Available online at www.sciencedirect.com
ScienceDirect
Gene 396 (2007) 46–58
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

SINEs, evolution and genome structure in the opossum

Wanjun Gu [a,1], David A. Ray [b,c,1], Jerilyn A. Walker [b], Erin W. Barnes [b], Andrew J. Gentles [d,e], Paul B. Samollow [f], Jerzy Jurka [e], Mark A. Batzer [b,2], David D. Pollock [a,*,2]

[a] *Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA*
[b] *Department of Biological Sciences, Biological Computation and Visualization Center, Center for Bio-modular Multi-scale Systems, Louisiana State University, Baton Rouge, LA 70803, USA*
[c] *Department of Biology, West Virginia University, Morgantown, WV 26505, USA*
[d] *School of Medicine, Stanford University, Stanford, CA 94305, USA*
[e] *Genetic Information Research Institute, 1925 Landings Drive, Mountain View, CA 94043, USA*
[f] *Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843-4458, USA*

Received 23 December 2006; received in revised form 15 February 2007; accepted 19 February 2007
Available online 19 March 2007

Abstract

Short INterspersed Elements (SINEs) are non-autonomous retrotransposons, usually between 100 and 500 base pairs (bp) in length, which are ubiquitous components of eukaryotic genomes. Their activity, distribution, and evolution can be highly informative on genomic structure and evolutionary processes. To determine recent activity, we amplified more than one hundred *SINE1* loci in a panel of 43 *M. domestica* individuals derived from five diverse geographic locations. The *SINE1* family has expanded recently enough that many loci were polymorphic, and the *SINE1* insertion-based genetic distances among populations reflected geographic distance. Genome-wide comparisons of *SINE1* densities and GC content revealed that high *SINE1* density is associated with high GC content in a few long and many short spans. Young *SINE1s*, whether fixed or polymorphic, showed an unbiased GC content preference for insertion, indicating that the GC preference accumulates over long time periods, possibly in periodic bursts. *SINE1* evolution is thus broadly similar to human *Alu* evolution, although it has an independent origin. High GC content adjacent to *SINE1s* is strongly correlated with bias towards higher AT to GC substitutions and lower GC to AT substitutions. This is consistent with biased gene conversion, and also indicates that like chickens, but unlike eutherian mammals, GC content heterogeneity (isochore structure) is reinforced by substitution processes in the *M. domestica* genome. Nevertheless, both high and low GC content regions are apparently headed towards lower GC content equilibria, possibly due to a relative shift to lower recombination rates in the recent *Monodelphis* ancestral lineage. Like eutherians, metatherian (marsupial) mammals have evolved high CpG substitution rates, but this is apparently a convergence in process rather than a shared ancestral state.

Keywords: Repetitive elements; SINE; Population genetics; Molecular evolution; Substitution rate; Isochores; *Monodelphis domestica*

1. Introduction

Transposable elements are involved in reshaping genomes through processes such as insertion mutagenesis (Deininger and Batzer, 1999; Kidwell and Lisch, 2001; Batzer and Deininger, 2002; Ostertag et al., 2003), insertion-mediated genomic deletions (Kidwell and Lisch, 2001; Gilbert et al., 2002; Symer et al., 2002; Callinan et al., 2005; Han et al., 2005), retrotransposition-mediated sequence transduction (Moran et al., 1999; Goodier et al., 2000; Pickeral et al., 2000; Miller and Capy, 2004; Xing et al., 2006), and post insertion recombination (Sen et al., 2006). The presence of several thousand to over one million nearly

Abbreviations: A, Adenine; bp, base pairs; C, Cytidine; CpG, Cytidine-phosphate-Guanosine Dinucleotide; G, Guanosine; kb, kilo base pairs; Mb, mega base pairs; MYA, Million Years Ago; myr, million years; SINEs, Short INterspersed Elements; spm, SINEs per Million base pairs; T, Thymidine; ts, transition rate; tv, transversion rate.

* Corresponding author. Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, MS 8101 (12801 17th Ave.), PO Box 6511, Aurora, CO 80045, USA. Tel.: +1 303 724 3234.
E-mail addresses: Wanjun.Gu@uchsc.edu (W. Gu), David.Ray@mail.wvu.edu (D.A. Ray), JAWalker@lsu.edu (J.A. Walker), AndrewG@stanford.edu (A.J. Gentles), PSamollow@cvm.tamu.edu (P.B. Samollow), Jurka@charon.girinst.org (J. Jurka), MBatzer@lsu.edu (M.A. Batzer), David.Pollock@uchsc.edu (D.D. Pollock).

[1] These two authors contributed equally and should be considered joint first authors.

[2] These two authors contributed equally as senior authors.

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.02.028

identical fragments of DNA scattered throughout a genome promotes recombination, including both equal and unequal crossover events (Deininger and Batzer, 1999; Gebow et al., 2000; Kidwell and Lisch, 2001). Segmental duplication (Bailey et al., 2003), gene conversion (Kass et al., 1995; Roy-Engel et al., 2002; Fischer et al., 2003), exon shuffling (Moran et al., 1999; Witte et al., 2001; Ejima and Yang, 2003; Jiang et al., 2004) and chromosomal rearrangements (Gray, 2000) have also been associated with highly repetitive mobile elements. All of these processes increase genetic variation in populations, and thus boost diversity within and among species. Studying the dynamics of mobile element amplification and evolution thus provides us with crucial insight into genomic diversification.

At least half of the newly sequenced *Monodelphis domestica* (gray short-tailed opossum) genome is comprised of highly repeated mobile elements (Gentles et al., in press). One of these elements, *SINE1*, is a short, 191 bp repeat with a poly-A tail, analogous to the *Alu* retrotransposons of primates. There are 603,385 copies of this repeat family in the opossum genome, making it about half as common as *Alu* elements in the human genome (Lander et al., 2001). Together with other short interspersed elements (SINEs), these elements account for more than 11% of the opossum genome. There has been considerable controversy over the role of *Alu* elements in human genome evolution, including questions concerning possible functional roles, and whether their affinity for GC-rich regions is the result of insertion bias, differential positive (e.g., adaptive) or negative (purifying) selection based on local GC content, or differential recombination and elimination following insertion (Grover et al., 2003; Grover et al., 2004; Jurka, 2004; Hackenberg et al., 2005; Cordaux et al., 2006). Given the similarity of opossum SINE structural features to those of *Alu* elements, it is of interest to determine whether some of their genomic distribution features are also similar, and whether the answers to the above questions about *Alu*s might also hold true for opossum SINE families. Also, the relatively low recombination rate (~0.2–0.3 cM/Mb; centiMorgans per megabase) in opossum (Samollow et al., 2004) may have historically created different dynamics in the *M. domestica* genome compared to the human genome.

The observation that older *Alu* elements are more common in GC-rich regions, whereas younger *Alu* elements tend towards slightly AT-rich regions, was originally interpreted to indicate positive selection to maintain *Alu* elements in GC-rich regions (Lander et al., 2001). There has been some discussion whether this model is theoretically feasible (Brookfield, 2001; Batzer and Deininger, 2002; The Chimpanzee Sequencing and Analysis Consortium, 2005), but a recent analysis of the location of fixed and polymorphic *Alu* elements in the very youngest *Alu* subfamilies concluded that recently integrated *Alu* elements are distributed randomly with regard to GC content, and that there is no detectable difference in surrounding GC content between polymorphic and recently fixed loci (Belle and Eyre-Walker, 2002; Cordaux et al., 2006). This effectively rules out GC content as a selective constraint on *Alu* element insertion and fixation, meaning that young *Alu* elements are essentially "neutral residents" in the genome with regard to GC content adjacent to the insertion site. A plausible alternative mechanism is that *Alu*s in AT-rich regions are preferentially removed by recombination, since gene densities are lower in AT-rich areas of the human genome, and there may be selection against deletion-causing *Alu–Alu* recombination in gene-rich regions (Pavlicek et al., 2001; Hackenberg et al., 2005). Assuming this is also true in the opossum, negative selection against deletion of coding sequences would act to "protect" *SINE1* elements that are integrated nearby.

Here, we begin to evaluate these questions in the opossum by analyzing the evolution and genomic distribution of *SINE1* elements in the *M. domestica* genome. We reconstruct the phylogenetic history of *SINE1* families, predict the youngest SINE subfamily, and report polymorphism levels in a panel of animals derived from five geographically distinct *M. domestica* populations. We relate *SINE1* density to regional GC content and to the GC content of adjacent sequence, and determine that substitution rates in *SINE1* elements sufficiently explain differences in GC content among region. There are remarkable similarities in genomic distribution patterns between the opossum *SINE1* and human *Alu* retrotransposon, and *SINE1* elements have considerable utility for analyzing mutation/substitution processes that shape genomic GC content and isochore structure.

