

Survey Sequencing Reveals Elevated DNA Transposon Activity, Novel Elements, and Variation in Repetitive Landscapes among Vesper Bats

Heidi J. T. Pagán^{1,5}, Jiří Macas², Petr Novák², Eve S. McCulloch³, Richard D. Stevens³, and David A. Ray^{1,4,*}

¹Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University

²Biology Centre ASCR, Institute of Plant Molecular Biology, Ceske Budejovice, CZ-37005, Czech Republic

³Department of Biological Sciences, Louisiana State University

⁴The Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University

⁵Present address: Harbor Branch Oceanographic Institute, Florida Atlantic University

*Corresponding author: E-mail: dray@bch.msstate.edu.

Accepted: 29 March 2012

Abstract

The repetitive landscapes of mammalian genomes typically display high Class I (retrotransposon) transposable element (TE) content, which usually comprises around half of the genome. In contrast, the Class II (DNA transposon) contribution is typically small (<3% in model mammals). Most mammalian genomes exhibit a precipitous decline in Class II activity beginning roughly 40 Ma. The first signs of more recently active mammalian Class II TEs were obtained from the little brown bat, *Myotis lucifugus*, and are reflected by higher genome content (~5%). To aid in determining taxonomic limits and potential impacts of this elevated Class II activity, we performed 454 survey sequencing of a second *Myotis* species as well as four additional taxa within the family Vespertilionidae and an outgroup species from Phyllostomidae. Graph-based clustering methods were used to reconstruct the major repeat families present in each species and novel elements were identified in several taxa. Retrotransposons remained the dominant group with regard to overall genome mass. Elevated Class II TE composition (3–4%) was observed in all five vesper bats, while less than 0.5% of the phyllostomid reads were identified as Class II derived. Differences in satellite DNA and Class I TE content are also described among vespertilionid taxa. These analyses present the first cohesive description of TE evolution across closely related mammalian species, revealing genome-scale differences in TE content within a single family.

Key words: transposon, survey sequencing, Chiroptera.

Introduction

Eukaryotes typically display high proportions of genomic content derived from transposable elements (TEs). These repetitive sequences are capable of movement within the genome and are classified according to their mode of transposition. Most mammalian insertions can be attributed to Class I TEs, also known as retrotransposons. Their copy-and-paste method of mobilization can lead to substantial accumulations in a genome. For example, Class I TEs comprise at least 45% of the human genome (Lander et al. 2001) and some estimates place that number above 60% (de Koning et al. 2011). The cut-and-paste mobilization mechanism of Class II TEs (DNA transposons) has likely contributed to their low

representation in the human genome, ~3%. Similarly, low proportions were identified in other mammals: <2% of dog and opossum and <1% of mouse and rat genomes (Waterston et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005; Mikkelsen et al. 2007). However, an apparent overall lack of Class II activity in mammals in the recent past is another factor limiting the contribution of DNA transposons to mammalian genomes. Observations of various mammalian models (human, mouse, rat, and dog) have suggested a generalized shutdown of Class II TEs during roughly the same time period, ~40 Ma (Lander et al. 2001; Waterston et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005). The first identified exception to this rule is the vespertilionid bats, in particular *Myotis lucifugus*.

© The Author(s) 2012. Published by Oxford University Press on behalf of the *Society for Molecular Biology and Evolution*.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

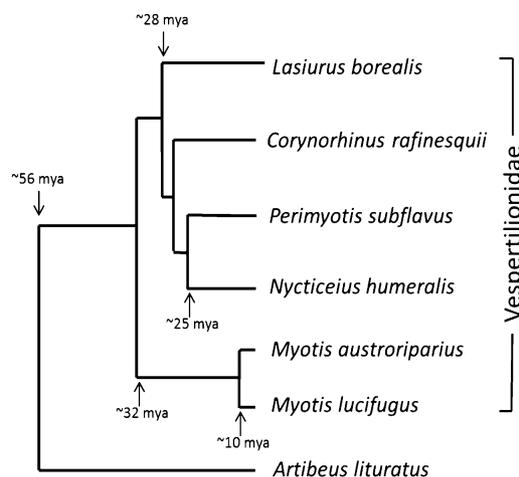


FIG. 1.—Most recent of several possible phylogenies for the surveyed taxa. Topology and vespertilionid divergence dates are taken from Lack and Van Den Bussche (2010). The date of the *Artibeus lituratus*/vespertilionid divergence is taken from Datzmann et al. (2010), and the *M. lucifugus*/*M. austroriparius* divergence is from Stadelmann et al. (2007).

While Class I TEs still dominate the overall TE landscape in *M. lucifugus*, Class II TEs have played a larger role when compared with other mammals (Ray et al. 2008). For example, members of the *Helitron* family, with their unique rolling circle amplification mechanism, have made significant contributions to genome content (Pritham and Feschotte 2007; Thomas et al. 2011). Many *Helitron* insertions as well as insertions from at least eight other Class II TE families occurred much more recently than 50 Ma and some may still be mobilizing. Interestingly, these recent DNA transposon invasions coincide with rapid diversification of *Myotis*, a genus with nearly worldwide distribution and more than 100 species (Wilson and Reeder 2005; Stadelmann et al. 2007). TE presence and activity can generate a diverse array of effects on coding sequence and expression of host genes (Kidwell and Lisch 1997; Deininger et al. 2003). In context of recent publications highlighting the capacity of TEs to drive speciation (Oliver and Greene 2009; Zeh et al. 2009; Rebollo et al. 2010), these findings provide a potential mechanism for the adaptive radiation of *Myotis*. On a larger scale, *Myotis* is a member of Vespertilionidae, the most species rich of all chiropteran families and the second most species rich family of mammals (Simmons 2005). Investigating the degree to which the elevated Class II activity found in *Myotis* extends to other bats is essential to future examinations into the potential role TEs have played in the evolution of Chiroptera as a whole and Vespertilionidae in particular.

Here, we present analyses of the TE landscapes for five additional vesper bats: *Myotis austroriparius*, *Perimyotis subflavus*, *Nycticeius humeralis*, *Lasiurus borealis*, and *Corynorhinus rafinesquii* (fig. 1). The genus *Myotis* diverged early from a monophyletic clade encompassing the other taxa, which represent a diverse sampling within Vespertilio-

nidae and thus allow us to determine if elevated Class II levels evolved singularly within *Myotis*. To serve as an outgroup, we also analyzed the phyllostomid bat, *Artibeus lituratus*. In our analyses, we applied 454-based sequencing to survey TE content. We demonstrate the utility of this method to investigating TE dynamics in nonmodel taxa that are unlikely to be the target(s) of full genome sequencing efforts, which will aid in determining the impact of Class II TEs on mammalian genome evolution.

