Journal of Heredity 2010:101(4):437–447
doi:10.1093/jhered/esq016
Advance Access publication March 23, 2010

Endangered Species Hold Clues to Human Evolution

Craig B. Lowe, Gill Bejerano, Sofie R. Salama, and David Haussler

From the Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064 (Lowe, Salama, and Haussler); the Department of Developmental Biology, Stanford University, Stanford, CA 94305 (Bejerano); the Department of Computer Science, Stanford University, Stanford, CA 94305 (Bejerano); and the Howard Hughes Medical Institute, University of California, Santa Cruz, CA 95064 (Salama and Haussler).

Address correspondence to David Haussler at the address above, or e-mail: haussler@soe.ucsc.edu.

Abstract

We report that 18 conserved, and by extension functional, elements in the human genome are the result of retroposon insertions that are evolving under purifying selection in mammals. We show evidence that 1 of the 18 elements regulates the expression of *ASXL3* during development by encoding an alternatively spliced exon that causes nonsense-mediated decay of the transcript. The retroposon that gave rise to these functional elements was quickly inactivated in the mammalian ancestor, and all traces of it have been lost due to neutral decay. However, the tuatara has maintained a near-ancestral version of this retroposon in its extant genome, which allows us to connect the 18 human elements to the evolutionary events that created them. We propose that conservation efforts over more than 100 years may not have only prevented the tuatara from going extinct but could have preserved our ability to understand the evolutionary history of functional elements in the human genome. Through simulations, we argue that species with historically low population sizes are more likely to harbor ancient mobile elements for long periods of time and in near-ancestral states, making these species indispensable in understanding the evolutionary origin of functional elements in the human genome.

Key words: *conservation biology, endangered species, exaptation, retroposons, transposons, tuatara*

Improvements in DNA sequencing technology have allowed the field of genomics to expand beyond humans and model organisms to begin sampling the diversity of life. Each genome project adds not only to our knowledge of the species being sequenced but also to our understanding of the human genome by increasing our statistical power to infer ancestral states on the human lineage. This ancestral history elucidates how each functional element introduced on the human lineage was created and refined, a list of genetic changes that have defined our evolution as a species and made us uniquely human.

Researchers working toward understanding the mutations that created each functional element in the human genome have discovered that both the protein-coding exons of genes and their surrounding regulatory regions occasionally have their origins in mobile element insertions (Brosius 1999; Lev-Maor et al. 2003; Bejerano et al. 2006; Wang et al. 2007; Sasaki et al. 2008; reviewed in Volff 2006; Feschotte 2008).

Mobile elements, often referred to as transposons or repeats, are sections of DNA that can generate multiple copies of themselves in the genome (reviewed in Kazazian 2004; Cordaux and Batzer 2009). When this occurs in the germ line, the resulting extra copy may be passed on to progeny, and this mutation can eventually become fixed in the population. Most mobile elements in the human genome are long interspersed nucleotide elements (LINEs), short interspersed nucleotide elements (SINEs), or long terminal repeats (LTRs), which all use a process known as retrotransposition to propagate. A functional retroposon must have an internal promoter that recruits a polymerase to transcribe an RNA copy of itself. This RNA copy must then recruit a fusion protein of an endonuclease and a reverse transcriptase that nicks the DNA and reverse transcribes the RNA copy back into the genome at another location, resulting in an additional genomic copy of the mobile element. LTRs and LINEs are "autonomous," meaning that their RNA copies encode the reverse transcriptase/endonuclease protein that is needed for reinsertion. SINEs are "nonautonomous," meaning that they do not encode a reverse transcriptase/endonuclease protein and therefore are dependent on recruiting a reverse transcriptase/endonuclease protein that has been translated from an RNA copy of a LINE for reinsertion into the genome.

These mobile element insertions can number in the millions for some families. As they spread throughout the genome, some carry with them functional modules, such as

transcription factor–binding sites to expand the regulatory network of some transcription factors (Wang et al. 2007) or an exon that can be added to genes (Lev-Maor et al. 2003). This process where a section of a mobile element consensus that has been under selection for transposition is placed under selection by the host for a new function falls under the term "exaptation" (Gould and Vrba 1982). More than 10 000 human functional elements have been exapted from LINEs, SINEs, DNA transposons, and LTRs that are identifiable in the extant human genome (Lowe et al. 2007). Mikkelsen et al. (2007) have estimated that at least 16% of eutherian-specific functional regions have their origins in mobile element insertions.