2. Materials and methods

2.1. Detecting recent SINE1 families

At the time this project was initiated, the *M. domestica* genome had only recently been sequenced. The sequence was minimally annotated, but no repeat library was available. We therefore used a novel method to identify high-copy repeat elements in assembly release MonDom1 (UCSC genome browser, 2004 Oct Release; Hinrichs et al., 2006)) based on oligonucleotide word counts (Gu et al., in preparation). In brief, we collected all 1,554,427 oligonucleotides of length 13 bp and having at least 200 copies in the opossum genome (out of 67,108,864 total reverse-complement unique 13mers), and built a "core" 16mer sequence (CCTGGGCAAGTCACTT) from the four highest copy oligonucleotides in this "high-copy" set. By extending this set to include all high-copy oligonucleotides that differed from the core 16mer at one or two sites we created a core probability cloud (P Cloud) of 632,602 sequences from the *M. domestica* genome. These sequences were extended in both directions wherever they matched oligonucleotides in the high-copy set (terminating after two consecutive low-copy oligonucleotides), and the 5449 hypothetical repeat sequences (0.8% of the total) that were longer than 170 bp were retained. These sequences were then aligned using ClustalX (Thompson et al., 1997), to obtain a consistently well-aligned region of 191 bp. This 191 bp alignment was used in a repeat subfamily classification program (Price et al., 2004), which classified them into eight subfamilies with corresponding consensus sequences (Fig. 1). The youngest of these families contained 454 members. We define it as young because the average genetic distance of its members from the consensus is only 0.068% (Kimura 2-parameter; (Kimura, 1980)), the low divergence suggesting recent retrotransposition activity in the genome. For a

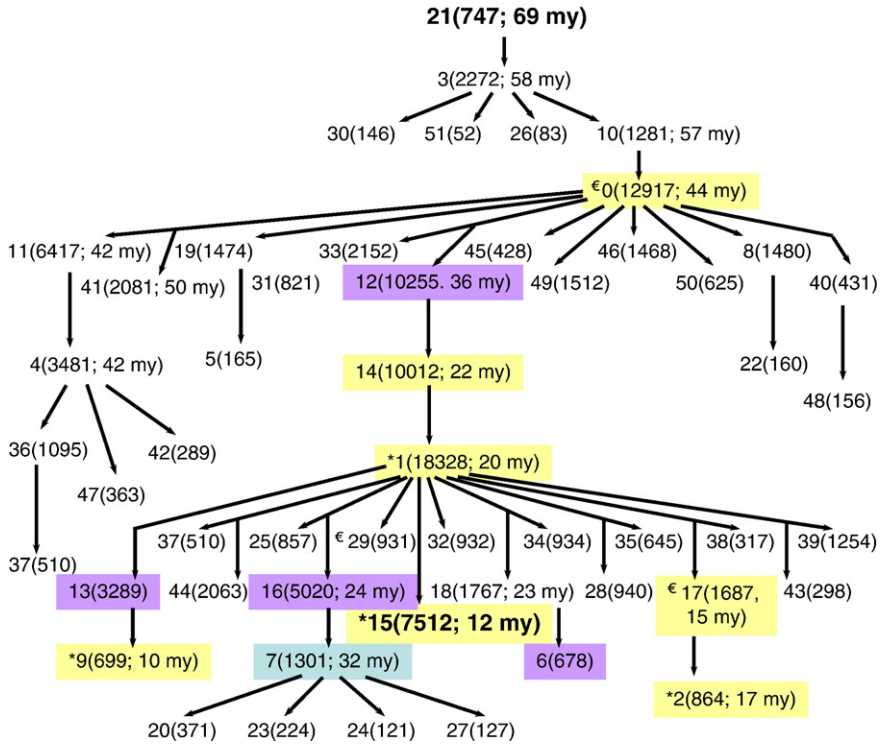

Fig. 1. Phylogenetic tree of subfamilies. The tree of subfamilies was obtained along with the subfamily affiliations and consensus sequences (Price et al., 2004) for a core 160 bp region of well-aligned sequences. The subfamily identification number is shown along with the number of loci in that subfamily in parentheses (and also the predicted mean age, in millions of years, for selected subfamilies). The arrows show the directions of ancestor to descendant relationships predicted by Price's program, but we note that these predictions are not always easily reconcilable with the mean age of the respective families. The predicted ancestral subfamily, and also the oldest, subfamily #21, is shown in bold, as is subfamily #15, which had the largest number of loci tested for polymorphism and was the largest very young family (<14 MYA). Subfamilies represented in the polymorphism test set are highlighted in yellow, while subfamilies with members originally identified as possible recent integrations (but which were not in the test set) are highlighted in light purple. Subfamilies with verified low or intermediate polymorphic loci are noted with an *, and subfamilies with a single verified high frequency locus are noted with a € symbol. Subfamily #29 had a single member in the polymorphism test set, which contained a high frequency polymorphism. Also note that subfamilies 0, 1, 2, 14, and 17 had tested loci that failed to provide assessable results for unknown reasons. Subfamily #7, discussed in the text for its aberrant adjacent sequence GC content and mean age in the center of a gap in subfamily production. Note that the ancestor/descendant relationships among this subfamily and subfamilies #16 and #1 appear incorrect based on the average ages of these subfamilies.

subset of 125 loci, PCR primers for amplification and polymorphism detection were designed using the proximal 500 bp on each side of these sequences in the opossum genome.

We subsequently tested for insertion polymorphism among *M. domestica* stocks originating from different locations across the species range. Specifically, DNA was extracted using standard procedures from a panel of 43 *M. domestica* liver samples derived from individuals bred at the Southwest Foundation for Biomedical Research (SFBR), San Antonio, Texas (USA). These individuals were chosen to represent five geographically distinct *M. domestica* populations (Fig. 2). Based on their distant geographic origins and preliminary genetic examinations (data not shown), population 1 and population 5 animals (pure stocks from regions 1 and 5, respectively) were believed to be the most genetically differentiated stocks in the SFBR colony. Detailed information concerning the origins and compositions of SFBR *M. domestica* stocks has been previously published (VandeBerg and Robinson, 1997; VandeBerg, 1999). Nine of the 43 samples were from a pure stock from region 1, four were from a pure stock from region 2, and 16 were from a pure stock from region 5. In addition, seven were from an admixed stock containing genetic material from regions 1 and 3, and seven were from an admixed stock with genetic material from regions 1 and 4.

PCR amplification was carried out in 25 μl reactions under the following conditions: 10–50 ng template DNA, 7 pM of each oligonucleotide primer, 200 mM dNTPs, in 50 mM KCl, 10 mM Tris–HCl (pH 8.4), 2.0 mM $MgCl_2$ and Taq DNA polymerase (1.25 units). An initial denaturation at 94 °C for 2 min was followed by 30–32 cycles of 94 °C for 15 s, the appropriate annealing temperature for 15 s, and 72 °C for 1 min and 10 s A final incubation at 72 °C for 5 min was included. PCR products were separated on 2% agarose gels, stained with ethidium bromide, and visualized using UV fluorescence. Interpretable genotypes (defined as amplicons with clearly distinguishable filled and/or empty site bands) were scored based on the presence or absence of a PCR product with a length corresponding to that predicted if the element were present or absent (Supplementary Table A). Polymorphic loci were classified based on the allele frequencies of the filled site in the total sample: fixed present (FP) loci were homozygous for the filled site in all tested individuals; high frequency (HF) loci had frequencies of 0.667 to less than 1.0; intermediate frequency (IF) loci had frequencies between 0.334 and 0.666; and low

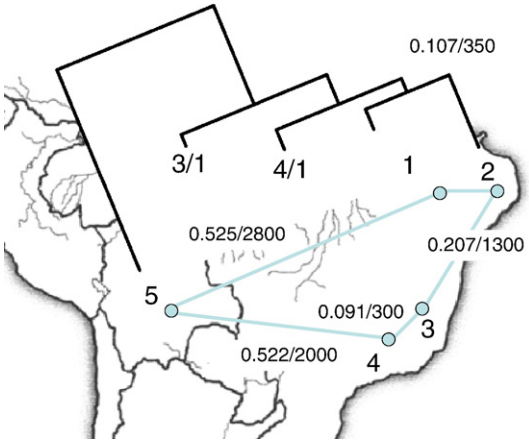

Fig. 2. Geographic and genetic distances between subpopulations of *M. domestica*. The map of South America shows the approximate locations of the source populations from which the laboratory stocks used in this study were derived. Lines between populations show the genetic distance/physical distance between those populations (genetic distances are based on shared polymorphism and determined by the program Gendist; physical distances are in kilometers). For populations #3 and #4, the genetic distances are for the stock populations, which are both admixed with individuals from population #1. The phylogenetic tree shown was obtained using the program Contml and the complete set of Gendist distances.

frequency (LF) loci had frequencies between 0.000 and 0.333. Primer pairs with multiple samples that did not produce a clear product, or exhibited products of aberrant size were excluded from further analysis, and pairs with apparent heterozygosity in a majority of individuals were identified as possible paralogs, as reported previously (Batzer et al., 1991). Putative paralogs were investigated using the most recent genome assembly, MonDom4 (UCSC genome browser, January 2006; (Hinrichs et al., 2006)).

Analysis of population structure using the polymorphic *SINE1* loci were performed using Structure 2.1 (Pritchard et al., 2000). For each individual, Structure 2.1 estimates the proportion of ancestry from each of *K* clusters. After determining appropriate run lengths as recommended (Pritchard et al., 2000), we used a burn-in of 50,000 iterations and a run of 10,000 replications. Five replicate runs were performed on the dataset using values of *K* ranging from 2 to 9. Genetic distances among populations were calculated using the CONTML and GENDIST programs in Phylip 3.6 (Felsenstein, 1989).

