The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage

John K. Pace, II and Cédric Feschotte

*Genome Res.* published online Mar 5, 2007; Access the most recent version at doi:10.1101/gr.5826307

Published online March 5, 2007 in advance of the print journal.

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](http://www.genome.org/subscriptions/)

Notes

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to: [http://www.genome.org/subscriptions/](http://www.genome.org/subscriptions/)

© 2007 Cold Spring Harbor Laboratory Press
The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage

John K. Pace II and Cédric Feschotte1

Department of Biology, University of Texas at Arlington, Arlington, Texas 76019, USA

Class 2, or DNA transposons, make up ~3% of the human genome, yet the evolutionary history of these elements has been largely overlooked and remains poorly understood. Here we carried out the first comprehensive analysis of the activity of human DNA transposons over the course of primate evolution using three independent computational methods. First, we conducted an exhaustive search for human DNA transposons nested within L1 and Alu elements known to be primate specific. Second, we assessed the presence/absence of 794 human DNA transposons at orthologous positions in 10 mammalian species using sequence data generated by the NIH/NISC ENCODE Project. These two approaches, which do not rely upon sequence divergence, allowed us to classify DNA transposons into three different categories: anthropoid specific (40–63 My), primate specific (64–80 My), and eutherian wide (81–150 My). Finally, we used this data to calculate the substitution rates of DNA transposons for each category and refine the age of each family based on the average percent divergence of individual copies to their consensus. Based on these combined methods, we can confidently estimate that at least 40 human DNA transposon families, representing ~98,000 elements (~33 Mb) in the human genome, have been active in the primate lineage. There was a cessation in the transpositional activity of DNA transposons during the later phase of the primate radiation, with no evidence of elements younger than ~37 My. This data points to intense activity of DNA transposons during the mammalian radiation and early primate evolution, followed, apparently, by their mass extinction in an anthropoid primate ancestor.

[Supplemental material is available online at www.genome.org.]
human genome sequence. This study concluded that, “there is no evidence for DNA transposon activity in the past 50 My in the human genome” (Lander et al. 2001). However, this statement should be taken with caution, since this conclusion was drawn solely from the average level of nucleotide divergence of individual copies to their reconstructed family consensus sequence. This approach has proven generally reliable to date the most prominent bursts of TE amplification and for comparing groups of elements that evolve at the same or almost the same rate. However, due to rapid fluctuations in substitution rates during the mammalian and primate radiation (Goodman 1985; Yi et al. 2002) and possible variations in the substitution rate of different types of TEs (for example, see Xing et al. 2004), the results of these analyses should be interpreted carefully. In addition, the method is entirely dependent on the reconstruction of an accurate consensus sequence, a process that is sensitive to the number and genomic distribution of the elements. Since human DNA transposons are anticipated to be of relatively ancient origin and have likely diversified over the entire course of mammalian evolution, sequence divergence should not be the sole method used for accurately dating these types of elements. Furthermore, there has been no published effort to characterize the tempo and mode of accumulation of every DNA transposon family in any mammalian genome. Therefore an accurate and detailed picture of DNA transposon history in humans and other mammals is still lacking.

Here we present the first detailed analysis of the age of the 125 DNA transposon families currently recognizable in the human genome. In particular, we sought to evaluate which human DNA transposon families were actively transposing during primate evolution. To this end, we used a combination of three independent computational methods, two of which do not rely upon sequence divergence. We estimate that at least 40 families of DNA transposons were active during the primate radiation. We conclude that ~98,000 individual elements were added to the primate genome in the last ~80 My of evolution. Eleven of these families, or ~23,000 individual elements, inserted into the primate genome between the split of prosimian primates and new world monkeys (~40 to 63 Mya). However, we found no evidence that any human DNA transposon family was active within the last ~37 My of primate evolution. Our results suggest an intriguing history of intense activity of diverse DNA transposons during the first half of the primate radiation, followed by a striking cessation of transposition activity in an anthropoid primate ancestor and no detectable germ-line reinfiltation of the primate lineages leading to humans over the last 37 My.

**Results**

**Census of DNA transposons in the human genome**

We began our investigation by assessing the diversity and copy number of all DNA transposon families currently recognized in the human genome. Copy numbers were calculated from the RepeatMasker tables of the May 2004 assembly of the human genome, available through the UCSC Genome Browser (http://genome.ucsc.edu). In agreement with previous reports (Smit and Riggs 1996; Smit 1999; Lander et al. 2001), we found that two superfamilies, hAT and Tc1/mariner, are predominant in the human DNA transposon population (Table 1). hAT elements account for approximately two-thirds of the census and more than half of the 125 families. Human Tc1/mariner elements account for one-third of the population and can be divided into three evolutionary distinct lineages: pogo-like, mariner-like, and Tc2-like (Smit and Riggs 1996; Robertson 2002). The former is the most abundant and diversified lineage, and includes eight families of transposase-encoding Tigger elements and 22 related MITE families. The prevalence of nonautonomous MITEs (74% of the total number of DNA TEs) over transposase-encoding elements (26%) is particularly striking in the human genome, and this phenomenon affects all superfamilies (Table 1). It is also a characteristic of the DNA transposon population of plants and nematodes (Feschotte et al. 2002).