Materials and Methods

454 Sequencing and Sequence Processing

DNA extractions were carried out on *M. austroriparius*, *P. subflavus*, *N. humeralis*, *L. borealis*, and *C. rafinesquii* using 5 PRIME ArchivePure DNA Tissue Kits. Sequencing was performed on genomic DNA at the Georgia Genomics Facility. Roche standard chemistry was used initially (*L. borealis*), but for the remaining vespertilionid samples, Titanium chemistry was utilized to accommodate lower DNA concentrations. Sample preparation and processing followed Roche protocols (October 2008).

The *A. lituratus* data was acquired separately (McCulloch and Stevens 2011); phenol–chloroform DNA extraction was used, and 454 Titanium sequencing was performed at Duke University Genome Sequencing and Analysis Core Facility using standard protocols.

Emulsion polymerase chain reaction (emPCR) drops containing only one unique template but multiple beads can produce sequencing artifacts consisting of duplicate sequences with nearly identical starting positions (Dong et al. 2011). Thus, all raw data were parsed locally using 454 Replicate Filter (<http://microbiomes.msu.edu/replicates/>) to remove these artifacts. Parameters were set at 0.95 sequence identity cutoff, 0 length difference requirement, and 3 beginning base pairs to check.

Reads derived from mitochondrial sequences were identified using BlastN. In some cases, we were able to reconstruct nearly complete mitochondrial genome sequences, and these have been discussed in a separate manuscript (Meganathan et al. 2012).

Repeat Discovery

To identify repeat content for each genome, we modified the pipeline developed by Macas et al. (2007) and updated by Novak et al. (2010). The methods were developed for plant genomic data but are applicable to mammalian genomes with minor modifications as described below. Briefly, the analysis consists of all-to-all comparison of 454 reads using mgbblast (Pertea et al. 2003) and representation of pair-wise sequence similarities exceeding the specified threshold (overlaps containing 55% or more of the longer read with 90% similarity) as edges in a virtual graph connecting the similar reads represented by graph nodes.

The reads representing different families of repetitive elements can then be distinguished as clusters (communities) of frequently connected nodes within the graph. These clusters are separated and the reads are further investigated, including their assembly into contigs using cap3 with “-o 100 -p 85” settings.

For each taxon in our analysis, a set of clusters consisting of contigs derived from overlapping reads was obtained. A cutoff was imposed to reduce the number of clusters analyzed to only include repeat families composing at least 0.01% of the respective genome of each species. Caution should be applied when extrapolating these data to the whole genome. Although our methods are apparently very good at identifying high copy number elements and moderate-to-low copy number families with high similarity, they will necessarily be inadequate for identifying very low copy number families and older highly diverged elements in a genome. In the former case, the reasoning is obviously due to the lack of whole genome coverage. In the latter case, the inadequacy is due to the combination of our assembly method and limiting ourselves to contigs with genome coverage of >0.01% of the genome. Highly divergent families or families with a large number of divergent subfamilies that each have low copy numbers would not assemble well in our analysis or produce multiple contigs that all fall below the 0.01% cutoff. Such scenarios would lead to underestimations in genome TE content.

Clusters may be representative of a particular TE family and every contig a possible consensus for a TE subfamily (Macas et al. 2007; Novak et al. 2010). However, the initial assembly resulted in individual clusters with large numbers of distinct contigs within them. For instance, *M. austroriparius* Cluster 1 contained 748 reads, 677 of which were assembled into 20 distinct contigs. Visual examination suggested that the contigs in a majority of these clusters could be assembled further to reduce the final data set without losing information. We reassembled these primary contigs in SeqMan (match size = 12, minimum percentage = 70, minimum length = 100) (DNASTAR, Madison, WI). Reassembly yielded a single contig for *M. austroriparius* Cluster 1 that was identified as the LINE element L1MAB_ML in RepBase. Similar results were obtained for other complex clusters in all examined taxa.

Consensus sequences from reassembled contigs were submitted to CENSOR to assist in classifying them into one of five categories: DNA, ERV/LTR, Non-LTR/LINE, Non-LTR/SINE, satellite, or unknown. In some cases, CENSOR returned hits to multiple TE families within a single contig. Such results could be caused by nested insertions or misassemblies and were addressed by splitting the contig into separate entries for the final library. Contigs were also queried with a custom library of bat-specific repeats derived from ongoing and previous analyses (Pritham and Feschotte 2007; Ray et al. 2007, 2008) using RepeatMasker. The library is available upon request.

Contigs from *M. austroriparius* were submitted to NCBI BlastN to query against the current whole genome shotgun (WGS) draft of *M. lucifugus* (AAPE00000000). Most contigs were found in their entirety multiple times, confirming their repetitive nature. For all taxa, contigs that could not be identified were queried against the NCBI nr database using BlastN and the protein database using BlastX.

To identify potential satellite repeats and tandem arrays, all unidentified contigs were also submitted to a local installation of Tandem Repeats Finder (Benson 1999), using the following parameters: match = 2, mismatch = 3, indels = 5, $P_M = 0.75$, $P_I = 0.20$, minimum period = 30, maximum period = 500. Output was then submitted to TRAP (Sobreira et al. 2006). Dotter (Sonhammer and Durbin 1995) allowed graphical confirmation of potential tandem repeats. The remaining unidentified contigs were submitted to TEclass (Abrusán et al. 2009), a tool that determines the likely mode of transposition and thus aids in identification of repeat type.

Potentially novel elements (contigs not identified via CENSOR, RepeatMasker, or Tandem Repeats Finder) were queried against the appropriate taxon sequence data using BlastN, which allowed us to generate a more accurate full-length consensus. If possible, the *M. lucifugus* 2× WGS was used to infer consensus sequences for TEs with low coverage in the 454 data. The top 40 hits were extracted with 200-bp flanking sequence (if available) using process_hits.pl (Smith and Ray 2011), a computational tool for TE mining which, in this case, was configured to combine hits with 50-bp overlaps and align them using MUSCLE. If the boundaries of the repeat element were not recovered, as evidenced by dissimilar sequence data at the 3' and 5' ends, then the outermost 150 bp of the consensus was used to query the data again and extend the alignment until the full-length TE could be assembled. Large contigs (>1000 bp) were submitted to open reading frame (ORF) Finder (<http://www.ncbi.nlm.nih.gov/projects/gorf/>) to identify potential reverse transcriptase, endonuclease, or transposase ORFs. Element names end with a two-letter taxon identifier to indicate the source of the consensus (i.e., *Mariner2_MI* was inferred from *M. lucifugus*).