2.2. SINE1 phylogeny, distribution, and GC content analysis

We parsed all the *SINE1* elements in the RepeatMasker annotation (Smit et al., 1996–2004) of the *M. domestica* genome (2006 January) from UCSC (Hinrichs et al., 2006). The density distribution of all *SINE1* elements along the genome was calculated by determining the distance between each set of 100 adjacent *SINE1* elements. The background GC content of the genome was calculated in 2 Kb segments, excluding all known *SINE1* elements, while the GC content of regions adjacent to *SINE1* elements was calculated based on 1 Kb on each side of an identified *SINE1* element. All full-length elements were then aligned with the *SINE1_MD* consensus sequence from the RepBase library (version 11.06; (Jurka et al., 2005)) using BLASTN (version 2.2.13; (Altschul et al., 1990)), and a well-aligned subset (covering the first 160 bp of the element) of these full-length *SINE1s* was selected for further analysis. This alignment was used to build *SINE1* subfamilies and phylogenies (Price et al., 2004).

2.3. Substitution/mutation rate analysis

We used a likelihood based estimation algorithm to estimate the substitution rate in different *SINE1* subfamilies (Arndt and Hwa, 2005). This method explicitly allows a different substitution rate for CpG dinucleotide pairs, which is essential for substitution rate analysis. The use of the substitution rate as a measure of the mutation rate requires an assumption of neutral evolution of *SINE1* sequences, which is plausible for sequences in the SINE family (Petrov and Hartl, 1999). Here, we used the consensus sequence of each subfamily as the ancestral sequence (Price et al., 2004), and compared them to each of the elements in the respective subfamily (over the aligned 160 bp fragment described above). In different analyses, we partitioned these elements by chromosome, region of chromosome 3, by age, and by GC content of adjacent sequences. In cases where different subfamilies were joined we determined a single average substitution process, but each element fragment was still compared to the predicted ancestor for its respective subfamily.

3. Results

3.1. Polymorphism and population structure

A family of 5449 high-copy number repeats were identified *de novo* from the MonDom1 genomic sequence (see Materials and methods), and later determined to be members of the *SINE1* family based on similarity to consensus sequences from RepBase (Jurka et al., 2005; Gentles et al., in press). We identified 454 of these elements as belonging to the most recently amplified subfamily based on the method of Price et al. (2004). Further evidence of the recent origin of these sequences was obtained from their sequence identity, indicating that there was little time for fixation of mutations subsequent to insertion (52.6% are identical to the consensus sequence, Supplementary Table A). Subsequent subfamily analysis of all RepeatMasker *SINE1* sequences (see below) put the majority of these in subfamily #15.

We determined genotypes for 120 insertion loci in a panel of 43 *M. domestica* individuals obtained from five different laboratory stocks derived (sometimes with admixture) from five geographically distinct populations and maintained to promote diversity (Fig. 2). We found that 50% of the 120 scorable loci were polymorphic (excluding twelve probable paralogs and four loci that exhibited no filled sites in any of the individuals in our panel (Supplementary Table B), indicating low-frequency, potentially *de novo*, or private insertions). Compared to the subfamily consensus sequence of the loci tested, the average distance for the fixed or high frequency inserts was $0.0111 +/- 0.0011$, and that of low or intermediate frequency

inserts was 0.078+/−0.0013. This may indicate that the mutation rate is fast enough and the time to fixation is slow enough that sequence mutations often accumulate during the insertion fixation process; alternatively, this pattern may reflect sequence diversity in the source elements.

Three loci were polymorphic only within the source population from which the individual for the *M. domestica* genome sequence was obtained (population #1) or in stocks admixed from this population. Furthermore, a total of 30 loci were either polymorphic or fixed for insertion in the source population, but were entirely absent from population #5, the most geographically remote from the source population (see below). These results support the hypothesis of recent amplification activity in the *SINE1s*.

Genetic distances among the three pure-origin stock populations (#1, #2, and #5) reflected geographic distance (Fig. 2). Among the three pure populations, population #2 was much closer to population #1 (genetic distance is 0.107) than was population #5 (genetic distance is 0.525). The admixed populations (#3 and #4) obviously have some degree of affiliation with population #1, regardless of the unknown similarities of the source populations. Their position on the branch leading to population #5 indicates that population #2 is closest to population #1, and that populations #3 and #4 are more genetically distinct, with genetic distances 0.100 and 0.076, respectively. The unknown level of admixture with population #1 prevents certainty as to how different they are. Although geographically closer to population #1, genetic distances place population #3 farther away on the tree than population #4. Population #3 may simply be more highly admixed, leading to the observed topology. The known population structure and number of populations was not provided to the Structure 2.1 program (Pritchard et al., 2000), but the data set was sufficiently large that there was consistent support for five populations. These five populations were typically clearly defined with 92.7% or greater of the individuals clustering with their correct cohort. Although the coherence of these populations may be artificially enhanced by their history as laboratory stocks (despite attempts to keep them outbred and maintain their genetic diversity), these results confirm the utility of SINE insertion polymorphisms for population genetic analyses in general, and the utility of the polymorphic *SINE1* markers we have identified for further dissection of natural population genetic structure in *M. domestica*.

Twelve loci exhibited amplification patterns (such as excess heterozygotes) indicative of paralogous insertions (i.e., elements that have inserted into duplicated regions), and we investigated these loci to identify possible evolutionary scenarios. In one case, MD0002_11, the *SINE1* element inserted into an LTR family element, ERV2-LTR. The primers hybridized equally well to as many as 107 members of this family, each with similarity to the original locus of 94% or greater, but only one contained the *SINE1* insertion. Similarly, MD0002_111 inserted into an L1 element. For example, one locus, MD0002_71, was heterozygous in all individuals, and has at least three potential paralogous insertion sites arising from genomic duplications. The reasons for unusual amplification patterns at the remaining loci were unclear, but some amplification patterns suggested paralogous insertions only in some subpopulations. For example, at four loci in population #5, all individuals were apparently heterozygous, while in the other populations all individuals were fixed for insertion. The likelihood of this occurring by chance is miniscule, which suggests that paralogous insertions may be involved. Since population #5 is the most distant population geographically and genetically, one of the paralogous loci may have become fixed in alternative states in the two populations.

3.2. Evolutionary history of SINE1 subfamilies

When a library of over 300 repetitive element subfamilies was deposited in RepBase and RepeatMasker annotation became available for MonDom4 (see Materials and methods), evolutionary analysis was carried out on the set of 603,385 *SINE1* elements (based on subsequent modification of RepBase, some of these elements are from a newly recognized subfamily, *SINE2*). The 376,378 *SINE1* elements with length greater than 175 bp were aligned with the RepBase *SINE1-MD* consensus sequence using BLAST. A well-aligned core region of 160 bp was identified, and we found that 114,302 elements aligned extremely well with the 160 bp consensus sequence over its entire length. Subfamily assignments were then determined (Fig. 1) using the method of Price et al. (2004). The number of members in each family ranged from 52 to 18,328. *SINE1* subfamilies are apparently also organized much like *Alu* subfamilies, with most sequences arising from ancestral "master" progenitor sequences that slowly diverge over time. This makes the *SINE1* elements very useful for analysis of substitution patterns along the recent opossum lineage. Since most elements evolve neutrally after landing in a new site in the genome and will be non-functional, many families have thousands or tens of thousands of independently evolving loci, each of which will replicate the substitution pattern leading from the common ancestor (Arndt et al., 2003b). Since unselected neutral substitution processes reflect the mutation process (Hartl and Clark, 1997), differences in mutation patterns can also be inferred for different chromosomal regions.

We selected a subset of 38 of these families with over 400 members for evolutionary rate analysis using the method of Arndt and colleagues (Arndt et al., 2003a; Arndt and Hwa, 2005). The model included six single site substitution rate parameters (symmetric rates on the two strands were assumed equal) plus a separate CpG substitution rate parameter. This method estimates substitution rates based on divergence from a consensus sequence, and allows for non-reversible models as well as inclusion of a separate CpG dinucleotide substitution rate. It also explicitly accounts for different numbers or even complete absence of CpG dinucleotides in the sequence. It is thus an essential extension from simpler but inaccurate general time reversible models, and can fully account for unequal nucleotide frequencies. A preliminary analysis had indicated that symmetric rates were nearly equal, and provided strong support for inclusion of a CpG rate parameter (based on log likelihood ratios). Average transversion rates in the families ranged from 0.0028 to 0.0197 per site (a factor of 7), and there was no evidence that the transition or transversion

substitution process was different in the young versus the old subfamilies (Fig. 3). CpG substitution rates are discussed in detail below. A:T⇒G:C transitions occurred at 2.65 times the average transversion rate, while C:G⇒T:A transitions occurred at 7.50 times the average transversion rate.

If it is assumed that the transversion rate in opossum *SINE1* is the same as in human *Alu* sequences (where an average transversion frequency of 0.01 corresponds to 35 myr; (Arndt et al., 2003b)), this would correspond to ages in the range of 10 to 69 million years ago (MYA) for these *SINE1* families. We note that while the constant relationship between transitions and transversions (Fig. 3) suggests that the transversion rate is stable over recent marsupial evolution, the equivalence of the marsupial and human transversion rates is an assumption made only for explanatory convenience, and does not affect any of the results presented below; should a more reliable absolute estimate of marsupial rates be obtained, the age estimates presented here should then simply be rescaled. The predicted ancestral subfamily (#21) was oldest (Fig. 1), and the youngest subfamily with large numbers (#15) provided nearly all of the sequences used for primer design and amplification. Subfamily #15 was used as an exemplar "young" subfamily in subsequent analyses. Interestingly, polymorphic loci were detected for members of a number of other subfamilies besides subfamily #15. This may indicate that there are still active "master" sequences for these subfamilies, although it should be considered that subfamily assignment is probabilistic, and some or all of these assignments could be "true" subfamily #15 members that suffered homoplasic sequence substitutions convergent with another subfamily that caused them to become misclassified.