**Analysis of DNA transposons nested into other elements**

In order to obtain a first assessment of human DNA transposon families that were active during the primate radiation, we took advantage of the fine-scale evolutionary histories of L1 and Alu elements in the primate lineage produced by others. We used these primate-specific families as historical markers for dating DNA transposons. We reasoned that any DNA transposon inserted or nested within a primate-specific L1 element (Khan et al. 2006) or a primate-specific dimeric Alu element (Kapitonov and Jurka 1996) should itself be primate specific. Using a Perl script and data from the UCSC Genome Browser RepeatMasker tables, we conducted an exhaustive search of human L1 and Alu retrotransposons that had suffered a nested insertion of a DNA transposon. A DNA transposon was considered to be nested within one of these retrotransposons if: (1) the upstream and downstream retrotransposon fragments were within 50 bp of the 5’ and 3’ ends of the DNA.

### Table 1. Summary of currently recognizable DNA transposons in the human genome with copy number >100

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Families</th>
<th>No. of families</th>
<th>Copy No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>hAT</td>
<td>Autonomous</td>
<td>Blackjack, Charlie, Cheshire, Zaphoed</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Nonautonomous</td>
<td>Arthur1, FordPrefect, MERT02, MERT106, MERT107, MERT112, MERT113, MERT115, MERT117, MERT119, MER1, MER20, MER3, MER30, MER33, MER45, MER58, MERS, MER63, MER69, MER81, MER91, MER94, MER96, MER99, ORSL</td>
<td>52</td>
</tr>
<tr>
<td>MuDR</td>
<td>Nonautonomous</td>
<td>Ricksha</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>3</td>
<td>985</td>
</tr>
<tr>
<td>piggyBac</td>
<td>Autonomous</td>
<td>Looper</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Nonautonomous</td>
<td>MER75, MER85</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>4</td>
<td>2,090</td>
</tr>
<tr>
<td>Tc1/mariner</td>
<td>Autonomous</td>
<td>HSMAR, Tigger, Kanga</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Nonautonomous</td>
<td>MADE, MARN, MERT04, MER2, MERT44, MERT46, MERT53, MERE, MERS, MERB2, MERB97</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>45</td>
<td>108,038</td>
</tr>
<tr>
<td>Unknown</td>
<td>MER103, MER105</td>
<td>2</td>
<td>7,567</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2</td>
<td>7,567</td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td>125</td>
<td>382,872</td>
<td></td>
</tr>
</tbody>
</table>
transposon, (2) the orientation of the upstream and downstream retrotransposon fragments were the same, and (3) the 3′ end of the upstream retrotransposon fragment and the 5′ end of the downstream retrotransposon fragment were within 20 bp of each other according to matching positions in their consensus nucleotide sequence.

This analysis revealed the presence of elements representing 10 distinct DNA element families that were inserted into primate-specific L1s (Table 2). Each of these insertions were validated by visual inspection of expected target-site duplications (TSDs) flanking the DNA transposon insertion based on alignment to the family consensus sequence (one example per family is shown in Table 2). The only exception was MER107 elements, whose insertion into L1 created a short deletion (from 11 to 19 bp) at the integration site accompanied by the consortion of unrelated “filler” DNA (11–27 bp), instead of the 8-bp TSD canonical of hAT elements (see Supplemental Fig. 1). The youngest L1 element that suffered a DNA transposon insertion belong to the LIPABA family, estimated to be between 42 and 50 My old (Khan et al. 2006; Supplemental Table 1). L1PA8A elements suffered insertions from two separate MADE1 elements previously proposed to be among the youngest DNA element families (Lander et al. 2001; Table 2).

In each case, we found that the length of the nested DNA transposons was similar to the length of its respective consensus sequence and that the nucleotide divergence of the L1 copy to its consensus sequence was congruent with the age of the corresponding L1 subfamily. For example, four distinct MER85 elements were found inserted into L1PA10 elements (one example is reported in Table 2). A TSD of the tetranucleotide sequence TTAA, a hallmark of piggyBac transposons, could be associated with each insertion. The length of the nested MER85 elements ranged from 126 to 130 bp, which is in good agreement with the length of the family consensus sequence (140 bp). The nested MER85 copies differ by 6.3% to 9.3% from their consensus sequence, while the average divergence for the family is 7.3% (as calculated from the May 2004 Reptamasker files from the UCSC Genome Browser, see Methods). The L1PA10 elements that suffered insertions by MER85 elements were 8.5% to 23.9% divergent from their consensus (with three of the four being 15.3% diverged or less), which is consistent with the average 14.7% pairwise divergence of the L1PA10 subfamily (Khan et al. 2006). These data indicate that both nested DNA transposon and disrupted L1 elements are fully representative of their respective families.

We found that six of the 10 DNA element families that had copies inserted into primate-specific L1 elements also comprise copies nested into dimeric Alu elements (Table 2), all of which are known to be primate specific (Kapitonov and Jurka 1996). One family, MER1A, was found to be nested within a dimeric Alu (Alu)b but was not found to be nested within a primate-specific L1. Here again, each DNA transposon insertion was validated as a bona fide transposition event by the presence of expected TSDs.