Age Analysis

Novel TEs were further analyzed to determine their approximate period of activity as described in Pagan et al. (2010). Consensus sequences were repeatmasked against the respective taxon from which each was inferred, either the appropriate filtered 454 data set or a quarter of the *M. lucifugus* 2× WGS. To ensure full-length hits could be acquired, the query sequences were trimmed to 300 bp; if possible, the fragment was selected from coding regions in autonomous TEs. RepeatMasker.align output files were processed by a perl script designed to calculate the Kimura 2-Parameter distances while excluding hypermutable CpG sites (Pagan et al. 2010). Output was parsed to only include hits that spanned at least

Table 1

454 Sequencing Summary

	Total Reads	Mean Read Length (bp)	Total Base Pairs	Estimated Genome Size	Percentage of Genome Coverage	After Sequencing Artifact Filter		
						Unique Reads	Percentage of Replicates	Percentage of Unique Genome Coverage
<i>Artibeus lituratus</i>	295660	397	1.01×10^8	2.70×10^9	3.75	255065	13.73	3.23
<i>Corynorhinus rafinesquii</i>	403317	285	1.15×10^8	2.42×10^9	4.75	317269	21.34	3.74
<i>Lasiurus borealis</i>	233826	368	8.60×10^7	2.56×10^9	3.36	169361	27.57	2.43
<i>Myotis austroriparius</i>	86583	285	2.47×10^7	3.26×10^9	0.76	67924	21.55	0.59
<i>Nycticeius humeralis</i>	135978	280	3.81×10^7	2.42×10^9	1.57	108535	20.18	1.26
<i>Perimyotis subflavus</i>	122395	265	3.24×10^7	2.26×10^9	1.44	99801	18.46	1.17

NOTE.—Percentage of Genome Coverage was approximated using mean read length and estimated genome size. A sequencing artifact filter was applied to data (Percentage of Unique Genome Coverage) before graph-based repeat discovery and RepeatMasker analyses to determine genome representation.

90% of the query sequence. Ages were estimated from the distances using the mammalian neutral mutation rate, 2.2×10^{-9} (Kumar and Subramanian 2002). A complete library of the full consensus sequences was also used to query all six 454 data sets using BlastN to test for lineage specificity.

Genome Representation

TE contribution to genome content is often quantified by the number of bases they occupy in sequenced genomes. This value could be estimated using contig length and read depth, as per Macas et al. (2007) for the pea (*Pisum sativum*) genome. However, unlike *Pisum*, in which the dominant TE is an LTR element, the primary TE components of mammalian genomes are LINEs which are often 5' truncated. This makes them difficult to reconstruct in their entirety from the limited coverage and short read lengths we obtained. This is also true for other large autonomous TEs, especially those with low copy numbers. SINE subfamilies are another major component of mammalian genomes. In these bats, the dominant SINE is Ves, with a consensus of just over 200 bp. Additionally, there are several short (<400 bp) nonautonomous DNA transposon families. Each of these observations suggests that using contig length and read depth might lead to inaccurate estimates of genome coverage. For example, our average read length was ~300 bp, longer than a typical full-length Ves. Thus, the assembled contig lengths would be longer than the actual elements and artificially inflate genome coverage calculations.

We therefore chose to focus on the proportion of total hits for each TE in the filtered data. We used a custom library of TE consensus sequences as identified above from each taxon to mask the respective filtered data set with RepeatMasker. Process_hits.pl was used to combine hits with 50 bp overlap, and then tally the number of and length of hits with a minimum length of 30 bp (the shortest 454 read lengths) in each taxon for each of five repeat categories (DNA, ERV/LTR, Non-LTR/LINE, Non-LTR/SINE, and satellite). Each read should represent random data from the genome. Thus, the proportion of the genome occupied by each TE

category and/or family was then extrapolated from the data.

Results

454 Sequencing

Approximately 3.97×10^8 bp of data were obtained. Genome sizes for all taxa were obtained from www.genomesize.com. C-values for *P. subflavus* and *M. lucifugus* were not available, but rather estimated from congeners. Genome coverage was calculated from the number of sequenced base pairs divided by the estimated genome size. Genome coverage ranged from ~0.76% for *M. austroriparius* to ~4.75% for *C. rafinesquii*. Read lengths ranged from 29 to 755 bp and averaged ~300 bp. The 454 replicate filter reduced the data by around 20%. For example, coverage was decreased to 0.59% for *M. austroriparius*. However, this level of coverage still allowed for identification of repeats present in >1,000 copies in the genome (Macas et al. 2007). For example, a 1,000 copy repeat in *M. austroriparius* will be found ~5.9 times in the data set, calculated as follows for 1.94×10^7 bp filtered data and 3.26×10^9 bp genome size: $1,000/[1/(1.94 \times 10^7/3.26 \times 10^9)]$. Information on the filtered and unfiltered reads is summarized in table 1. The raw data are available from the Dryad Repository: <http://dx.doi.org/10.5061/dryad.83164r7v>.

Repeat Discovery and Distribution

Myotis lucifugus is the best characterized bat with regard to TE content (Pritham and Feschotte 2007; Ray et al. 2007, 2008). Although we were unable to obtain a *M. lucifugus* sample for this sequencing survey, the inclusion of the congener, *M. austroriparius*, allows validation of our methods. The estimated 9.9 Myr divergence time (Stadellmann et al. 2007) between the two species suggests we should find similar TE landscapes. Indeed, we identified all major *M. lucifugus* TE families in *M. austroriparius*. Most contigs that were not initially classified using Censor or RepeatMasker were identified as either tandem repeats or mitochondrial DNA. Less than 0.5% of the *M. austroriparius* repeat content was labeled

Table 2Comparison of RepeatMasker Output from *Myotis austroriparius* 454 Data and the WGS for *M. lucifugus*

Element Class/Family	<i>M. austroriparius</i>		<i>M. lucifugus</i>	
	Percentage of RM Hits	Percentage of 454 Sequence Data	Percentage of RM Hits	Percentage of WGS
DNA/hAT	10.75	2.07	12.95	2.29
DNA/Helitron	15.13	2.78	16.23	2.57
DNA/Mariner	3.19	0.68	3.32	0.67
DNA/piggyBac	1.14	0.27	0.65	0.16
DNA/TcMar-Tigger	0.09	0.02	0.27	0.05
ERV/LTR	10.49	2.35	9.17	2.22
Non-LTR/LINE	29.10	9.21	17.49	6.02
Non-LTR/SINE	30.02	5.31	39.86	6.27
Non-LTR/unknown	0.10	0.03	0.04	0.02

NOTE.—Percentage of RM hits = proportion of total RepeatMasker hits to any given TE type. Percentage of 454 sequence data indicates proportion of bases masked from *M. austroriparius* survey sequence data. Percentage of WGS indicates proportion of bases masked in the *M. lucifugus* WGS.

“unknown.” Most contigs could be classified as satellite, DNA, LINE, SINE, or LTR elements and were found either to be previously identified or, if not already characterized, were shown to be repetitive in *M. lucifugus*. Indeed, our estimates of genome coverage for multiple element classes using the WGS of *M. lucifugus* and the collected 454 reads for *M. austroriparius* are a close match (table 2). The only appreciable deviation between the two is for the non-LTR/LINES. Harismendy et al. (2009) performed a comparison of next generation sequencing platforms and found overall that Roche 454 data had fairly even treatment of unique versus repetitive sequences, but did note a 1.25 overrepresentation of LINES. It is possible that we are observing this bias here, but it would be expected to occur equally across all taxa, and no apparent bias is observed for the Class II families. In combination with similar analyses on pea (Macas et al. 2007) and snake genomes (Castoe et al. 2011), these data suggest that our approach is appropriate for estimating the TE landscape despite limited genome coverage.