3.3. Distribution and GC content adjacent to SINE1 sequences

The density of *SINE1* sequences was observed to vary among different regions of the *M. domestica* chromosomes ($N=9$), following at least two distinct patterns. Density was measured by determining the spacing between every 100 *SINE1* elements, and while much of the genome had highly variable densities, ranging from 50 to 300 *SINE1s* per Mb(spm), there were extended regions

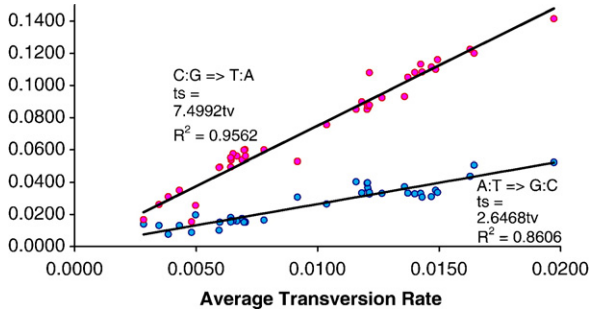

Fig. 3. Change in relative rates over time. The average transition rates (C:G⇒T:A in pink, A:T⇒G:C in blue) for *SINE1* subfamilies with more than 400 members are shown versus the average transversion rate in the same subfamily. All rates were calculated using the method of Arndt and Hwa (2005), including explicit accounting of accelerated CpG dinucleotide transition rates (not shown). The linear regression of each transition rate (ts) on transversion rate (tv) is also shown; the highly significant regressions indicate that the relative rates have not substantially changed over time. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

with consistently higher densities averaging in the range of 300 to 600 spm. Examples are shown for chromosomes 3 and X in Figs. 4 and 5, respectively (other chromosomes are not shown). These broad regions of high *SINE1* density had higher GC content minima than the low-density regions, and occasionally had much higher GC content (50–55%; Fig. 4). The 2 kilobase (Kb) of sequence adjacent to *SINE1* elements tended to be about 4% higher in GC content than the average background, indicating that *SINE1s* tend to be located in regions of high GC content. In the *SINE1*-dense regions, however, sequences adjacent to *SINE1s* were not much more than 37–46% GC, even when some background sequence regions were as high as 55% GC. This indicates that the GC preference might be better described as a tendency to avoid regions low in AT content. (We note that here and elsewhere in the manuscript, *SINE1* elements that happen to lie in the adjacent regions are not considered in the GC content calculations; thus, these measurements are not affected by the density of *SINE1* insertion).

In general, the telomeres also tended to be relatively GC-rich (Figs. 4 and 5 and Supplementary Fig. A), but this was not necessarily reflected in increased SINE density (e.g., compare the background GC versus *SINE1* densities at the ends of chromosome 3 in Fig. 4). The relationship between spm (μ) and GC content was well described by a power function ($\%GC = 22.547 * \mu^{0.0981}$, $R^2=0.5033$, $n=814$; or $\%GC = 0.183\mu + 33.82$, $R^2=0.4754$, $n=814$; probabilities were infinitesimal in both cases; see Fig. 4D inset). It is also notable that the *SINE1* sequences themselves have different frequencies depending on whether they are in *SINE1*-dense regions or not (Fig. 4). The X chromosome, although relatively short [61 Mb; (Mikkelsen et al., in press)], has extended regions at the beginning and middle that are relatively high in background GC content, tend to have higher GC content in SINE-adjacent sequences, and tend to have higher *SINE1* densities (Fig. 5). Notably, even very short regions of high background GC content are regularly matched by short regions of higher *SINE1* density (Fig. 5).

The GC preference of *SINE1* elements could be due to preferential insertion in GC-balanced regions or preferential loss in AT-rich regions. We found, however, that for the loci that were genotyped, the GC content in the 1 Kb of sequence adjacent to polymorphic loci was 36.5% (+/−1.4%), while in sequence adjacent to fixed loci it was 37.3% (+/−1.1%). These differences are not significantly different from each other or from the genome average of 37.7% (Mikkelsen et al., in press). We note that the tendency is that sequences next to polymorphic loci are slightly less GC-rich than adjacent to fixed loci, and it is possible that a large increase in the number of loci tested would detect a significant difference if the difference is real but small.

We also considered how adjacent GC content was distributed among *SINE1* subfamilies of different ages, and found a significant regression (*P*-value $B=4.8*10^{-6}$) of GC content on subfamily age (Supplementary Fig. B). Subfamily ages were estimated based on average transversion rate estimates (see below) and the assumption that transversion rates are the same in marsupials as they are in eutherian mammals. These ages are mostly clustered in two groups, from 9.9–27.3 MYA and from 36.3–52.3 MYA, with a gap of 9 million years (myr) between

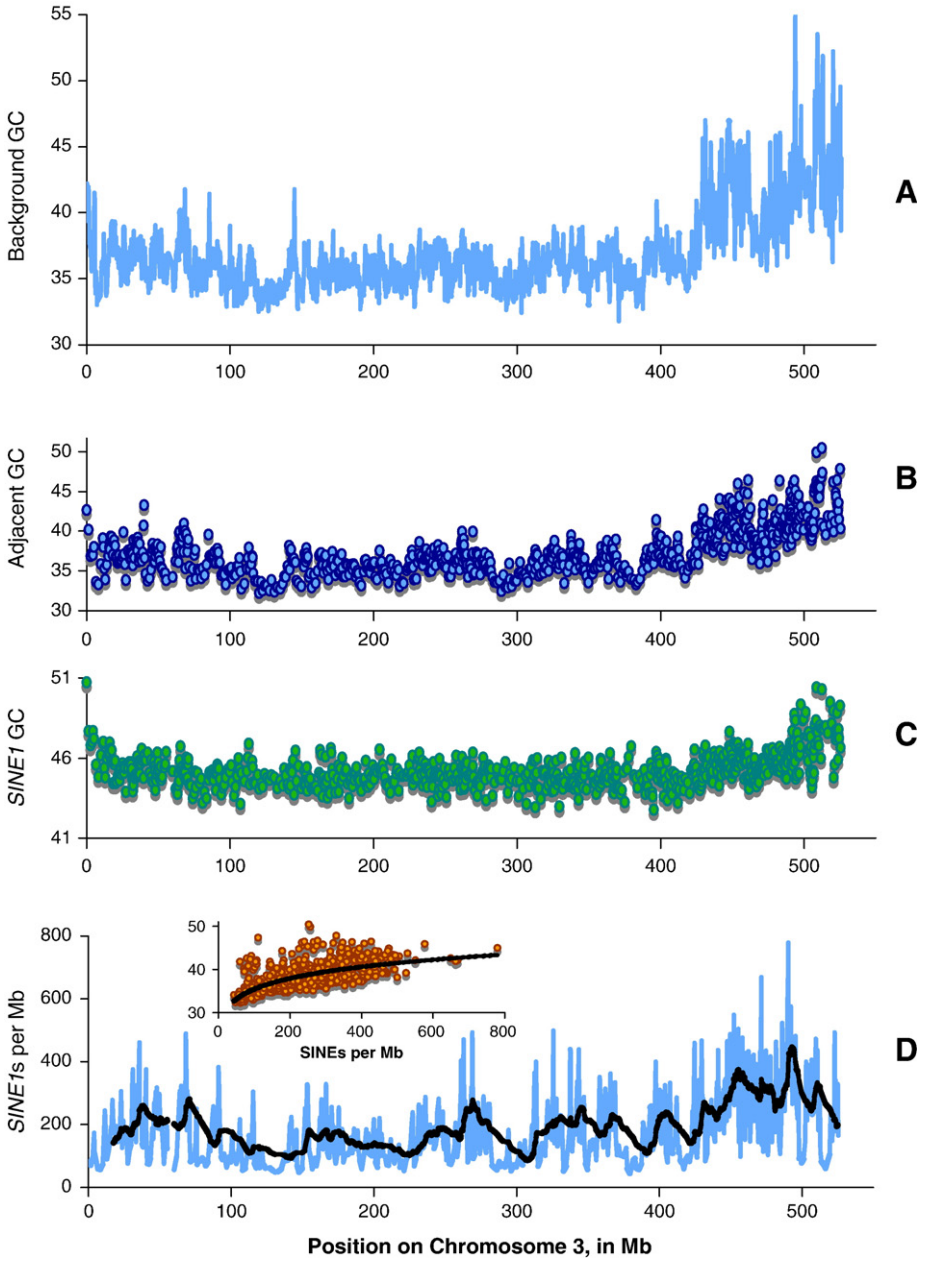

Fig. 4. GC content and *SINE1* density on chromosome 3. (A) The background GC content in 2 Kb chunks (excluding regions with *SINE1* elements). Points are averaged over 100 chunks. (B) The GC content of the 2 Kb adjacent to each *SINE1* element; points are averaged over 100 *SINE1* elements. (C) The GC content in *SINE1* elements, averaged over 100 *SINE1* elements. (D) The density of *SINE1* elements, in *SINE1s* per Mb. Each point represents the scaled inverse of the distance between every 100 *SINE1* elements. In the inset, the GC content in the 2 Kb adjacent to each *SINE1* (averaged over 100 *SINE1* loci) is graphed compared to the density of *SINE1s* per Mb. The power relationship curve shown is $y=22.55x^{0.0981}$ ($R^2=0.5033$).

them (Supplementary Fig. B). There are also three somewhat older families. The average GC content of the younger cluster is 37.0%, slightly below the genome average, but the average GC content for the older cluster, 37.8%, is significantly higher (and right about at the genome average). The average GC content of the oldest set of families, 39.5%, is higher still (also significant, and well above the genome average). Interestingly, there is one outlier subfamily, at 32.1 MYA in the center of the 9 myr "gap" between subfamily clusters, with an extremely high average GC content of 42.9% (Fig. 6). Altogether this indicates that the *SINE1* GC preference is limited to older SINE families, and that young SINEs are essentially "neutral" genomic residents, neither preferentially inserted according to GC content, nor selected against based on GC content during the fixation process.