### Table 2. DNA element insertions into primate-specific L1s, Alu, and LTRs

<table>
<thead>
<tr>
<th>DNA element</th>
<th>Superfamily</th>
<th>LINE</th>
<th>Target site duplication</th>
<th>Age of LINE</th>
<th>Avg % div. of LINE</th>
<th>Position (May 2004 Assembly)</th>
<th>No. of nestings found</th>
</tr>
</thead>
<tbody>
<tr>
<td>MER85</td>
<td>piggyBac</td>
<td>L1PA10</td>
<td>TAA/TTAA</td>
<td>46.4*</td>
<td>13.03</td>
<td>chr1:87694072–87694198</td>
<td>4</td>
</tr>
<tr>
<td>MER107</td>
<td>hAT</td>
<td>L1PB3</td>
<td>None</td>
<td>73.5*</td>
<td>17.84</td>
<td>chr2:166664067–166664265</td>
<td>1</td>
</tr>
<tr>
<td>MER75</td>
<td>hAT</td>
<td>L1M2</td>
<td>Ta/TA</td>
<td>65.8*</td>
<td>15.75</td>
<td>chr7:144252811–144253077</td>
<td>1</td>
</tr>
<tr>
<td>MADE1</td>
<td>Tc1/meriner</td>
<td>L1PA8A</td>
<td>TA/TT</td>
<td>41.7*</td>
<td>12.92</td>
<td>chr6:91273479–91273559</td>
<td>2</td>
</tr>
<tr>
<td>HMSAR1</td>
<td>Tc1/meriner</td>
<td>L1M1</td>
<td>TA/TA</td>
<td>61.6*</td>
<td>15.58</td>
<td>chrX:85338493–85339772</td>
<td>2</td>
</tr>
<tr>
<td>Charlie3</td>
<td>hAT</td>
<td>L1PA16</td>
<td>CtgTATCC/CtATCC</td>
<td>79.7*</td>
<td>18.35</td>
<td>chr12:83980806–83990697</td>
<td>1</td>
</tr>
<tr>
<td>MER30</td>
<td>hAT</td>
<td>L1M1</td>
<td>Ttctaatgc/tttcaatg</td>
<td>61.6*</td>
<td>15.58</td>
<td>chr8:43680827–43681054</td>
<td>2</td>
</tr>
<tr>
<td>MER309</td>
<td>hAT</td>
<td>L1M3</td>
<td>TccagGAT/tccGgAT</td>
<td>68.1*</td>
<td>12.96</td>
<td>chr3:117825182–11782536</td>
<td>1</td>
</tr>
<tr>
<td>MER75</td>
<td>piggyBac</td>
<td>L1PA13</td>
<td>TAA/TTAA</td>
<td>65.8*</td>
<td>15.75</td>
<td>chr10:110019741–11002055</td>
<td>3</td>
</tr>
<tr>
<td>MER1B</td>
<td>hAT</td>
<td>L1M1</td>
<td>Gtttagat/Gtttagct</td>
<td>61.6*</td>
<td>15.58</td>
<td>chrX:33586269–33586602</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DNA element</th>
<th>Superfamily</th>
<th>Alu</th>
<th>Target site duplication</th>
<th>Age of Alu</th>
<th>Avg % div. of Alu</th>
<th>Position (May 2004 Assembly)</th>
<th>No. of nestings found</th>
</tr>
</thead>
<tbody>
<tr>
<td>MER85</td>
<td>piggyBac</td>
<td>Alu</td>
<td>b</td>
<td>TAA/TTAA</td>
<td>56*</td>
<td>16.57</td>
<td>chr1:15696507–15696630</td>
</tr>
<tr>
<td>MER107</td>
<td>hAT</td>
<td>Alu</td>
<td>d</td>
<td>None</td>
<td>60*</td>
<td>17.43</td>
<td>chr16:15634659–15634717</td>
</tr>
<tr>
<td>MER75b</td>
<td>piggyBac</td>
<td>Alu</td>
<td>b</td>
<td>TAA/TTAA</td>
<td>56*</td>
<td>16.57</td>
<td>chr2:202795190–202795404</td>
</tr>
<tr>
<td>MADE1</td>
<td>Tc1/meriner</td>
<td>Alu</td>
<td>x</td>
<td>Ta/TA</td>
<td>39.8*</td>
<td>12.17</td>
<td>chr2:18793090–18793161</td>
</tr>
<tr>
<td>MER30</td>
<td>hAT</td>
<td>Alu</td>
<td>o</td>
<td>Ggctagag/Ggctagag</td>
<td>60*</td>
<td>17.47</td>
<td>chr8:25981852–25981851</td>
</tr>
<tr>
<td>MER1A</td>
<td>hAT</td>
<td>Alu</td>
<td>b</td>
<td>Gcttagag/Gcttagag</td>
<td>56*</td>
<td>16.57</td>
<td>chrX:3362201–3362641</td>
</tr>
<tr>
<td>MER18</td>
<td>hAT</td>
<td>Alu</td>
<td>b</td>
<td>Gcttagag/GcTtagag</td>
<td>56*</td>
<td>16.57</td>
<td>chr19:33538374–33538704</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DNA element</th>
<th>Superfamily</th>
<th>LTR</th>
<th>Target site duplication</th>
<th>Avg % div. of LTR</th>
<th>Position (May 2004 Assembly)</th>
<th>No. of nestings found</th>
</tr>
</thead>
<tbody>
<tr>
<td>MER75</td>
<td>piggyBac</td>
<td>MSTB</td>
<td>TttAA/TTAA</td>
<td>16.86</td>
<td>chr5:520183–520422</td>
<td>1</td>
</tr>
<tr>
<td>MADE1</td>
<td>Tc1/meriner</td>
<td>THE1B</td>
<td>Ta/TA</td>
<td>14.75</td>
<td>chr2:27628868–27628948</td>
<td>1</td>
</tr>
<tr>
<td>MER30</td>
<td>hAT</td>
<td>MSTA</td>
<td>GgcTcaGGc/GgcTcaGGc</td>
<td>14.75</td>
<td>chr9:117723954–117724179</td>
<td>2</td>
</tr>
<tr>
<td>MER75</td>
<td>piggyBac</td>
<td>MSTA</td>
<td>TAA/TTAA</td>
<td>14.75</td>
<td>chr5:36127535–36128098</td>
<td>2</td>
</tr>
<tr>
<td>MER1A</td>
<td>hAT</td>
<td>MSTA</td>
<td>Gctgaacc/Gcttacc</td>
<td>14.75</td>
<td>chr5:35215634–35216182</td>
<td>6</td>
</tr>
<tr>
<td>MER18</td>
<td>hAT</td>
<td>MSTA</td>
<td>Gcttagat/GgTtagat</td>
<td>14.75</td>
<td>chr7:35906363–35906685</td>
<td>6</td>
</tr>
</tbody>
</table>