Even though a mobile element family can have millions of copies in the genome, typically a very small percentage of these instances will have the ability to retrotranspose. When the reverse transcriptase is inserting a new instance into the genome, the enzyme often halts before the end of the RNA and often there are rearrangements in the inserted sequence (Cordaux and Batzer 2009). This leaves almost all new genomic instances inactive on arrival because they lack an internal promoter at their 5′ end. Thus, an autonomous family of mobile elements can be inactivated by mutations disabling the promoter or the protein-coding sequence of the few active copies that exist. The nonautonomous families can be inactivated by either disabling mutations in their promoters or inactivating the autonomous family they depend on. An example of the latter is the inactivation of the mammalian interspersed repeat SINE when the LINE2 element was inactivated in eutherians and marsupials (Warren et al. 2008).

Exaptation events older than the mammalian radiation are often difficult to detect using the extant human genome because most mobile elements that were active hundreds of millions of years ago have been inactivated somewhere along the lineage leading to humans. All their remaining insertions that were not exapted by the host have neutrally decayed to the point of being unrecognizable as mobile element insertions in the extant genome. However, when a species outside of mammals is sequenced that still harbors the ancient transposon in a near-ancestral state, the repeat consensus from the newly sequenced species can be used as an estimate of the consensus that was once active on the lineage leading to humans. This allows researchers to identify exaptation events on the human lineage associated with repeat families that have been inactive in humans for a long time but are recently active in another species. This methodology has been used in several recent studies that have taken advantage of genomic sequence becoming available from previously unsequenced clades (Bejerano et al. 2006; Nishihara et al. 2006; Mikkelsen et al. 2007).

If a species is the only remaining organism to harbor an ancestral version of a mobile element that was once active on the human lineage, this species may be the only means by which we can understand the evolutionary history of the human functional elements that were exapted from its insertions.

The tuatara, *Sphenodon punctatus* and *Sphenodon guntheri*, is a reptile that is currently restricted to small islands in northern New Zealand and has been recognized as threatened and protected since 1895. Yet, 10 of the 40 populations have become extinct during the last century (Daugherty et al. 1990), and it is possible that without conservation efforts during the past 115 years that the remaining populations would be extinct. The tuatara may continue to battle extinction because their temperature-dependent sex determination may be ill suited for a warming planet when coupled with their long generation time, low genetic diversity, and small island habitats that prevent southward migration (Mitchell et al. 2008).

Materials and Methods

Sequences Searched

When looking for sequence similarity to the endangered (EDGR)-LINE, we used LASTZ (Harris 2007) to search all sequence with taxonomic information from the National Center for Biotechnology Information trace archives and GenBank. When an assembled genome was available, we used the assembly in lieu of the trace and GenBank sequences. Prior to searching with LASTZ, we soft masked the target sequences for low complexity. We used a sensitive set of LASTZ parameters that include seeding on 6 identical nucleotides and using the HOXD55 matrix to score matches and mismatches in the alignments (Chiaromonte et al. 2002).

E Value Calculations for LASTZ Alignments

Because LASTZ does not compute significance values to accompany alignment scores, we calculated E values by generating 10 000 shuffles for each consensus sequence and searching these against the human genome. This simulation allowed us to assign an E value to a score by counting the number of alignments from the 10 000 shuffled sequences that scored as well or better than the score of interest and divide this number by 10 000 to arrive at the E value for the given score. With this method, we are able to assign E values down to 0.0001.

To ensure that our E value calculations are robust, we also implemented a second method that does not shuffle the consensus sequence but rather generates random sequences of the same length as the consensus from a first-order Markov model that was trained on the human genome. These random sequences were used in the same manner as the shuffling method to create E values for alignment scores. As with the shuffling methods, all tuatara alignments are significant and in all cases more significant than matches to any of the other consensus sequences (Supplementary Table S2 online).

To ensure that our results were not biased by the method used to reconstruct the consensus sequences (Supplementary Text online) or our parameter settings for LASTZ, we completed a third analysis by conservatively assuming that an individual a priori knows the location of the 18 exaptations in the human genome from the EDGR-LINE and that the EDGR-LINE may be found in the lizard, frog,

or coelacanth. We used WU-basic alignment search tool (Gish 2006) with default parameters, except a seed of 6 bases to aid in sensitivity, to separately align these human regions to the current assemblies for *Anolis carolinensis* and Xenopus tropicalis and all available sequence (65 Mb) for *Latimeria menadoensis*. Even with the statistical advantage of knowing where the mobile element insertion is, 3 of the human elements show no sign of homology in the other extant genomes; all exaptations show multiple regions of homology in the 32 Mb of sequence data available for Sphenodon punctatus (Supplementary Table S3 online).

