanking on either side from the *C. jacchus* genome using BLAT (Kent 2002). Orthologous sequences to the flanking sequence were retrieved from the human (hg19), chimpanzee (panTro2/4), orangutan (ponAbe3), and rhesus macaque (rheMac2/3) genome assemblies using BLAT, and a multiple species alignment was performed with ClustalW (Hall 1999). We designed primers with Primer3 (Rozen and Skaltsky 2000) using the default settings with the following exceptions: only mononucleotide repeats with up to four consecutive identical nucleotides were permitted, the minimum nucleotide number for primers was increased to 20, the annealing temperature range was set to 57°C–61°C, and the allowed PCR amplicon size was increased to 1400 bp. All primer pairs that passed our downstream screening (Supplemental Methods S4) and wet bench-based phylogenetic analyses (for species and origin of DNA, see Supplemental Table S1A), we performed PCRs in a 96-well format using a BioRad iCycler thermocycler in a final volume of 25 µL. Each PCR reaction contained 25 ng of template DNA; 200 nM of each oligonucleotide primer; 1.5 mM MgCl₂; 1X PCR buffer (50 mM KCl; 10 mM TrisHCl, pH 8.3); 0.2 mM dNTPs; and 2 units Taq DNA polymerase. PCR reactions were performed using the following conditions: initial denaturation for 90 sec at 94°C, followed by 32 cycles of denaturation for 30 sec at 94°C, annealing at optimal temperature for 20 sec, and extension for 30 sec at 72°C. PCRs were terminated with a final extension for 3 min at 72°C. The PCR amplicons (20 µL of each PCR product) were size fractionated in a horizontal gel chamber on a 2% agarose gel containing 0.1 µg/mL ethidium bromide for 50 min at 200 V. DNA fragments were visualized with UV-fluorescence and images were saved using a BioRad ChemiDoc XRS imaging system.

**Data access**

Sequencing data generated for this study have been submitted to GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) under accession numbers KT427526–28 and are shown in the alignment of Supplemental Figure S5.

**Acknowledgments**

We thank Jerilyn A. Walker and Catherine C. Fontenot for discussion, support, and advice. This research was supported by the National Institutes of Health R01 GM59290 (M.A.B.), E.L.A. and S.S. were supported in part by a grant to Louisiana State University from the Howard Hughes Medical Institute (HHMI) through the Precollege and Undergraduate Science Education Program. S.S. was also supported in part by the Louisiana Board of Regents Supervised Undergraduate Research Experience (SURE) program (B.U., M.K.K., M.A.B.). The computational analyses were in part supported through National Science Foundation grant CNS-1126739 (B.U., M.A.B., M.K.K.). Portions of this research were conducted with high-performance computing resources provided by LONI and LSU HPC.

**Author contributions:** M.K.K. found the element, designed the studies, performed analyses, and wrote the manuscript. B.U. performed computational analyses and was involved in the writing of the manuscript; E.L.A. and S.S. performed phylogenetic and population genetic analyses; E.L.A., S.S., and S.A.B. were involved in Platy-1 characterization. R.H. and A.F.A.S. were involved in the identification and characterization of the element. M.A.B. provided experimental input, support, and helped write the manuscript.

**References**


Platy-I, a New World monkey-specific repeat family


Received September 3, 2015; accepted in revised form February 23, 2016.
Discovery of a new repeat family in the *Callithrix jacchus* genome


*Genome Res.* 2016 26: 649-659 originally published online February 25, 2016
Access the most recent version at doi:10.1101/gr.199075.115

---

**Supplemental Material**

http://genome.cshlp.org/content/suppl/2016/04/07/gr.199075.115.DC1

**References**

This article cites 88 articles, 24 of which can be accessed free at:
http://genome.cshlp.org/content/26/5/649.full.html#ref-list-1

**Creative Commons License**

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

**Email Alerting Service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

---

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2016 Konkel et al.; Published by Cold Spring Harbor Laboratory Press