*Age from Khan et al. 2006.

**Age from Smit et al. 1995 and Khan et al. 2006.

1Age from Price et al. 2004.

2Age from Xing et al. 2004.
the galago (Table 2), except for MER107, which once again created a short deletion upon insertion rather than a typical TSD (Supplemental Fig. 1). The youngest Alu element that suffered a DNA element insertion was a member of the AluSx subfamily, which amplified ∼44 Mya (Xing et al. 2004). Four unambiguous instances were found of an AluSx element that had suffered an insertion of a MADE1 transposon (one example is shown in Table 2). Each MADE1 was integrated at a different position within each AluSx element, indicating that they all resulted from independent MADE1 transposition events rather than from propagation of a composite element.

Though the evolutionary history of primate LTR elements has not been analyzed as fully as those of L1 and Alu elements, we drew upon the available literature to provide further validation for the nested insertion analysis. Smit (1993) demonstrated that five families of LTR retrotransposons (THE1A, THE1B, THE1C, MSTIA, and MSTB) were primate specific. We found that six of the 11 DNA elements that had nested within the primate-specific L1 and Alu had nested within these LTR elements (Table 2). Each insertion was validated by the presence of the expected TSD. The youngest primate-specific LTR to suffer an insertion from a DNA element was THE1B.

Cross-species genomic analysis of orthologous insertions

To refine the age of the DNA transposon families, we turned to another method that did not rely on sequence divergence. It is possible to ascertain when individual TE insertions occurred by investigating their presence/absence at orthologous genomic regions in multiple species whose phylogenetic relationships are confidently established. If an element is present and fixed within the population in one species, but is absent at the orthologous position in another species, then the element must have been transposing some time after the split of the two species. Note that for DNA transposons, which transpose through a cut-and-paste mechanism, this pattern could be caused by either the insertion of the element or by its excision in one of the two species examined. In either case, nonetheless, the absence of the element in one of the two species can be interpreted as a manifestation of transposon activity after the divergence time of the two species. Conversely, if an element is present at orthologous positions in two different species, it is almost certain that this insertion predates the divergence of the two species. This is because the probability of two elements of the same family inserting at the exact same position independently in two different lineages is extremely small, especially for DNA elements, as these occur in relatively low copy-number families in mammalian genomes.

Taking advantage of the ongoing NISC-NIH ENCODE Project and of other genomic resources accessible through the UCSC Genome Browser, we assessed the presence/absence of 794 human DNA transposon copies at orthologous positions in at least eight (and up to 10) other mammalian species, including three primate species (see Methods). These elements represent 111 of 125 families known in the human genome.

We found members of 11 DNA element families that were present at orthologous positions in human, Rhesus macaque (Macaca mulatta), and marmoset (Callithrix jacchus), but absent in the galago (Otolemur garnettii), a prosimian primate. We were able to identify “empty sites” in the respective orthologous galago loci for members of each DNA element family (one example per family is shown in Table 3) except for MER75B, since there is only one copy of MER75B within an ENCODE region and there is a large deletion within the galago lineage at that locus. These were clear “empty sites,” with only one copy of the TSD and no additional sequences indicative of transposon excision. DNA transposon excisions are typically imprecise in that they leave behind one of the two TSD and/or a few terminal nucleotides of the transposon at the excision site (for example, see Plasterk 1991; for review, see Craig et al. 2002). This data suggests that these elements were inserted in the anthropoid lineage, rather than excised in the galago lineage. We assume that these individual copies are representative of their respective families, because their individual percent divergences are all within the 95% confidence interval for average divergence of the family (average divergence ± 1.96 SD). As such, these elements do not represent statistical outliers. These data demonstrate that at least 11 families of human DNA transposons were transpositionally active after the split of prosimians and anthropoid primates, or during the last ∼63 My (Goodman 1999). Seven of these 11 families also include copies nested within L1 or Alu elements known to be primate specific (see above; Table 2). Assuming that these individual copies are representative of their respective families and that their activity is contemporary to the activity of their entire family, these 11 families make up a total of 23,570 transposons in our genome. Therefore, these data imply that many thousands of DNA transposons were inserted in the lineage of anthropoid primates, that is, within the last ∼63 My.