3.4. Do SINE1 substitution rates explain local GC content?

To understand the causal reasons for the genomic structure described above, it is important to know the relationship between substitution rates and genomic structure, and in particular whether local or large-scale chromosomal structure dominates in determining substitution processes. GC content is,

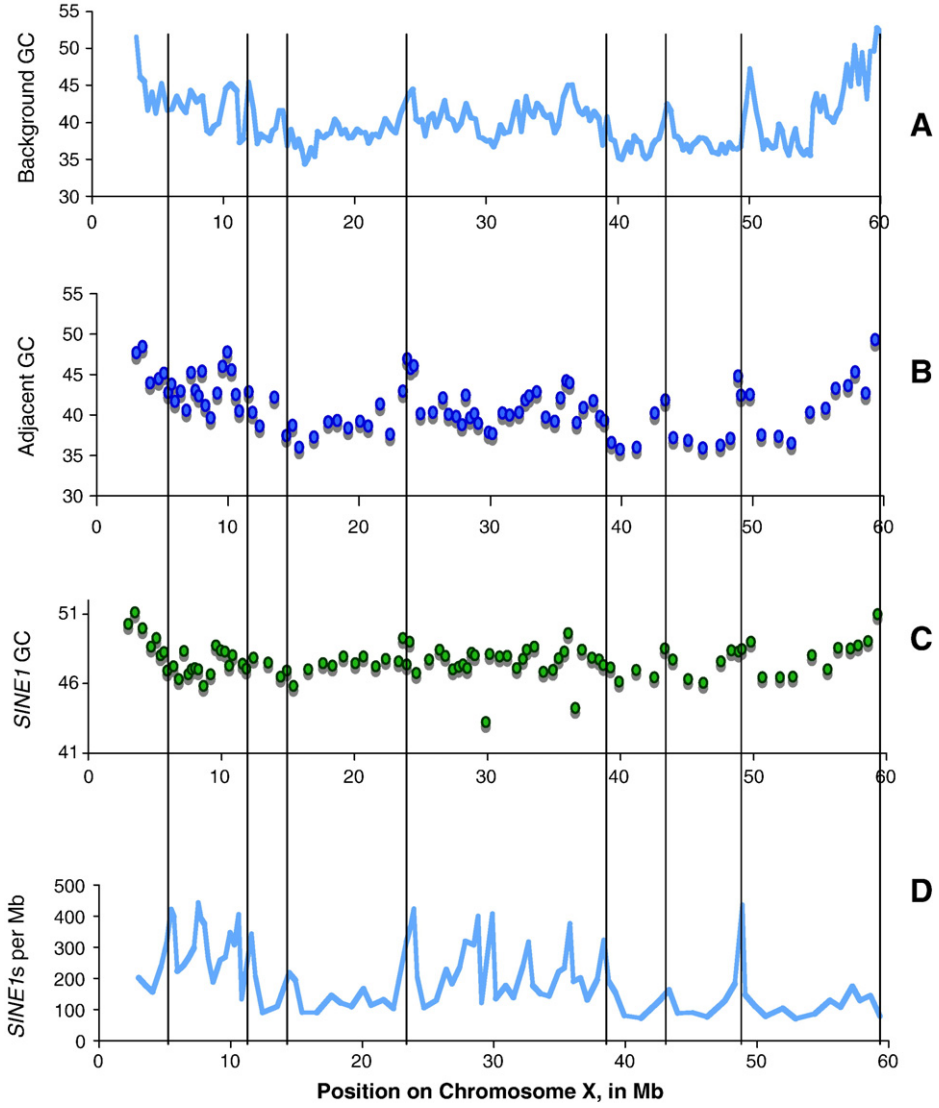

Fig. 5. GC content and *SINE1* density on chromosome X. Subfigures A, B, C, and D are as in Fig. 4. The vertical lines show the correspondence between peaks in the respective subfigures.

of course, only an outcome of evolutionary processes, and studying variation in the processes themselves can be far more enlightening than studying GC content alone. We evaluated the relationship between substitution rates and local GC content by determining evolutionary rates for *SINE1* elements clustered into bins according to the GC content of adjacent sequences (Fig. 7; note that as mentioned earlier, *SINE1* elements that happen to lie in the adjacent regions are not considered). Variation in the substitution process is strongly related to GC content. It is particularly notable that the rate of A:T ⇒ G:C transitions is nearly 2.5 times greater in the highest GC content bin than in the lowest GC content bin, in sync with A:T ⇒ C:G transversions, while G:C ⇒ A:T transitions and G:C ⇒ T:A transversions decrease in a nearly reflective manner. In contrast, A:T ⇒ T:A and C:G ⇒ G:C transversions are relatively indifferent to GC content (Fig. 7). This is consistent with predictions from a putative biased gene conversion process that affects all GC content-altering substitutions simultaneously, but leaves content-neutral substitutions unaffected.

To test whether there were further effects arising from large-scale chromosomal structure, we compared the somewhat unusual X chromosome (Mikkelsen et al., in press) to the autosomes, and also compared two regions of chromosome 3 that have markedly different *SINE1* insertion densities. It has been observed (based on substitution rates at 3rd codon positions) that genes with conserved placement on the X chromosome in the human and opossum genomes evolve about 20% faster than other genes (Mikkelsen et al., in press). Based on a comparison of SINE substitution rates among chromosomes, the high GC content X chromosome is markedly different (Table 1). In particular, five substitution rates in the X chromosome are more than two standard deviations greater than the rates measured in the other chromosomes, and four of these differences act to increase GC content.

Overall, whether autosomal rates are averaged by chromosomes, or determined for the whole genome, there is a 23% increase in the expected stationary GC content of the X chromosome compared to the autosomes (Table 1). The predicted

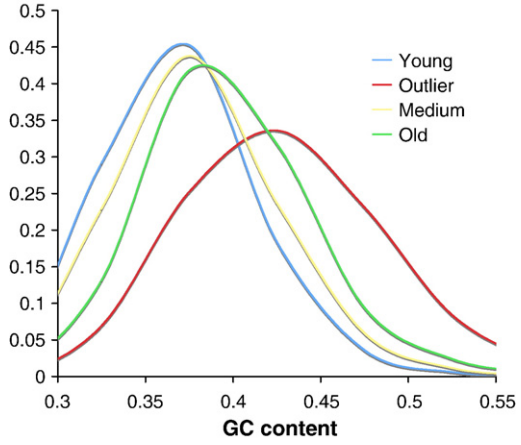

Fig. 6. Distribution of GC content in *SINE1*-adjacent sequence according to subfamily age. The GC content distribution was averaged over three sets of young (9.9–27.3 MYA), medium (36.3–52.3 MYA), and old (57–69 MYA) subfamilies. The GC content distribution for the outlier subfamily #7 (32.1 MYA) is also shown.

Table 1
Substitution rates per site on chromosome X versus substitution rates across the autosomes in *SINE1* subfamily #15

	Autosome average	+/−	Joint non-X	X	X/ auto	X/ non-X
A:T→C:G	0.0025	0.000478	0.0024	0.0039[a]	1.61	1.62
A:T→T:A	0.0024	0.000390	0.0025	0.0014[a]	0.56	0.56
C:G→G:C	0.0026	0.000406	0.0026	0.0028	1.09	1.08
G:C→T:A	0.0064	0.000776	0.0064	0.0059	0.91	0.92
Mean tv	0.0035	0.000391	0.0035	0.0035	1.01	1.01
A:T→G:C	0.0127	0.000924	0.0130	0.0158[a]	1.25	1.21
C:G→T:A	0.0262	0.000806	0.0263	0.0238[a]	0.91	0.90
Mean ts	0.0194	0.000677	0.0196	0.0198	1.02	1.06
CpG→CpA or TpG	0.3629	0.029961	0.3671	0.3233[a]	0.89	0.88
ts/tv	5.6	0.67	5.7	5.7	1.01	1.01
CpG/ts	18.7	1.45	18.7	16.3[a]	0.88	0.87
CpG/tv	104.6	12.6	105.9	92.4	0.88	0.87
StationaryGC	0.28	0.011	0.28	0.35[a]	1.23	1.23

[a] Values for the X chromosome differing from the autosomal average by more than two standard deviations.

stationary GC contents are quite low (0.28 for the autosomes, 0.35 for the X chromosome), and autosome sequences are generally 2–7% above equilibrium values, while X chromosome sequences are generally closer to equilibrium. These observations are consistent with the idea that the marsupial X chromosome is evolving more quickly than the autosomes. The generally high AT content in most of the genome (twice the GC content) means that increases in the A or T ⇒ G or C per site substitution rates will lead to a large number of excess observed substitutions. Such substitutions will further accelerate the substitution rate indirectly by creating fodder for the extremely rapid CpG methylation-deamination mutation process.