As described in Novak et al. (2010), graph conformation of a specific cluster revealed features of the respective repeat family. Reads, presented as vertices, are connected by edges to other reads, which they overlap. A summary of the five largest clusters for each taxon can be found in table 3.

We were able to confirm some previous PCR analyses that probed for Class II TEs in other vespertilionid taxa and identified several (*piggyBac1*, *hAT2*, *hAT3*) that initially appear to be limited to *Myotis* (Ray et al. 2008). However, two TE families previously thought to be confined to *Myotis* were identifiable in other taxa: *hAT1_M1* was identified in *N. humeralis* and *piggyBac2_M1* was observed in data from *C. rafinesquii* and *L. borealis*. This was likely due to mispriming from the internal primers of the earlier analysis and highlights the advantage of survey sequencing for a more accurate inspection of repetitive DNA. If we assume that the vespertilionid phylogeny described by Lack and Van Den Bussche (2010) (see fig. 1) is accurate, the presence of *hAT1_M1* in *N. humeralis* but not in *C. rafinesquii*, *L. borealis*,

or *P. subflavus* may result from two independent invasions of *hAT1_M1* into the lineages leading to *N. humeralis* and *Myotis* spp. However, alternative phylogenetic hypotheses exist (Hoofer and Van Den Bussche 2003) and correct inference of independent invasions will depend on a reliable phylogeny of the group. Several novel elements were also identified and their key features are summarized in table 4. These novel elements have been submitted to RepBase.

Most Class II TEs were categorized according to terminal inverted repeat (TIR) length and target site duplications (TSDs) after extending and assembling the full repeat consensus (see Repeat Discovery in the Materials and Methods section). Blast hits to potential ORFs were also used for identification. *Tc1/mariners* have 25- to 29-bp TIRs and TA TSDs, while *hATs* typically have 16-bp TIRs and 8-bp TSDs with central TA dinucleotides. *Helitrons* are characterized by a 5' TC, 3' CTRR, an AT target site, and a 3' 18-bp palindrome; elements are identified according to >80% similarity at 3' (family) and 5' (subfamily) 30 bp (Yang and Bennetzen 2009). All *Helitrons* identified in this study were from the *Helibat* family (Pritham and Feschotte 2007), and several fell within a single unique subfamily according to the 5' 30 bp (*nHeliBat1_Lb/Nh/Ps/Cr*). The observation that we could not identify the probe sequences used by Thomas et al. (2011) in the consensus sequences of these elements suggests that they fall within a separate but similar lineage. Also of note is *Mariner1_M1*, which included the full *Mariner2_M1* within TIRs of extended length. Although initially identified in *P. subflavus*, the consensus sequences were inferred from *M. lucifugus* to obtain adequate coverage. Both elements contained an ORF and a nonautonomous variant was also recovered with 67-bp TIRs from the *P. subflavus* data set (*nMariner1_Ps*).

Unidentified clusters from most taxa were generally composed of low numbers of reads. The exception to this pattern was *P. subflavus*, in which 19 of 66 clusters (~6% of the repetitive content) could not be identified by CENSOR or through BlastX and BlastN searches against NCBI databases.

Table 3

Top Clusters for Each Taxon

	Cluster Number	Original Number of Reads	Number of Reads Used in Contigs	Number of Cluster-Based Contigs	Number of SeqMan Contigs	Number of RepeatMasker Reads	Element Name	Element Family
<i>Corynorhinus rafinesquii</i>	CL1	9595	8625	283	7	24730	L1MAB_ML	Non-LTR/LINE
						2347	ERV2X1A_I_ML	ERV/LTR
	CL2	3820	3526	61	1	4842	HAL1-1A_ML	Non-LTR/LINE
	CL3	2582	2538	3	1	2814		mtDNA
	CL4	2469	2249	37	1	3775	HAL1-1A_ML	Non-LTR/LINE
<i>Lasiurus borealis</i>	CL5	1755	1601	77	4	3343	HAL1-1A_ML	Non-LTR/LINE
	CL1	3324	2919	102	1	12262	L1MAB_ML	Non-LTR/LINE
	CL2	2174	1956	80	1	2973	HAL1-1A_ML	Non-LTR/LINE
	CL3	1076	847	92	4	4182	HAL1-1A_ML	Non-LTR/LINE
	CL4	625	531	44	3	1467	L1MAB_ML	Non-LTR/LINE
<i>Myotis austroriparius</i>	CL5	510	380	15	1	1663	HAL1-1A_ML	Non-LTR/LINE
	CL1	748	677	20	1	2197	L1MAB_ML	Non-LTR/LINE
	CL2	644	599	14	1	960	HAL1-1B_ML	Non-LTR/LINE
	CL3	563	330	16	3	2882	VES	Non-LTR/SINE
	CL4	303	226	6	1	423	Tandem Repeat	Satellite
<i>Nycticeius humeralis</i>	CL5	262	248	4	1	510	L1MAB2_ML	Non-LTR/LINE
	CL1	1818	1093	34	4	10436	VES	Non-LTR/SINE
	CL2	614	521	28	2	1397	HAL1-1A_ML	Non-LTR/LINE
	CL3	470	399	26	2	2101	L1MAB_ML	Non-LTR/LINE
						226	ERV2X1A_I_ML	ERV/LTR
<i>Perimyotis subflavus</i>	CL4	432	357	10	1	229	L1MAB_ML	Non-LTR/LINE
						512	ERV2X1A_I_ML	ERV/LTR
	CL5	345	260	37	2	3218	nHelitron1_Nh	DNA/Helitron
	CL1	2092	1634	65	3	2934	Tandem Repeat	Satellite
	CL2	1596	1430	54	1	5329	L1MAB_ML	Non-LTR/LINE
<i>Artibeus lituratus</i>	CL3	1408	1151	88	6	4994	nHelitron1_Ps	DNA/Helitron
	CL4	1282	1157	37	1	2002	HAL1-1A_ML	Non-LTR/LINE
	CL5	830	790	7	1	926	Tandem Repeat	Satellite
	CL1	5933	5225	154	4	24398	L1-4_PVa	Non-LTR/LINE
	CL2	5299	4563	169	5	11493	HAL1-3_ML	Non-LTR/LINE
<i>Artibeus lituratus</i>	CL3	2688	2498	20	1	3131	Tandem Repeat	Satellite
	CL4	2454	2385	7	1	2609	Tandem Repeat	Satellite
	CL5	1482	1269	41	3	2321	Tandem Repeat	Satellite

NOTE.—Information regarding the content of the graph-based clusters is provided, including the original number of contigs, which were submitted to SeqMan. The SeqMan contigs were then submitted to CENSOR for identification and used to RepeatMask the respective taxonomic 454 data set to determine genome representation.