Members of the remaining 100 human DNA transposon families represented in the ENCODE regions were found at orthologous positions in human, marmoset, and galago. To investigate which of these families was primate specific, each of the 316 copies, along with at least 100 bp flanking both the 5’ and 3’ ends, was visually inspected in the UCSC Genome Browser for their presence/absence at orthologous positions in at least five nonprimate mammals using the ENCODE Comparative Genomics tracks. Phylogenetically, these five species form two separate outgroups to the primates, with the mouse/rat/rabbit lineage being closer to the primates than the cow/dog lineage (Murphy et al. 2004). Hence, a transposon present in primates but absent in the five other mammals most likely resulted from an insertion in the primate lineage. A less parsimonious explanation for this pattern would be that the element inserted in a eutherian (i.e., placental) ancestor and subsequently excised twice, independently, in the mouse/rat/rabbit lineage and in the dog/cow lineage. We found 163 elements from 23 families that were clearly present in humans as well as three of the five other primate genomes (chimp, baboon, rhesus, marmoset, and galago), but absent in all other mammals examined. Thus, these 23 families were classified as primate specific.

The remaining 77 families were represented by elements present at orthologous sites in all the primates and at least one of the other eutherian mammals; thus, these families were classified as eutherian-wide. It should be noted, however, that several families (such as Charlie1 and Charlie1a) included some copies that were apparently primate specific, as well as copies present at orthologous positions in primates and at least one of the nonprimate mammals. It could well be that the activity of these families initiated prior to the emergence of the primates, but continued in a primate ancestor, generating primate-specific insertions. Alternatively, lineage-specific sorting of ancient alleles with or without the insertions could also account for these patterns. In the absence of further evidence to distinguish between these possibilities, we adopted a conservative classification of such families as eutherian-wide.
To determine whether the other 14 DNA element families not represented within the ENCODE regions were primate specific, we searched the finished mouse, rat, and dog genomes for the presence or absence of TEs at orthologous positions to the human DNA elements (see Methods). We found that six families of elements (MER6C, Ricksha, Ricksha_b, Ricksha_c, Tigger5a, and Tigger5b) were present in primates, but clearly absent in the mouse, rat, and dog. These families were additionally classified as primate specific. The other eight families were found to be present in the primates and in the mouse, rat, or dog genomes. These families were classified as eutherian-wide. Together, the cross-species analysis of orthologous insertions suggests that a minimum of 40 distinct DNA transposon families, which accounts for 98,300 DNA elements currently fixed in the human genome, were active in the primate lineage, i.e., within the last ~80 My.

**Age of DNA transposons based on sequence divergence**

The results above allowed us to distinguish among DNA transposon families that have been active in an anthropoid (40–63 Ma), primate (63–80 Mya), or eutherian ancestor (80–150 Mya) (Goodman 1999; Murphy et al. 2004). In order to obtain a better resolution of the age of each family, we used the median of these three evolutionary periods (51.5, 72, and 115.5 My, respectively) to determine the median substitution rate per million years of all DNA transposons within each age class (anthropoid, primate, or eutherian). We also calculated a low- and high-substitution rate using the upper and lower bounds of the age class. This was done by calculating the corrected number of nucleotide substitutions per site in all individual copies to their respective family consensus sequence (as given by Repeatmasker), and excluding consensuses Cpg positions, using the REV model (Tavare 1986) (see Methods). The REV, or general reversible model, calculates five different rate parameters based upon nucleotide composition, thereby correcting for differing nucleotide compositions between DNA element families. The same analysis was performed separately with the L1s and Alus falling into the anthropoid-or primate-specific class ages (plus an additional class of <40 Ma) and four eutherian-wide L1 elements (L1PA17, PB4, and MA4-S), following the most recently published dating of L1 (Khan et al. 2006) and Alu subfamilies (Price et al. 2004; Xing et al. 2004).

This approach allowed us to calculate a substitution rate per million years that takes into account possible fluctuations in evolution rates during the different time periods and among the three types of elements (see Methods; Supplemental Table 2). The results reveal that DNA transposons, L1s, and Alus each have different average substitution rates in the different evolutionary periods (Table 4). There are statistically significant differences among the three types of TEs within the same period. For example, within the anthropoid-specific lineage (41–63 My), which was the only period for which mutation rates could be estimated for all three types of TEs, the differences were statistically significant (P < 0.05, ANOVA). These differences may be attributed to several factors, such as biased genomic distributions (e.g., Alus preferentially accumulate in GC-rich regions), compositional biases, replication mechanisms (reverse transcriptase for RNA elements, DNA polymerase for DNA transposons), and amplification dynamics (subfamily structure). Regardless, this data provides a rationale for treating DNA transposons separately from the other types of TEs in the human genome in these calculations as opposed to using substitution rates estimated for other type of elements or for other neutrally evolving sequences such as pseudogenes (for example, see Robertson and Martos 1997).

We next applied the substitution rate of each type of element for each period to estimate the age of individual families based on their average nucleotide divergence to the respective consensus sequence, excluding Cpg sites (see Methods). This method differed slightly from other datings of TEs in which a single, constant substitution rate for the entire span of primate evolution was used (for examples, see Price et al. 2004; Khan et al. 2006). Our estimates for the age of L1 and Alu subfamilies using
these period-specific substitution rates differed somewhat from recently published datings (Fig. 1; Price et al. 2004; Xing et al. 2004; Khan et al. 2006). This is to be expected since all three published datings used the Kimura 2-parameter correction not the REV model (Kimura 1980). Overall, L1 elements appeared to be slightly older, due to their higher consensus AT content, and Alu elements appeared to be slightly younger, due to their lower consensus AT content, using the REV model rather than using the Kimura model. The resulting estimated ages of human DNA transposon families less than or equal to 80 My is given in Table 5, and the dating for all 125 families is available in the Supplemental Table 1. Table 5 gives the median age for the family and the upper and lower boundaries of the age.