To determine if these rate changes were specific to the X chromosome, or were a feature of SINE-dense regions in general, we compared the SINE substitution process for loci on the left 3/4 of chromosome 3 versus the right 1/4 (see Fig. 4; the left end is defined as the end where chromosome position numbers begin, and which is closest to the centromere). The predicted differences in substitution rates (Table 2) were quite similar to the differences seen for the X versus autosome comparisons. The only notable difference was that for the chromosome 3 comparison, the SINE-dense region also had significantly fewer G:C ⇒ T:A transversions (in the X comparison there were fewer than in the autosomes, but within two standard deviations of the average autosome estimate). There is also 23% greater predicted stationary GC content in the SINE-dense region of chromosome 3.

Given these results, it is prudent to consider whether localization in SINE-dense regions affects substitution rate beyond the fact that these regions tend to be GC-rich. We addressed this by considering whether sequences with similar adjacent GC content (0.3–0.4) had different substitution rates depending on

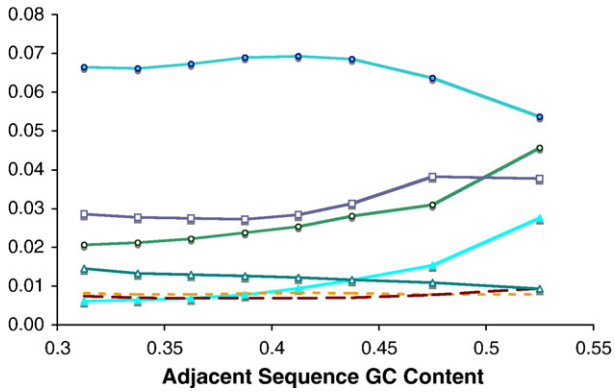

Fig. 7. Substitution rates versus adjacent GC content. All seven substitution rates in *SINE1* elements are shown as a function of adjacent nucleotide sequence GC content. *SINE1s* were grouped into approximately equal-sized bins, and the average GC content of each bin is graphed. Transition rates are shown with circles; A:T ⇒ G:C circles are open, C:G ⇒ T:A circles are closed. Transversion rates that alter GC content are shown with triangles; G:C ⇒ T:A transversions are shown with open triangles, and A:T ⇒ C:G are shown with closed triangles. Transversion rates that do not affect GC content are shown without symbols (long dash is A:T ⇒ T:A, short dash is C:G ⇒ G:C). CpG substitution rates are shown with open squares, and were divided by 10.0 to allow visualization on the same graph.

Table 2
Substitution rates in SINE-dense and SINE-deficient regions of chromosome 3 in *SINE1* subfamily #15

	Chromosome 3, all	Chromosome 3, left 3/4	Chromosome 3, right 1/4	Right/ left
A:T→C:G	0.0025	0.0021	0.0034[a]	1.61
A:T→T:A	0.0027	0.0028	0.0024[a]	0.85
C:G→G:C	0.0025	0.0026	0.0025	0.97
G:C→T:A	0.0062	0.0065	0.0055[a]	0.85
Mean tv	0.0035	0.0035	0.0034	1.07
A:T→G:C	0.0127	0.0120	0.0143[a]	1.19
C:G→T:A	0.0266	0.0275	0.0245[a]	0.89
Mean ts	0.0197	0.0198	0.0194	1.04
CpG→CpA or TpG	0.3550	0.3566	0.3512	0.98
ts/tv	5.6	5.7	5.6	1.00
CpG/ts	18.1	18.0	18.1	1.00
CpG/tv	101.9	102.0	101.8	1.00
StationaryGC	0.28	0.26	0.32[a]	1.23

[a] Values for the left and right of chromosome 3 differing from each other by more than two standard deviations of the autosomal average.

Table 3
Substitution rates in SINE-dense and SINE-deficient regions of chromosome 3; includes *SINE1s* in all subfamilies with adjacent GC frequencies in the range of 0.3 to 0.4

	Chromosome 3, left 3/4	Chromosome 3, right 1/4	Right/ left
A:T→C:G	0.0067	0.0078	1.16
A:T→T:A	0.0082	0.0076	0.93
C:G→G:C	0.0070	0.0064	0.91
G:C→T:A	0.0138	0.0117	0.85
Mean tv	0.0089	0.0084	0.94
A:T→G:C	0.0223	0.0228	1.02
C:G→T:A	0.0686	0.0637	0.93
Mean ts	0.0455	0.0432	0.95
CpG→CpA or TpG	0.2648	0.3007	1.14
ts/tv	5.0856	5.1643	1.02
CpG/ts	5.8254	6.9530	1.19
CpG/tv	29.6260	35.9071	1.21
Stationary GC	0.2609	0.2619	1.00
Mean GC	0.3545	0.3686	1.04

where they were located (i.e., in a SINE-dense or SINE-deficient region). Pooling results from all families and comparing the left 3/4 of chromosome 3 to the right 1/4 of chromosome 3, the differences in substitution rates are quite small (Table 3). In particular, the A:T⇒G:C transition rate is only 3% larger in the right half than the left. The observed differences can be explained by the slight differences in mean GC content in the two bins. Thus, the differences in substitution processes between SINE-dense and SINE-deficient regions appear to reflect different frequencies of high-GC regions rather than processes that are unique to each region.

The above comparison also provides us with a further estimate of the relative CpG substitution rate and how it changes over time. The CpG substitution rate for subfamily #15, the largest young family, was 18.7 times greater than the average transition rate, and 105.9 times the average transversion rate (Table 1). This high relative rate (about double to that in humans) helps to explain the extremely low frequency of CpG dinucleotides in the opossum genome (Mikkelsen et al., in press). The relative CpG rate averaged across other young subfamilies is, however, somewhat lower (46–48 compared to the transversion rate, and depending on adjacent GC content), and older subfamilies have still lower CpG rates (32–40 times the transversion rate; Table 4). This indicates a general increase in CpG rates over time.

4. Discussion

Despite the observation that *SINE1* and *Alu* are distinctly different repetitive elements in taxon groups (eutherian and metatherian mammals) separated by at least 160 myr of evolution, their common biology as non-autonomous LINE-driven elements creates many similarities in their genomic features. From age estimates of *Alu* subfamilies, the oldest dimeric *Alu* subfamily (*AluJo*) was dated at 82 MYA, and the youngest (the broadly defined *AluY* set of elements) at 30 MYA (Arndt et al., 2003b). If the average transversion rate in marsupials has been similar (but note caveats above), the *SINE1* family in *M. domestica* genome arose 13 myr later than dimeric *Alu*s did in primates. As with *Alu* elements (Belle and Eyre-Walker, 2002; Cordaux et al., 2006), young *SINE1* element locations are neutral to adjacent GC content, while old *SINE1* are more

Table 4
Substitution rates and stationary GC content across categories of low (<42.5%) and high (>42.5%) adjacent GC content and young (9.9–27.3 MYA), medium (36.3–52.3 MYA), and old (57–69 MYA) subfamily clusters

Rates	Young-low	Young-high	YL/YH	Med-low	Med-high	ML/MH	Old- low	Old-high	OL/OH
A:T→C:G	0.005	0.008	1.755	0.011	0.019	1.795	0.015	0.020	1.394
A:T→T:A	0.005	0.005	0.980	0.011	0.010	0.889	0.017	0.014	0.819
C:G→G:C	0.005	0.005	0.993	0.010	0.009	0.911	0.015	0.013	0.899
G:C→T:A	0.008	0.007	0.791	0.018	0.014	0.754	0.023	0.016	0.698
TA/GC tv	1.77	0.80	0.45	1.74	0.73	0.42	1.57	0.79	0.50
Mean tv	0.006	0.006	1.074	0.012	0.013	1.036	0.017	0.016	0.917
Realized tv	0.011	0.012	1.126	0.024	0.026	1.087	0.034	0.032	0.945
Estimated age	19.777	21.237		43.266	44.817		60.609	55.602	
A:T→G:C	0.015	0.019	1.228	0.032	0.037	1.156	0.046	0.055	1.195
C:G→T:A	0.045	0.042	0.920	0.093	0.082	0.884	0.130	0.109	0.838
TA/GC ts	2.96	2.21	0.75	2.91	2.22	0.76	2.80	1.96	0.70
Mean ts	0.030	0.030	0.998	0.063	0.060	0.953	0.088	0.082	0.932
Realized ts	0.026	0.029	1.115	0.054	0.058	1.060	0.078	0.080	1.027
Sum rate	0.037	0.041	1.118	0.078	0.083	1.068	0.112	0.112	1.002
CpG→CpA/TpG	0.269	0.277	1.031	0.448	0.460	1.028	0.555	0.640	1.153
GC content									
Predicted stationary	0.25	0.32	1.29	0.25	0.33	1.30	0.26	0.33	1.28
Observed mean	0.36	0.45	1.25	0.37	0.46	1.24	0.38	0.46	1.22
Rate ratios									
CpG/tv	47.60	45.72		36.21	35.94		32.04	40.27	
CpG/ts	8.91	9.21		7.16	7.71		6.30	7.79	
ts/tv	5.34	4.96		5.06	4.66		5.09	5.17	
ts_{GC}/tv	2.70	3.09		2.59	2.89		2.68	3.49	
ts_{AT}/tv	7.98	6.84		7.53	6.43		7.50	6.85	

commonly located in GC-rich regions. We have demonstrated that *SINE1* elements are recently active in the *M. domestica* genome, but there is no significant difference in adjacent GC frequencies between polymorphic and fixed *SINE1* loci. Patterns of varying GC content in the opossum are not organized into clear isochores, but rather there are a few long stretches of high GC content, while most of the genome has low GC content peppered with short bursts of high GC content sequence. Nevertheless, *SINE1* elements are dense in the high GC content regions wherever they occur, except at the chromosomal termini.