Dotter and TRF analyses did not identify the contigs as tandem repeats, and visual inspection showed no indication of sequencing artifacts. Many of the unrecognized contigs were >500 bp, and the ends were not recovered. Cluster 50, for instance, contained an 864-bp contig, and attempts to identify the ends using Blast were unsuccessful. With no similarity to known TEs and lack of 5' and 3' ends, which often contain the defining features of the various repeat families, we were unable to discern if these might be novel TEs. Identifying these contigs is the subject of ongoing investigations.

A potentially confounding artifact in these types of analyses was also observed in *P. subflavus*. Many of the graph-based clusters contained only a few reads, yet RepeatMasker output indicated a large number of hits to the contig. Cluster 50 contained 18 reads of which only two

were used to generate the cluster-based contig, yet RepeatMasker identified 794 hits. In this case, the RepeatMasker data was inflated with hits primarily to a 6-bp tandem repeat embedded within the contig. This suggests that future analyses will require passing even identifiable TEs through Tandem Repeat Finder prior to genome coverage analyses.

Finally, repeat analysis of outgroup *A. lituratus* suggests that the elevated Class II TE content does not extend to Phyllostomidae. Less than 0.5% of the data set was identified as Class II and no novel or potentially recently active families were observed. Like most other mammals observed to date, Class I TEs comprise more than 25% of the genome. The major TE clades present in *A. lituratus* were Ves SINEs and L1 (3% and 15% of the filtered data set, respectively).

Table 4

Characteristics and Ages of Novel TEs

Element	Length (bp)	TIR (bp)	ORF (aa)	N ^a	Average K2P	Standard Error	Average Age (Myr) ^b
Mariner2_Ml	803	28	235	349	0.0188	0.0005	8.5
nhAT1_Nh	192	16		404	0.0194	0.0006	8.8
Mariner1_Lb	2294	25	347	23	0.0197	0.0024	9.0
nhAT4_Nh	203	16		127	0.0223	0.0018	10.1
nhAT2_Nh	246	16		61	0.0228	0.0012	10.4
nMariner2_Lb	231	25		518	0.0268	0.0006	12.2
nHeliBat1_Ps	1207			33	0.0416	0.0041	18.9
nhAT3_Nh	213	16		47	0.0509	0.0066	23.2
nMariner1_Lb	184	29		54	0.0639	0.0032	29.1
nHeliBat1_Lb	993			209	0.0905	0.0019	41.1
nHeliBat1_Nh	1183			34	0.0916	0.0055	41.7
nHeliBat2_Ps	220			39	0.1119	0.0113	50.8
nHeliBat1_Cr	364			74	0.1208	0.0041	54.9
Mariner1_Ps	1293	32	345				
nMariner1_Ps	279	67					
Mariner1_Ml	1211	198	235				
nhAT5_Nh	337	16					

NOTE.—Elements shown in **bold** are lineage-specific. Names preceded by an “n” are nonautonomous. Age estimations are only shown if >20 hits of appropriate length were obtained for analysis. Final two letters denote data set from which consensus was inferred (e.g., Lb-*L. borealis*).

^a Number of RepeatMasker hits, which are at least 90% of the query length; see Materials and Methods.

^b Average mammalian neutral mutation rate (2.2×10^{-9}).

Age Analysis of Selected Elements

The three newly described Mariner elements from *L. borealis* as well as the five novel *nhATs* from *N. humeralis* appear to be lineage specific and have been active in the relatively recent past (table 4). The age estimate (average 8.5 Myr) for *Mariner2_Ml* suggests it would be specific to *Myotis*, yet it was identifiable via BlastN analysis in all vesper 454 data sets. These contrasting results were further investigated by determining the activity periods of *Mariner2_Ml* in each taxon. Due to limited copy numbers in the 454 data, all RepeatMasker hits were used (instead of only hits within 90% length of the query, as for table 4). Average age estimates were as follows: *M. austroriparius* 11 Myr ($N = 40$), *P. subflavus* 16 Myr ($N = 21$), *C. rafinesquii* 17 Myr ($N = 68$), *N. humeralis* 19 Myr ($N = 17$), and *L. borealis* 23 Myr ($N = 44$). These estimates suggest activity of *Mariner2_Ml* in each taxon following the split from *Myotis* 32 Ma (fig. 1). Several *Helitrons* appear to predate the divergence of the five vespertilionid taxa, with the oldest having been active roughly 51 (*nHeliBat2_Ps*) and 55 (*nHeliBat1_Cr*) Ma. BlastN analysis supported the presence of similar fragments (E value ≤ -65) of both TEs in all but the outgroup, *A. lituratus*, suggesting further evidence that at least the *Helitron* phase of the Class II invasion began in the common ancestor of vesper bats (Thomas et al. 2011).

Genome Representation

As would be expected for mammals, Class I elements dominated the TE landscape for all six taxa (fig. 2, table 5). The

highest LINE content (nearly 15% of the genome) was observed in the phyllostomid, *A. lituratus*. This was accompanied by the lowest SINE complement (3%). *Nycticeius humeralis* exhibited the reverse situation with decreased LINE content (7%) alongside elevated SINE levels (6%), revealing an inverse relationship between the full-length LINES and the nonautonomous SINEs ($r = -0.90809$, all six taxa). The contribution of LTRs across all taxa was low, roughly 1.0% or lower. Finally as with *M. lucifugus*, elevated Class II levels were observed for the five vespertilionids (ranging from 3% in *L. borealis* to 5% in *P. subflavus*), but not for the phyllostomid bat (<1% in *A. lituratus*). A broader examination of genome-wide TE relationships is depicted between Class I and Class II elements in figure 3 ($r = -0.84632$, $P = 0.03361$).

Discussion

We have modified a methodology originally applied to plant genomes to identify distinct TE landscapes within five vespertilionids and a single phyllostomid bat. Comparison of a congener of the well-characterized *M. lucifugus* suggests that the method provides an accurate estimate of the TE landscape. Of course, this assumes that no major changes in TE dynamics have occurred in either lineage since their divergence ~ 10 Ma (Stadelmann et al. 2007).