This data was used to generate the plot of the age of DNA transposons and L1s as a function of their copy number (Fig. 1). This representation shows that our dating of DNA transposons is in good agreement with the nested insertion analysis, but provides a better resolution of the age of an individual family. For example, MER85 and MADE1 are among the youngest DNA elements, with estimated ages of ∼37 and ∼46 Mya, respectively, and both families included members nested within L1PA10 elements (Table 2), a subfamily that we dated as ∼51-My old (in agreement with Khan et al. 2006). Conversely, there are no DNA transposon families significantly younger than L1PA8, as indeed we found no instances of any DNA element nested within L1PA8 or younger. However, we could detect several L1PA8 elements nested within MER85 and MADE1 transposons (data not shown).

Our dating of individual DNA transposon families based on sequence divergence and calculated age is also largely congruent with the cross-species analysis. All elements classified as eutherian-wide from this analysis were found to be older than 65 My based on sequence divergence, with all but one family (MER53) being older than 70 My. MER53 is an outlier due to its unusually high 70% AT content (Table 5). All DNA transposon families classified as primate specific by the cross-species analysis were estimated to be between 57 and 78 My based on sequence divergence.

Figure 1 reveals that there were two bursts of DNA transposon activity in the time period between the mammalian radiation and the split of New World Monkeys from the primate ancestor. The first peak is the most pronounced and involves primarily members of the hAT superfamily and spanned a period of ∼40 My from a pre-primate era to early primate evolution (∼70 Mya). The second subsequent peak is strictly primate specific (from 80 to 63 Mya) and mostly implicated Tc1/mariner elements, although the greatest diversity of DNA transposons was active during this time, including hAT, MuDR, and PiggyBac elements (Fig. 1). The relatively more recent activity of Tc1/mariner elements compared with the most abundant hAT elements has been previously noticed (Lander et al. 2001). Our data indicates that the burst of Tc1/mariner activity reached a plateau that seems to coincide with the emergence of anthropoid primates, at ∼63 Mya. Three superfamilies (hAT, Tc1/mariner, PiggyBac) continued to be active in the anthropoid lineage, but there seems to be a sudden loss of activity of all DNA transposons shortly after the emergence of the new world monkeys – 40 Mya.

Discussion

While the evolutionary history of human Alu and L1 retroposons has been studied intensively, the history of DNA transposons has largely been overlooked. In this study we have combined three different approaches to determine the average age of all 125 DNA transposon families known in the human genome. The results of the three approaches converge to reveal that a substantial fraction of human DNA transposon families (at least 40 and up to 69 families, see Tables 5, 6), representing at least ∼98,000 elements in our genome, were transpositionally active in the primate lineage. Below we first discuss the value of combining various methods for estimating the age of TEs, then turn to the specific implications of our findings for primate genome evolution and for understanding the forces underlying the amplification dynamics of TEs in mammalian genomes.

Combined methods provide a detailed estimate of the age of human DNA transposons

As new genome sequences are released, different methods for dating transposable elements are being developed that allow greater accuracy in estimating the age of TEs (Price et al. 2004; Salem et al. 2005; Caspi and Pachter 2006). This is a crucial aspect of genome research because TEs not only provide a rich fossil record to determine the pace and mode of molecular evolution (for example, see Waterston et al. 2002), but they are also major players in the structural evolution and regulation of genes and genomes (Feschotte et al. 2002; Deininger et al. 2003; Kazazian 2004; Cordaux et al. 2006). To our knowledge, our study is the
first that uses a combination of three independent methods to evaluate the age of a broad population of TEs on a genome-wide scale.

The first method, nested insertion analysis, capitalizes on the well-characterized history of L1 and Alu elements and provides an estimation of the relative age of TE families. This method does not rely on the molecular clock and can be performed even in the absence of genomic sequences for other closely related species. The shortcoming of this approach is that not every TE family member will necessarily suffer an insertion from every other TE family, thereby leading to gaps in the data, especially for TE families with low copy numbers. The second method (cross-species analysis of orthologous insertions) takes advantage of the large amount of sequence data recently generated for several mammalian species and is also independent of the molecular clock. These first two methods deliver a rough, yet robust evaluation of the time periods when the elements were inserted even in the absence of genomic sequences for other closely related species.

Sixty-nine families of DNA transposons were predicted to be primate specific based on sequence divergence data and confirmed to be primate specific by at least one of the two alternative methods. Sixty-nine families of DNA transposons were predicted to be primate specific by all three methods (Fig. 2). Since nested insertion analysis does not allow us to examine all DNA transposon families, only a subset of them, one cannot expect a complete overlap between the results produced by the three different methods.

Of the 125 DNA element families currently recognizable in the human genome, a total of 11 families could be classified as primate specific by all three methods (Fig. 2). Since nested insertion analysis does not allow us to examine all DNA transposon families, only a subset of them, one cannot expect a complete overlap between the results produced by the three different methods. Sixty-nine families of DNA transposons were predicted to be primate specific based on sequence divergence data and thus, calculated age, alone (Table 5), and 40 of these families were confirmed to be primate specific by at least one of the two alternative methods. Hence, we believe that this set of 40 families provides a reliable, yet conservative estimate of DNA transposons that were integrated during primate evolution. The corresponding families range in age from MER85 (37 My) to Tigger5c (78 My) and have contributed 98,300 elements (totaling ∼33 Mb of their ancestral consensus sequence. There is only partial overlap between the results gathered by the three methods (Fig. 2), which emphasizes the value of combining the three methods to derive a reliable history of the entire population of human DNA transposons.