We infer that simple point mutations have not changed much over the course of recent marsupial evolution, since there is little evidence of differences in the substitution process between old and young *SINE1* families (Fig. 3). Despite this, most sequences do not appear to be in stationary equilibrium. Compared to relative transition rates in the human genome, *SINE1* elements have a slightly lower A:T⇒G:C transition rate (2.65 in *Monodelphis SINE1* elements versus 2.74 in Human *Alu* elements) and a much larger C:G⇒T:A transition rate (7.5 in *Monodelphis SINE1* elements versus 5.5 in human *Alu* elements; (Arndt et al., 2003b)). The difference in these two transitions helps to explain the low GC content in opossum genomes (Mikkelsen et al., in press).

Substantial differences in substitution processes, both transition and transversion substitutions, are a likely cause of the differences in GC content in different regions of the genome. The different substitution rates on X chromosomes versus autosomes (Mikkelsen et al., in press) appears to be almost entirely driven by the same underlying process. It has been recently observed that GC content at 3rd codon positions in the *RAG1* gene vary widely among marsupial lineages (Gruber et al., 2007). Since *RAG1* 3rd codon positions have a fairly high GC content (58–59% in genus *Monodelphis,* higher in most other genera), this indicates that the underlying GC-rich substitution process may be highly variable over evolutionary time, even though the average (mostly AT-rich) process in the recent ancestral *Monodelphis* lineage is relatively constant.

In the human genome, the GC-dependence of substitution rates appears to have been consistent with isochore structure (alternating GC-rich and GC-poor regions) only in the distant past (before 90 MYA). For the last 90 myr there is only weak correspondence between substitution rates and consequent stationary equilibria, and the observed GC content; regions with high GC content are therefore predicted to eventually disappear from the human genome (Duret et al., 2002; Arndt et al., 2003b). In contrast, *SINE1* substitution rates in *Monodelphis* are expected to maintain isochore structure (although GC content in all regions is expected to decrease).

Although isochore structure is consistent with substitution processes in the opossum, we did not see clear isochore structure in the sense of alternating patches of GC content, as in most mammalian genomes (Bernardi, 2000; Arndt et al., 2003b; Gu and Li, 2006; Webster et al., 2006). Instead, much of the genome consists of long stretches of GC-poor regions only occasionally peppered with short GC-rich stretches, and there are relatively few extended regions with generally high GC-rich segments. It has been suggested that the opossum does not have clear genome-level isochore structure because the reduction in recombination rates in the opossum due to reduction in chromosome number ($2n=18$) would have led to a reduction in biased gene conversion (Gu and Li, 2006). Since biased gene conversion is thought to be a key factor causing GC-biased substitution rates in GC-rich regions, the logic is that reduction in this factor will reduce the GC bias. Our results do not contradict the possibility of biased gene conversion as a causative agent, however, since reduction of the average recombination rate appears to equally affect all regions, reducing predicted GC equilibria, but not destroying local isochore structure. Indeed, our substitution analyses in GC-rich regions are strongly consistent with biased gene conversion.

Although the root cause of isochore-like patches in eutherian and metatherian mammals remains a mystery, the strong linkage observed in this study between *SINE1* density, GC content, gene density (Mikkelsen et al., in press) and substitution processes that lead to high GC content are suggestive. The gene density factor (see Introduction) seems to be the one causal agent that could lead to increased *SINE1* density over long periods of time (by selecting against homologous recombination among *SINE1* elements) and also to greater exposure (reduced histone binding) during translation that could lead to differences in replication timing and increased opportunities for recombination, biased gene conversion, and mutation.

There has been only a slight increase in the relative CpG dinucleotide transition rate in recent *SINE1* elements compared to more ancient elements over most of the ancestral *Monodelphis* lineage, a very recent doubling, evident in the most recent large subfamily. Although reminiscent of the inferred CpG rate increase in primates, the difference is not nearly as great as in primates (a 4–8 fold increase) because unlike in primates, the ancestral *Monodelphis* CpG rate was already fairly high. This ancestrally high CpG substitution rate and somewhat higher modern CpG rates, along with lower levels of CG, helps to explain the extremely low CpG frequency in *M. domestica* (Mikkelsen et al., in press). We note that the relative CpG rate we estimated in opossum genome ranges from 32 to 47, which is quite close to modern eutherian mammals (39 in human *Alu*s; (Arndt and Hwa, 2005)) and chickens (Webster et al., 2006), but significantly higher than in fish (8 in the zebrafish genome; (Arndt and Hwa, 2005)) and what is predicted in ancestral mammals (90–250 MYA). If the fish rate is truly ancestral and these rate estimates are all correct, there must have been two independent jumps to a higher rate on the eutherian and metatherian lineages, and either a third jump on the ancestral avian (or archosaur) lineage, or a temporary reversion to a slower rate on the ancestral mammalian lineage.

The average transversion rates in *SINE1* subfamilies also points to fluctuations in the rate of subfamily generation along the ancestral *Monodelphis* lineage. Assuming that the average *Monodelphis* transversion rate is the same as in humans, subfamily generation was common between 9.9 and 27.3 MYA, and between 36.3 and 52.3 MYA, but was quite uncommon in the 9 myr gap in between. There is only one subfamily, *SINE1_7*, with fewer than 7000 estimated members, that

appears to have arisen in this gap, at 32.1 MYA. The phenomenological observation that this same family has an aberrantly high average adjacent GC content of 43% is striking. More evidence would be invaluable, but we note that this is consistent with the idea that the GC preference (or AT avoidance) of *SINE1* sequences is caused by a post homologous recombination mechanism. A general increase in homologous recombination rates during the 9 myr gap would tend to have preferentially targeted elements in the AT-rich regions that were young at the time (since homologous recombination depends on sequence similarity), and thus reduce the rate of generation of new subfamilies. Even small divergences between substrates have been shown to result in dramatic decrease in the efficiency of homologous recombination (Yang and Waldman, 1997; Lukacsovich and Waldman, 1999). Conversely, a pause in retrotransposition of *SINE1* elements during the 9 myr period would have induced a decrease in recombination rates, at least between *SINE1s*, due to a reduction in the availability of highly identical elements. This does not appear to be a likely causal explanation, however, since lower recombination rates would also have slowed the relative accumulation of *SINE1s* in gene-rich regions of the genome.

It is notable that based on the (possibly dubious) assumption of equivalency with the human transversion rate, the average age of even the youngest *SINE1* subfamilies in the opossum is fairly old at 9–10 MYA. Since we have demonstrated that these subfamilies are polymorphic, this might indicate that *M. domestica* subpopulations are quite old, that the average transversion rate in opossums is much higher than in humans (and that the age estimates are thus far too old), that the long-term effective breeding population size is much larger than in humans, that the subfamilies remain active for a longer time than in humans, or some combination of these explanations. Regardless of the resolution to this question, we also demonstrated that the set of polymorphic loci that we identified can be useful in resolving unknown population structure in *M. domestica*, and that our method can be used to identify active repeats in un-annotated genomes. Because of their long history and wide distribution in a variety of taxa, SINE insertions have been shown to be very effective as markers for resolving population and phylogenetic patterns (Shedlock and Okada, 2000; Okada et al., 2004; Shedlock et al., 2004; Ray et al., 2007). We expect that, as with SINE-based analyses of primate relationships (Salem et al., 2003; Ray and Batzer, 2005; Schmitz et al., 2005; Xing et al., 2005), and relationships among other taxa ranging from fish to whales (Takahashi et al., 1998; Nikaido et al., 2001; Takahashi et al., 2001; Sasaki et al., 2004; Nishihara et al., 2005), *SINE1* families will make highly accurate phylogenetic markers for resolution of deeper *Monodelphis* and other marsupial relationships.

Acknowledgements

This research was supported in part by: National Science Foundation grants BCS-0218338 (MAB) and EPS-0346411 (MAB and DDP); National Institutes of Health grants R33GM065612 (DDP), RO1GM59290 (MAB), R24RR014214 (PBS) and P41LM006252-09 (JJ); and the State of Louisiana Board of Regents Support Fund (MAB and DDP).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.02.028.