NonLTR retrotransposons were the most abundant TEs in all species, as is typical of mammals. This is generally attributable to L1 elements. A large contribution of satellite DNA was noted in the *P. subflavus* genome (6%), as well

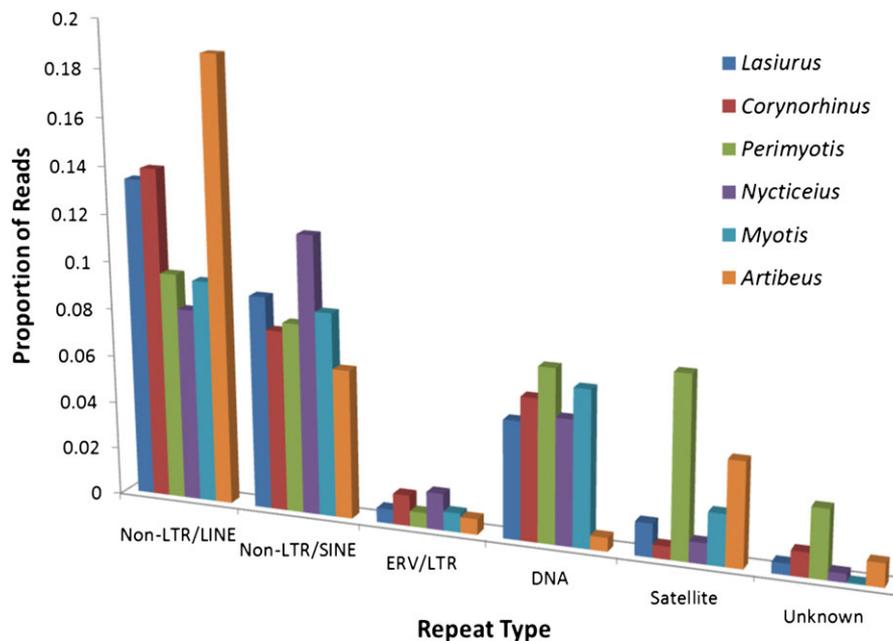


FIG. 2.—Genome representation of the TE classes. The inclusion of outgroup *Artibeus* suggests elevated DNA transposon activity is limited to the vesper taxa, while other aspects of their repetitive landscapes differ within the family.

as a considerable number of unidentifiable contigs across several repeat clusters. Unlike most mammals, Class II content was consistently elevated throughout Vespertilionidae, with ~3% or greater contribution to genome content in all five taxa. Class II elevation was not observed in the phyllostomid outgroup taxon, providing additional support to the hypothesis that vesper bats are unique within Chiroptera in their ability to tolerate and/or host DNA transposons (Thomas et al. 2011). At the very least, these data provide evidence that the surge of DNA transposon activity observed in Vespertilionidae arose following the divergence of Vespertilionidae and Phyllostomidae ~56 Ma (Datzmann et al. 2010).

As noted by Pritham and Feschotte (2007) and Thomas et al. (2011), the *Helitron* superfamily is a prevalent component of the vesper bat TE landscape. Our data demonstrate that *Helitrons* were active during the early diversification of Vespertilionidae. Analyses suggest that two *Helitrons* had peak activity over 50 Ma, which would indicate activity in

the common ancestor of Vespertilionidae and Phyllostomidae. However, these elements could not be recovered from the *A. lituratus* data. It should be noted that any elements with very low copy numbers (<1000) could be missed by our analyses. However, Thomas et al. (2011) failed to identify *Helitron* elements in Miniopteridae, suggesting that *Helitron* activity is indeed limited to Vespertilionidae. Assuming their hybridization and PCR-based results are accurate, this raises some issues regarding some of our activity period estimations. *nHeliBat1_Cr* and *nHeliBat2_Ps* were both estimated to have been active >50 Ma. Yet, the miniopterid divergence from Vespertilionidae is estimated to have occurred ~43 Ma (49–38 Ma) (Miller-Butterworth et al. 2007), suggesting that these two families should be present in miniopterid genomes. The problem likely arises from attempting to apply an average mammalian mutation rate (2.2×10^{-9}) to a wide range of taxa. Lack and Van Den Bussche (2010) noted that substitution rates in vesper bats are highly variable and that non-*Myotis* vespertilionids have consistently higher

Table 5

Genome Representation Determined Using RepeatMasker and a Custom Repeat Library Compiled for Each Taxon

	Non-LTR/LINE (%)	Non-LTR/SINE (%)	ERV/LTR (%)	Total Class I (%)	Total Class II (%)
<i>Artibeus lituratus</i>	14.83	2.90	0.93	18.66	0.38
<i>Lasiurus borealis</i>	11.74	4.02	0.42	16.18	2.56
<i>Corynorhinus rafinesquii</i>	11.93	3.91	0.97	16.81	3.12
<i>Nycticeius humeralis</i>	7.16	6.04	1.02	14.22	3.11
<i>Myotis austroriparius</i>	8.46	4.48	0.53	13.48	3.52
<i>Perimyotis subflavus</i>	9.33	4.18	0.69	14.20	4.45

NOTE.—Primary Class I repeat types are shown, and final two columns depict Class I versus Class II content.

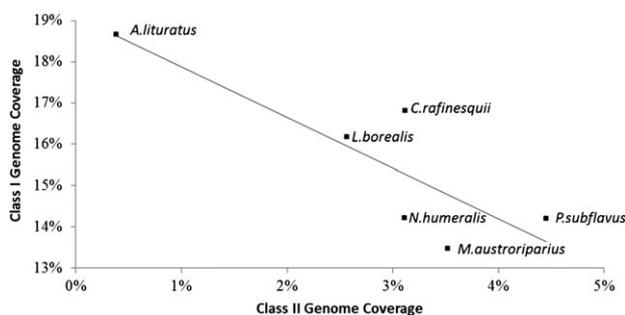


FIG. 3.—Correlation of Class I and Class II TE activity. Initial data suggest that TE activity may be inversely related between the two classes such that higher Class II genome representation is accompanied by a decrease in Class I content ($r = -0.85$, $P < 0.05$).

substitution rates. Thus, we might reasonably expect *Perimyotis* and *Corynorhinus* to exhibit inflated substitution rates. Calculating and applying lineage-specific rates to each taxon was beyond the scope of this study. However, future studies should incorporate such analyses. Future studies will also include samples from family Miniopteridae, which was recently elevated to the status of family and is more closely related to Vespertilionidae (Miller-Butterworth et al. 2007) and would therefore be appropriate for defining the limits of DNA transposon activity in these groups.

While *Helitrons* were active during the early stages of vesper bat diversification, other DNA transposon families have since invaded and been active in these genomes. For example, multiple *hAT*, *piggyBac*, and *Tc1/Mariner* elements, many of them novel to this study, exhibit activity profiles ranging from ~8 to 30 Ma (table 3) (Ray et al. 2008). One striking observation is from the *Mariner* family. Age analysis suggests that *Mariner2_M1* has been active the most recently, within the past 10 Myr in *M. lucifugus*. However, BlastN analyses of the available data indicate that this element is present in all five vesper taxa, which might suggest instead that *Mariner2_M1* was an older element with activity prior to the divergence ~32 Ma. Class II TEs generally have a short period of activity in a genome before accumulating inactivating mutations (Brookfield 2005). Likewise, although Class I TEs persist over longer timespans, they accumulate mutations and diverge into different subfamilies (Cordaux et al. 2004). A possible explanation for *Mariner2_M1* might be repeated reinvasion of vespertilionid genomes. However, at this time, we can only speculate.