### Table 5. DNA elements less than 80 million years

<table>
<thead>
<tr>
<th>Category</th>
<th>RepName</th>
<th>Corrected Age (My)</th>
<th>% Div. (excluding CpGs)</th>
<th>AT content of consensus sequence (excluding CpGs)</th>
<th>AT content (excluding CpGs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>26</td>
<td>4</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

...
sequence) to the human genome (Fig. 2). Furthermore, the results of divergence and cross-species and/or nested insertion analysis confirm that nearly one-fourth of these transposons (23,462 elements, ~5 Mb) have likely been inserted since the split of anthropoid primates from prosimian primates, or within the last ~63 My (according to Goodman 1999).

Thirty families were predicted to be primate specific based on their sequence divergence and calculated age but were shown to be eutherian-wide by cross-species analysis. However, for some of these families, such as MERS3, Charlie5, and MER33 (65, 72, 72 My, respectively), we could detect copies present at orthologous positions in all eutherian species examined, which strongly indicates that at least a subset of family members inserted prior to the divergence of placental mammals. Thus, sequence divergence alone may not always be an accurate measure of the age of TEs. It could be that, for various reasons, members of these families evolve more slowly overall than other families. A non-mutually exclusive explanation is that these families include a subfamily of primate-specific elements as well as older elements. Further analyses are required to distinguish between these possibilities.

A general extinction of DNA transposon activity in the anthropoid lineage

An interesting phenomenon is observed when looking at the overall history of DNA transposons in the mammalian and primate lineages (Fig. 3). Eighty-five families, or ~291,000 DNA transposons, are shared between primates and other mammals. In contrast, 29 families, or ~74,000 elements, were active specifically in primates prior to the split of emergence of anthropoids, and 11 families, or ~23,000 elements, were integrated in anthropoid species. Thus, there was a steady decline in the activity of DNA transposons during primate evolution (see Figs. 1, 3). According to our combined age calculations, we found no evidence for DNA transposon families significantly younger than the divergence of new world monkeys, that is ~40 Mya (Fig. 1; Table 5). Furthermore, we conducted a systematic survey for the presence/absence of human DNA transposons at orthologous positions in the nearly complete genome of the Rhesus macaque (an old world monkey) and could not uncover a single conclusive instance of a DNA transposon copy present in human, but missing in the macaque (data not shown). Thus, there is no evidence for the activity of any DNA transposons after the emergence of old world monkeys. The last active DNA transposon families represented in the human genome seem to have all become extinct in the relatively short evolutionary window (~23 My) that separated prosimians and new world monkeys (Fig. 1). Yet, our study predicts that at the dawn of this extinction, some ~40–55 Mya, there were at least 11 families from three different superfamilies active in the anthropoid genome (Fig. 1; Table 5). The majority of these elements were from the hAT superfamily (six families), while two Tcl/meriner and three piggyBac families comprised the rest of this group (Figs. 1, 3). This suggests that at least three distinct sources of transposases, with some of them represented by hundreds of seemingly intact copies (e.g., Hsmar1) (Robertson and Zampano 1997; Cordaux et al. 2006) were shut down around the same evolutionary period.

What could have provoked the extinction of DNA transposons that would not have affected the propagation of L1, which continued to thrive even after the emergence of new world monkeys (see Fig. 1; Khan et al. 2006)? A distinctive feature of the life cycle of DNA transposons is their apparent propensity for horizontal transmission (Silva and Kidwell 2000; Robertson 2002). Theory predicts that horizontal transfer is indeed critical for the maintenance of DNA transposons (Hickey 1982; Hartl et al. 1997). In contrast, horizontal transfer of LINEs occurs rarely, if ever (Eickbush and Malik 2002) and it is probably not an essential component to their maintenance (Wei et al. 2001; Kulpa and Moran 2006). It is tempting to speculate that the extinction of the DNA transposon population in the anthropoid lineage was linked to a sudden incapacity of these elements to undergo horizontal transmission. This could be due to the emergence of a host barrier aimed against the cellular entrance of TEs and other forms of invasive DNA. This would also explain the parallel regression of endogenous retroviruses during the same period of primate evolution (Lander et al. 2001). Interestingly, several defense mechanisms have been recently characterized that restrict retroviral activity in primates (Emerman 2006; Zenou and Bieniasz 2006). It could be that similar mechanisms have also compromised the propagation of DNA transposons in anthropoid primates.

Contribution of DNA transposons to primate genome evolution

One of the most striking findings of our study is that at least ~74,000 of the DNA transposons now fixed in the human genome (~33 Mb of DNA) were integrated during a period of ~17 My, prior to the emergence of prosimian primates (~63 Mya) but...
after the divergence of a primate ancestor from the closest mammalian clades represented in our data set (rat, mouse, and rabbit; ∼75–85 Mya, see Fig. 1). This is almost twice the number of L1 elements inserted during the same period and now fixed in the human genome (∼43,000 elements from the subfamilies L1PA15–16, L1PB3, L1MA2–3). Clearly, early primate evolution was a period of intense activity for DNA transposons. During the next phase of the primate radiation (63–40 Mya), i.e., after the split of prosimians, but prior to the emergence of new world monkeys, we estimated that ∼23,000 DNA elements were inserted and fixed in the human genome, adding at least ∼5 Mb to an ancestral anthropoid genome (Figs. 1, 3). Hence, the quantitative contribution of DNA-mediated transposition to the primate genome is far from negligible.