References

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Arndt, P.F., Burge, C.B., Hwa, T., 2003a. DNA sequence evolution with neighbor-dependent mutation. J. Comput. Biol. 10, 313–322.
Arndt, P.F., Petrov, D.A., Hwa, T., 2003b. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol. Biol. Evol. 20, 1887–1896.
Arndt, P.F., Hwa, T., 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21, 2322–2328.
Bailey, J.A., Liu, G., Eichler, E.E., 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. 73, 823–834.
Batzer, M.A., Deininger, P.L., 2002. *Alu* repeats and human genomic diversity. Nat. Rev., Genet. 3, 370–379.
Batzer, M.A., Gudi, V.A., Mena, J.C., Foltz, D.W., Herrera, R.J., Deininger, P.L., 1991. Amplification dynamics of human-specific (HS) *Alu* family members. Nucleic Acids Res. 19, 3619–3623.
Belle, E.M.S., Eyre-Walker, A., 2002. A test of whether selection maintains isochores using sites polymorphic for Alu and l1 element insertions. Genetics 160, 815–817.
Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.
Brookfield, J.F.Y., 2001. Selection on Alu sequences? Curr. Biol. 11, R900–R901.
Callinan, P.A., Wang, J., Herke, S.W., Garber, R.K., Liang, P., Batzer, M.A., 2005. Alu retrotransposition-mediated deletion. J. Mol. Biol. 348, 791–800.
Cordaux, R., Lee, J., Dinoso, L., Batzer, M.A., 2006. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene 373, 138–144.
Deininger, P.L., Batzer, M.A., 1999. *Alu* repeats and human disease. Mol. Genet. Metab. 67, 183–193.
Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., Galtier, N., 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics 162, 1837–1847.
Ejima, Y., Yang, L., 2003. Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. Hum. Mol. Genet. 12, 1321–1328.
Felsenstein, J., 1989. PHYLIP — Phylogeny Inference Package (Version 3.2). Cladistics 5, 164–166.
Fischer, S.E., Wienholds, E., Plasterk, R.H., 2003. Continuous exchange of sequence information between dispersed Tc1 transposons in the *Caenorhabditis elegans* genome. Genetics 164, 127–134.
Gebow, D., Miselis, N., Liber, H.L., 2000. Homologous and nonhomologous recombination resulting in deletion: effects of p53 status, microhomology, and repetitive DNA length and orientation. Mol. Cell. Biol. 20, 4028–4035.
Gentles, A.J., et al., in press. Evolutionary dynamics and biological impact of transposable elements in the short-tailed opossum *Monodelphis domestica*. Genome Res.
Gilbert, N., Lutz-Prigge, S., Moran, J.V., 2002. Genomic deletions created upon LINE-1 retrotransposition. Cell 110, 315–325.
Goodier, J.L., Ostertag, E.M., Kazazian Jr., H.H., 2000. Transduction of 3′-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. 9, 653–657.
Gray, Y.H., 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet. 16, 461–468.
Grover, D., Majumder, P.P., C, B.R., Brahmachari, S.K., Mukerji, M., 2003. Nonrandom distribution of Alu elements in genes of various functional

categories: insight from analysis of human chromosomes 21 and 22. Mol. Biol. Evol. 20, 1420–1424.
Grover, D., Mukerji, M., Bhatnagar, P., Kannan, K., Brahmachari, S.K., 2004. Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. Bioinformatics 20, 813–817.
Gruber, K.F., Voss, R.S., Jansa, S.A., 2007. Base-compositional heterogeneity in the *RAG1* locus among didelphid marsupials: implications for phylogenetic inference and the evolution of GC-content. Syst. Biol. 56, 83–96.
Gu, J., Li, W.-H., 2006. Are GC-rich isochores vanishing in mammals? Gene 385, 50–56.
Hackenberg, M., Bernaola-Galvan, P., Carpena, P., Oliver, J.L., 2005. The biased distribution of Alus in human isochores might be driven by recombination. J. Mol. Evol. 60, 365–377.
Han, K., et al., 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic Acids Res. 33, 4040–4052.
Hartl, D.L., Clark, A.G., 1997. Principles of Population Genetics, 3rd ed. Sinauer Associates, Inc., Sunderland, MA.
Hinrichs, A.S., et al., 2006. The UCSC genome browser database: update 2006. Nucleic Acids Res. 34, D590–D598.
Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., Wessler, S.R., 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature 431, 569–573.
Jurka, J., 2004. Evolutionary impact of human Alu repetitive elements. Curr. Opin. Genet. Dev. 14, 603–608.
Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.
Kass, D.H., Batzer, M.A., Deininger, P.L., 1995. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. Mol. Cell. Biol. 15, 19–25.
Kidwell, M.G., Lisch, D.R., 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution 55, 1–24.
Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.
Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Lukacsovich, T., Waldman, A.S., 1999. Suppression of intrachromosomal gene conversion in mammalian cells by small degrees of sequence divergence. Genetics 151, 1559–1568.
Mikkelsen, T.S., et al., in press. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. Nature.
Miller, W.J., Capy, P., 2004. Mobile genetic elements as natural tools for genomic evolution. Methods Mol. Biol. 260, 1–20.
Moran, J.V., DeBerardinis, R.J., Kazazian Jr., H.H., 1999. Exon shuffling by L1 retrotransposition. Science 283, 1530–1534.
Nikaido, M., et al., 2001. Evolution of CHR-2 SINEs in cetartiodactyl genomes: possible evidence for the monophyletic origin of toothed whales. Mamm. Genome 12, 909–915.
Nishihara, H., Satta, Y., Nikaido, M., Thewissen, J.G.M., Stanhope, M.J., Okada, N., 2005. A retroposon analysis of afrotherian phylogeny. Mol. Biol. Evol. 22, 1823–1833.
Okada, N., Shedlock, A.M., Nikaido, M., 2004. Retroposon mapping in molecular systematics. In: Miller, W.J., Capy, P. (Eds.), Mobile Genetic Elements: Protocols and Genomic Applications. Humana Press, Totowa, NJ, pp. 189–226.
Ostertag, E.M., Goodier, J.L., Zhang, Y., Kazazian Jr., H.H., 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am. J. Hum. Genet. 73, 1444–1451.
Pavlicek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J., Bernardi, G., 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.
Petrov, D.A., Hartl, D.L., 1999. Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proc. Natl. Acad. Sci. U. S. A. 96, 1475–1479.
Pickeral, O.K., Makalowski, W., Boguski, M.S., Boeke, J.D., 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. 10, 411–415.
Price, A.L., Eskin, E., Pevzner, P.A., 2004. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res. 14, 2245–2252.
Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.
Ray, D.A., Batzer, M.A., 2005. Tracking Alu evolution in New World primates. BMC Evol. Biol. 5, 51.
Ray, D.A., Xing, J., Salem, A.-H., Batzer, M.A., 2007. SINEs of a nearly perfect character. Syst. Biol. 55, 928–935.
Roy-Engel, A.M., et al., 2002. Non-traditional *Alu* evolution and primate genomic diversity. J. Mol. Biol. 316, 1033–1040.
Salem, A.H., Kilroy, G.E., Watkins, W.S., Jorde, L.B., Batzer, M.A., 2003. Recently integrated Alu elements and human genomic diversity. Mol. Biol. Evol. 20, 1349–1361.
Samollow, P.B., et al., 2004. First-generation linkage map of the gray, short-tailed opossum, *Monodelphis domestica*, reveals genome-wide reduction in female recombination rates. Genetics 166, 307–329.
Sasaki, T., Takahashi, K., Nikaido, M., Miura, S., Yasukawa, Y., Okada, N., 2004. First application of the SINE (short interspersed repetitive element) method to infer phylogenetic relationships in reptiles: an example from the turtle superfamily Testudinoidea. Mol. Biol. Evol. 21, 705–715.
Schmitz, J., Roos, C., Zischler, H., 2005. Primate phylogeny: molecular evidence from retroposons. Cytogenet. Genome Res. 108, 26–37.
Sen, S.K., et al., 2006. Human genomic deletions mediated by recombination between Alu elements. Am. J. Hum. Genet. 79, 41–53.
Shedlock, A.M., Okada, N., 2000. SINE insertions: powerful tools for molecular systematics. Bioessays 22, 148–160.
Shedlock, A.M., Takahashi, K., Okada, N., 2004. SINEs of speciation: tracking lineages with retroposons. Trends Ecol. Evol. 19, 545–553.
Smit, A.F.A., Hubley, R., Green, P., 1996–2004. RepeatMasker Open-3.0. at http://repeatmasker.org.
Symer, D.E., et al., 2002. Human l1 retrotransposition is associated with genetic instability in vivo. Cell 110, 327–338.
Takahashi, K., Terai, Y., Nishida, M., Okada, N., 1998. A novel family of short interspersed repetitive elements (SINEs) from cichlids: the patterns of insertion of SINEs at orthologous loci support the proposed monophyly of four major groups of cichlid fishes in Lake Tanganyika. Mol. Biol. Evol. 15, 391–407.
Takahashi, K., Terai, Y., Nishida, M., Okada, N., 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Mol. Biol. Evol. 18, 2057–2066.
The Chimpanzee Sequencing and Analysis Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.
VandeBerg, J.L., 1999. The laboratory opossum (*Monodelphis domestica*). In: Poole, T., English, P. (Eds.), UFAW Handbook on the Management of Laboratory Animals. Terrestrial Vertebrates, vol. 1. Blackwell Science, Ltd., Oxford, U.K., pp. 193–209.
VandeBerg, J.L., Robinson, E.S., 1997. The laboratory opossum (*Monodelphis domestica*) in laboratory research. ILAR J. 38, 4–12.
Webster, M.T., Axelsson, E., Ellegren, H., 2006. Strong regional biases in nucleotide substitution in the chicken genome. Mol. Biol. Evol. 23, 1203–1216.
Witte, C.P., Le, Q.H., Bureau, T., Kumar, A., 2001. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc. Natl. Acad. Sci. U. S. A. 98, 13778–13783.
Xing, J., et al., 2005. A mobile element based phylogeny of Old World monkeys. Mol. Phylogenet. Evol. 37, 872–880.
Xing, J., Wang, H., Belancio, V.P., Cordaux, R., Deininger, P.L., Batzer, M.A., 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc. Natl. Acad. Sci. U. S. A. 103, 17608–17613.
Yang, D., Waldman, A.S., 1997. Fine-resolution analysis of products of intrachromosomal homeologous recombination in mammalian cells. Mol. Cell. Biol. 17, 3614–3628.