Thus far, no evidence for elevated or recent Class II TE activity in bats has been found outside of the vesper lineage (Ray et al. 2008; Thomas et al. 2011). RNAi has been shown to specifically target TIRs to prevent transposon integration (Sijen and Plasterk 2003), but these defenses can be evaded when distinct subfamilies are present in low copy numbers (Plasterk 2002). The Class II TE expansion in *M. lucifugus* has been diverse, from *Helitron* and *Tc1/Mariner* superfamilies to various subfamilies of *hATs* and *piggyBacs* (Ray et al. 2008).

Similar findings of TE diversity for the taxa described here suggest that vesper bats in general are predisposed to accommodate invasion by novel TEs. While the following suggestion is open to further study, the capacity of vespertilionid bats to harbor active DNA transposons may be linked to another feature of *M. lucifugus*. A BlastN query of the newly released 7× *M. lucifugus* WGS using multiple mammalian Piwi homologs (list available upon request) and a search of the *Myotis* Ensembl database suggests that only two Piwi homologs are present, PIWIL2 (ENSM-LUG00000002115) and PIWIL4 (ENSM-LUG00000002018). This lies in stark contrast to the presence of all four homologs in the WGS of the megabat, *Pteropus vampyrus* (ENSPVAG00000010030, ENSPVAG00000009878, ENSPVAG00000016875, and ENSPVAG00000007245). Mammalian genomes are protected from TE integration in the germline by piRNA mediated methylation (O'Donnell and Boeke 2007; Aravin et al. 2008; Obbard et al. 2009), and loss of a single Piwi homolog has been linked to upregulated transposition (Carmell et al. 2007). Additional work to determine if the PIWI homologs missing in *M. lucifugus* are also missing in other affected bats would be an avenue worth pursuing. Loss of Piwi RNA genes may provide hypotheses to explain how TEs have managed to thrive in vesper bats. However, it raises an interesting question. Are vesper bats more susceptible to invasion or are they exposed to more potential invaders? It may be that Vespertilionidae is particularly susceptible to invasion by DNA transposons via their role as a host for a diverse array of parasites (Marinkelle and Grose 1972; Calisher et al. 2006; Wibbelt et al. 2010). Further research to identify patterns among bats may help answer these questions.

Several lineage-specific activity patterns were observed, suggesting differential activity in each lineage for particular transposon families and potential horizontal transfer events. As described above, at least two cases of potential horizontal transfer can be identified from this data. However, identifying horizontal transfer is dependent on overlaying the taxonomic distributions of TEs onto a well-established phylogeny. Vespertilionid phylogeny, unfortunately, has been rather intractable to both morphological and molecular data and is a well-known problem within the phylogenetic community (Stadelmann et al. 2007; Lack and Van Den Bussche 2010). Thus, while we suspect based on the most recent phylogenetic hypothesis presented by Lack and Van Den Bussche that both *hAT1_M1* and *piggyBac2_M1* have been transferred laterally into multiple vespertilionid genomes, we must be vigilant and work to generate a more robust phylogeny before making strong statements. That being said, both *hAT* elements in general and *piggyBac2_M1* in particular have been implicated in multiple horizontal transfers (Pace et al. 2008; Gilbert et al. 2010; Pagan et al. 2010).

Our initial interest in the vesper lineage was spurred by the elevated Class II activity in genus *Myotis*. However, the methods we describe allow for characterization of all TEs with relatively high copy numbers in a genome. Therefore,

we could also note differences in Class I content. For example, the *A. lituratus* genome exhibited the lowest SINE content (3%) and the highest LINE contribution (15%). A much larger SINE-to-LINE ratio was observed in *N. humeralis*, which may suggest an adaptation in recent Ves subfamilies to more efficiently utilize the LINE enzymatic machinery in this taxon. Such a scenario is the opposite of that seen in the recent analysis of the orangutan genome, in which the primate SINE, *Alu*, has apparently lost its ability to efficiently mobilize (Locke et al. 2011). The autonomous/nonautonomous relationship suggests a possible interaction between LINEs and SINEs as they compete with one another for use of needed enzymatic machinery (Brookfield 2005; Le Rouzic and Capy 2006).

Our data indicate that the rise in Class II TE activity may have been accompanied by a decreased Class I TE genome contribution. Vesper bat genera *Perimyotis* and *Myotis* displayed the highest Class II content (5% and 4%) and the lowest Class I content (14% each). This trend is amplified when the phyllostomid bat is included (fig. 3, $r = -0.85$), in which Class II content is at the low end of the spectrum while Class I content is the highest of the six taxa. However, while these results are suggestive of a trend, they still represent only six data points and should be taken with caution.

Our investigation is the first step in isolating any potential links between elevated Class II TE activity and the evolution of vesper bats. Variation in TE landscapes may be partially derived from population subdivision and genetic drift (Jurka et al. 2011). While the primate lineage has been examined extensively to elucidate the potential role of TEs in diversification, the focus was largely constrained to ancestrally derived Class I elements and remnants of extinct Class II TEs (Kim et al. 2004; Oliver and Greene 2009, 2011), although there are a few cases of recent Class II invasion (Gilbert et al. 2010). However, continued activity of both TE classes combined with horizontal transfer and novel TE invasions have furnished the vespertilionid family with a variety of elements with potential for facilitating species-specific adaptations. Finally, we note that the methods described here are conceptually similar to those described in a recent analysis of two snakes (Castoe et al. 2011) and multiple amphibian genomes (Sun et al. 2012). The major differences are with the precise computational methods used and not with the type of data analyzed. This suggests a strong interest in the evolutionary biology community in investigating the dynamics of TEs in large samples of relatively closely related organisms. Comparisons of mammalian TE landscapes have, until now, typically encompassed relatively diverse taxa. Inferences drawn from a limited sampling of genomes consisting mostly of model organisms are often broadly applied across taxa. This strategy is imposed primarily by the substantial costs of whole genome sequencing. However, the advent of next generation sequencing techniques has provided a leap forward in terms of gaining genome-level data (if not entire genome assemblies) for nonmodel organisms. Here,

we have demonstrated the utility of survey sequencing for generating sufficient data for comparative analyses and descriptions of novel TEs and have gathered data suggesting an extensive history of Class II TE activity throughout a broader sample of Vespertilionidae.

Acknowledgments

Research in the Ray laboratory is supported by the National Science Foundation (NSF) (MCB-0841821, DEB-1020865, and MCB-1052500) and the Mississippi Agricultural and Forestry Experiment Station. R.D.S. was funded by NSF (DEB-1020890) and the Louisiana Board of Regents (LEQSF-2006-09). J.M. and P.N. were supported by the Academy of Sciences of the Czech Republic (AVO-Z50510513). We acknowledge the LSU Museum of Natural Sciences for specimens used in this study, with special thanks to Donna Dittman, Robb T. Brumfield, and Mark S. Hafner.