Prior to this study, the history of human DNA transposons has been largely neglected relative to those of retroelements (Alus and L1s). One reason for this is the common belief that active DNA transposon families have long been extinct and that they are currently only represented by very ancient molecular “fossils” immobilized in the genome. Indeed, unlike Alus and L1s (Deininger et al. 2003), there are no known cases of de novo insertion of any human DNA transposon. Our results support the conclusion that there has been little, if any, activity of DNA transposons in the ape lineage. On the other hand, our study demonstrates that many thousands of DNA elements have integrated and become fixed during the first half of primate evolution and that several high copy number families with >90% nucleotide identity among copies remain in the human genome (see average sequence divergence in Table 5). It is tempting to speculate that these primate-specific bursts of DNA transposition have had a strong impact on the structural evolution of primate genomes. DNA transposons have been frequently implicated in chromosomal rearrangements in plant and animal species, including deletions, inversions, duplications, translocations, and chromosome breakage mediated by interelement recombination or aberrant transposition events (for examples, see Lim and Simmons 1994; Caceres et al. 1999; Gray 2000; Zhang and Peterson 2004). Given the medical and evolutionary importance of chromosomal rearrangements in humans (Inoue and Lupski 2002; Eichler and Sankoff 2003; Feuk et al. 2006), the possible role of DNA transposons in shaping primate genomes warrants further investigation.

Methods
Calculation of average percent divergence from RepeatMasker output
The average percent divergence of each transposable element family was calculated using the RepeatMasker rmsk files from the UCSC Genome Browser for the May 2004 assembly. The percent divergence (milliDiv) of each distinct element within a transposable element family was weighted by the length of the element. The average percent divergence was weighted to account for the vast difference in sizes between currently recognized elements of the same family. To calculate the average percent divergence for each family, the percent divergence calculated by RepeatMasker was multiplied by the length of the element and the sum of all elements in each family was divided by the sum of the total length of all elements in the family.

Nested insertions
To find nested insertions, the RepeatMasker files from the UCSC Genome Browser and a Perl script (see Supplemental Material) were used to automate the analysis. Each Alu, L1, and DNA element in the RepeatMasker files was evaluated to see whether there was a repeat within 50 bp upstream and a repeat with 50 bp downstream. If so, the upstream and downstream repeats were compared to see whether the repName and Strand fields were the same. Next, the repeat positions of the ends of the upstream and downstream TEs were examined. If the upstream and downstream TEs both began within position 1–20 or both ended within 20 bp of the consensus length, the case was discarded, suggesting that they might represent a cluster of two elements of the same family independently inserted in close proximity and surrounding another element, rather than a single element disrupted by a nested element. Next, the coordinates of the two flanking repeats were checked to verify that the end of the first repeat was within ±20 bp of the start of the second repeat according to the family consensus sequence. If all of these conditions were met, the Alu, L1, or DNA element was classified as nested within each other.

Cross-species genomic analysis of orthologous insertions
To search for the presence or absence of DNA elements in marmoset and galago, sequences for each human DNA element present in an ENCODE region (http://www.nsc.nih.gov) were retrieved from the UCSC Genome Browser in September 2005. BAC sequences for each of the orthologous ENCODE regions in the marmoset and galago were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) using the accession ID given on the ENCODE website. From these sequences, custom BLAST databases were built for both the marmoset and galago using BLAST Version 2.2.11 (Altschul et al. 1990) and the sequences
from the GenBank accessions. Each human DNA transponson, along with 100 bp of flanking sequences, were used as queries in BLASTN searches of the marmoset and galago custom databases. A repeat was classified as present at the orthologous position if at least one of the two flanking regions and at least 50% of the element from the human sequence were found in the marmoset or galago.

Calculation of substitution rates and dating according to sequence divergence

The May 2004 human genome sequence (hg17) was downloaded from the UCSC Genome Browser. TEs were masked locally using RepeatMasker version 3.0 with the March 14, 2006 library from RepBase Update. A Perl script (see Supplemental Material) was used to parse the RepeatMasker alignments and generate a single, concatenated sequence for each different chromosomeal repeat along with the corresponding consensus sequence. The concatenated sequences had all CG dinucleotides (for + strand) and GC dinucleotides (for − strand), as well as non-ATGC characters, removed. These chromosomal repeat and consensus sequences were then combined and analyzed using PAML version 3.15 (Yang 1997). Each file was analyzed using the REV model with the clock = 1 option. The corrected number of substitutions per site was calculated as one-half of the branch length, since the consensus sequence does not evolve.

To calculate the substitution rate, the corrected substitutions per site was divided by the median age for the class (anthropoid, primate, or eutherian specific). The rate for each TE within the age class was weighted by the percentage of the total number of bases that TE comprised of the total base length for the entire class. These weighted rates were then summed, giving a corrected substitution rate for the entire class. The age of the family was calculated by multiplying the corrected substitution rate by the corrected percent divergence for the family.

Acknowledgments

We thank Ellen Pritham for critical reading of the manuscript, Don Hucks for support with PAML, and members of the Genome Biology Group at UTA for stimulating discussions. This work was supported by funds from the University of Texas at Arlington.

References


Evolutionary history of human DNA transposons

205: 219–228.


Received August 1, 2006; accepted in revised form January 9, 2007.