Literature Cited

- Abrusán G, Grundmann N, DeMester L, Makalowski W. 2009. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 25(10):1329–1330.
- Aravin AA, et al. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell*. 31(6):785–799.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27(2):573–580.
- Brookfield JF. 2005. The ecology of the genome—mobile DNA elements and their hosts. *Nat Rev Genet.* 6(2):128–136.
- Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T. 2006. Bats: important reservoir hosts of emerging viruses. *Clin Microbiol Rev.* 19(3):531–545.
- Carmell MA, et al. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell.* 12(4):503–514.
- Castoe TA, et al. 2011. Discovery of highly divergent repeat landscapes in snake genomes using high throughput sequencing. *Genome Biol Evol.* 3:641–653.
- Cordaux R, Hedges DJ, Batzer MA. 2004. Retrotransposition of Alu elements: how many sources? *Trends Genet.* 20(10):464–467.
- Datzmann T, von Helversen O, Mayer F. 2010. Evolution of nectarivory in phyllostomid bats (Phyllostomidae Gray, 1825, Chiroptera: Mammalia). *BMC Evol Biol.* 10(1):165.
- de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 7(12):e1002384.
- Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev.* 13(6):651–658.
- Dong H, et al. 2011. Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System. *Acta Biochim Biophys Sin.* 43(6):496–500.
- Gibbs RA, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428(6982):493–521.
- Gilbert C, Schaack S, Pace li JK, Brindley PJ, Feschotte C. 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* 464(7293):1347–1350.

- Harismendy O, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10(3):R32.
- Hooper SR, Van Den Bussche RA. 2003. Molecular phylogenetics of the chiropteran family Vespertilionidae. *Acta Chiropt.* 5(Suppl):1–63.
- Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure, and the origin of species. *Biol Direct.* 6(44):1–16.
- Kidwell MG, Lisch D. 1997. Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A.* 94(15):7704–7711.
- Kim T-M, Hong S-J, Rhyu M-G. 2004. Periodic explosive expansion of human retroelements associated with the evolution of the hominoid primate. *J Korean Med Sci.* 19(2):177–185.
- Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A.* 99(2):803–808.
- Lack JB, Van Den Bussche RA. 2010. Identifying the confounding factors in resolving phylogenetic relationships in Vespertilionidae. *J Mammal.* 91(6):1435–1448.
- Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
- Le Rouzic A, Capy P. 2006. Population genetics models of competition between transposable element subfamilies. *Genetics* 174(2):785–793.
- Lindblad-Toh K, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069):803–819.
- Locke DP, et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* 469(7331):529–533.
- Macas J, Neumann P, Navratilova A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics* 8:427.
- Marinkelle CJ, Grose ES. 1972. A review of bats as carriers of organisms which are capable of infecting man or domestic animals. *Mitt Inst Colombo-Aleman Invest Cient.* 6:31–51.
- McCulloch ES, Stevens RD. 2011. Rapid development and screening of microsatellite loci for *Artibeus lituratus* and their utility for six related species within Phyllostomidae. *Mol Ecol Resour.* 11(5):903–913.
- Meganathan PR, Pagan HJT, McCulloch ES, Stevens RD, Ray DA. 2012. Complete mitochondrial genome sequences of three bats species and whole genome mitochondrial analyses reveal patterns of codon bias and lend support to a basal split in Chiroptera. *Gene* 492(1):121–129.
- Mikkelsen TS, et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* 447(7141):167–177.
- Miller-Butterworth CM, et al. 2007. A family matter: conclusive resolution of the taxonomic position of the long-fingered bats, *miniopterus*. *Mol Biol Evol.* 24(7):1553–1561.
- Novak P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* 11(1):378.
- O'Donnell KA, Boeke JD. 2007. Mighty Pivis defend the germline against genome intruders. *Cell* 129(1):37–44.
- Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. 2009. The evolution of RNAi as a defence against viruses and transposable elements. *Philos Trans R Soc B Biol Sci.* 364(1513):99–115.
- Oliver K, Greene W. 2011. Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. *Mob DNA.* 2(1):1–17.
- Oliver KR, Greene WK. 2009. Transposable elements: powerful facilitators of evolution. *Bioessays* 31(7):703–714.
- Pace JK, Gilbert C, Clark MS, Feschotte C. 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Evolution* 105(44):17023–17028.
- Pagan HJT, Smith JD, Hubley RM, Ray DA. 2010. PiggyBac-ing on a primate genome: novel elements, recent activity and horizontal transfer. *Genome Biol Evol.* 2(0):293–303.
- Perlea G, et al. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19(5):651–652.
- Plasterk RHA. 2002. RNA silencing: the genome's immune system. *Science* 296(5571):1263–1265.
- Pritham EJ, Feschotte C. 2007. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. *Proc Natl Acad Sci U S A.* 17(4):422–432.
- Ray DA, Pagan HJT, Thompson ML, Stevens RD. 2007. Bats with hATs: evidence for recent DNA transposon activity in genus *Myotis*. *Mol Biol Evol.* 24:632–639.
- Ray DA, et al. 2008. Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. *Genome Res.* 18:717–728.
- Rebollo R, Horard B, Hubert B, Vieira C. 2010. Jumping genes and epigenetics: towards new species. *Gene* 454(1–2):1–7.
- Sijen T, Plasterk RHA. 2003. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* 426(6964):310–314.
- Simmons NB. 2005. Order Chiroptera. In: Wilson DE, Reeder DM, editors. *Mammal species of the world: a taxonomic and geographic reference*. Baltimore (MD): The Johns Hopkins University Press. p. 312–529.
- Smith J, Ray D. 2011. Expedited batch processing and analysis of transposon insertions. *BMC Res Notes.* 4(1):482.
- Sobreira TJP, Durham AM, Gruber A. 2006. TRAP: automated classification, quantification and annotation of tandemly repeated sequences. *Bioinformatics* 22(3):361–362.
- Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167(1–2):GC1–GC10.
- Stadelmann B, Lin LK, Kunz TH, Ruedi M. 2007. Molecular phylogeny of New World *Myotis* (Chiroptera, Vespertilionidae) inferred from mitochondrial and nuclear DNA genes. *Mol Phylogenet Evol.* 43(1):32–48.
- Sun C, et al. 2012. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol Evol.* 4(2):168–183.
- Thomas J, Sorourian M, Ray D, Baker RJ, Pritham EJ. 2011. The limited distribution of Helitrons to vesper bats supports horizontal transfer. *Gene* 474(1–2):52–58.
- Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.
- Wibbelt G, Moore MS, Schountz T, Voigt CC. 2010. Emerging diseases in Chiroptera: why bats? *Biol Lett.* 6(4):438–440.
- Wilson DE, Reeder DM, editors. 2005. *Mammal species of the world: a taxonomic and geographic reference*. Baltimore (MD): The Johns Hopkins University.
- Yang L, Bennetzen JL. 2009. Structure-based discovery and description of plant and animal Helitrons. *Proc Natl Acad Sci U S A.* 106(31):12832–12837.
- Zeh DW, Zeh JA, Ishida Y. 2009. Transposable elements and an epigenetic basis for punctuated equilibria. *Bioessays* 31(7):715–726.

Associate editor: Richard Cordaux