Validation of Nonsense-Mediated Decay–Causing Alternatively Spliced Exon in *Asxl3*

We seeded undifferentiated KH2 mouse embryonic stem cells (Beard et al. 2006) at 15 000 cells per square centimeter in gelatin-coated 6-well dishes. We then followed the neural differentiation protocol from Ying and Smith for 5 days to differentiate the cells into neural precursors (Ying and Smith 2003). After 5 days, the cells were treated with 100 μg/ml of emetine for 4 h according to the methods of Ni et al. (2007). We isolated RNA from each sample using Trizol reagent (Invitrogen) according to the manufacturer's protocol. To compare the relative amounts of *Asxl3* transcripts, including or excluding the EDGR-LINE–derived exon, we converted 2.5 μg of RNA into complementary DNA using random hexamer primers and Superscript III (Invitrogen) according to the manufacturer's recommendations. We then performed 30 cycles of polymerase chain reaction with primers residing in the exons flanking the alternatively spliced exapted exon (actccccaatgacagcaaag and tccccgactcgagtgttagt) and then quantified the ratios of the splice forms with an Agilent Bioanalyzer using the DNA 1000 kit. As a control, we measured the effect of emetine on the well-characterized neural polypyrimidine tract binding protein cassette exon (Boutz et al. 2007) using primers (agctggtggcaatacagtcc and cccatcagccatctgtatca) described by Ni et al. (2007).

Results

We have discovered a previously unknown Chicken repeat 1 (CR1)-like LINE in the tuatara's genome that we termed the EDGR-LINE for endangered LINE. Due to inefficiencies in reverse transcription, the vast majority of LINE insertions are 5′ truncated (Deininger et al. 2003). Our consensus sequence for the EDGR-LINE represents the 2576 bases at the 3′ end of the LINE, including the 823 C-terminal amino acids of the second open reading frame (ORF2), which contains the endonuclease/reverse transcriptase protein (Supplementary Text online). The tuatara's genome is projected to be approximately 5 GB in size (Olmo 1981); yet, there is currently only 32 Mb of sequence data available (3 Mb in sequences longer than 1 kb) so the consensus may be expanded as more of the genome is sequenced. We searched the EDGR-LINE consensus from tuatara against all sequence data from GenBank, the trace archives, and assembled genomes with sensitive parameters and discovered additional novel LINEs in lizard (*A. carolinensis*), frog (*Xenopus tropicalis*), and coelacanth (*L. menadoensis*) that have sequence similarity to the EDGR-LINE in tuatara (see Materials and methods and Supplementary Text online). We translated the consensus sequences and conducted a phylogenetic analysis on their ORF2 proteins and all vertebrate ORF2 sequences from a previous study analyzing the diversity of CR1-like LINEs (Kapitonov and Jurka 2003). With a posterior probability of 0.99, our 4 novel LINEs coalesce to a common ancestral sequence outside of the set of all previously known proteins, suggesting that the EDGR-LINE is a novel lineage (Supplementary Text and Figure S1 online).

The most parsimonious history is for all extant EDGR-LINEs to have a common ancestral LINE in the sarcopterygian ancestor (lobe-finned fish and tetrapods), where it is inherited by descent, becoming inactivated prior to the last common ancestor of crocodiles, turtles, and birds and independently inactivated in the mammalian ancestor (Figure 1A).

Exaptations of the EDGR-LINE in the Mammalian Ancestor

When searching the EDGR-LINE consensus from *Sphenodon* against the human genome, 18 significant matches emerge (Table 1 and Supplementary Table S1 online). We calculate E values by shuffling the bases of the consensus 10 000 times and searching each shuffled sequence against the human genome. This allows us to assign score cutoffs to E values between 1 and 0.0001 (see Materials and methods). All the exaptations have E values below our significance threshold of 0.01. These exaptation events happened sometime after the separation of birds and reptiles from the human lineage because none were present in the amniote ancestor but all were present on the human lineage before the radiation of mammals. The exaptations have been evolving under strong purifying selection as evidenced by the level of constraint on their rate of evolution in a 28-way vertebrate alignment (Siepel et al. 2006; Miller et al. 2007). This level of evolutionary constraint would be needed to align a section of extant DNA to its ancestral form before the mammalian radiation.

The exaptations come from all different regions of the EDGR-LINE consensus (Supplementary Figure S2 online) and consist of an alternatively spliced exon in the *ASXL3* transcript and 17 putative gene regulatory regions that show strong purifying selection but do not appear in any known mature transcripts. The insertion that gave rise to an alternatively spliced cassette exon in *ASXL3* is included between the second and third exons in the 12-exon gene and contains stop codons in all 3 frames (Figure 2). The stop codons are much further upstream from the last exon–exon junction than the 50–55 nucleotides needed to cause nonsense-mediated decay (NMD) of the transcript (Maquat 2005), so the inclusion of this exon likely causes the transcript to be degraded and no functional protein to be produced. Lareau et al. (2007) and Ni et al. (2007) have both shown that the expression level of proteins can be finely controlled by the inclusion or exclusion of

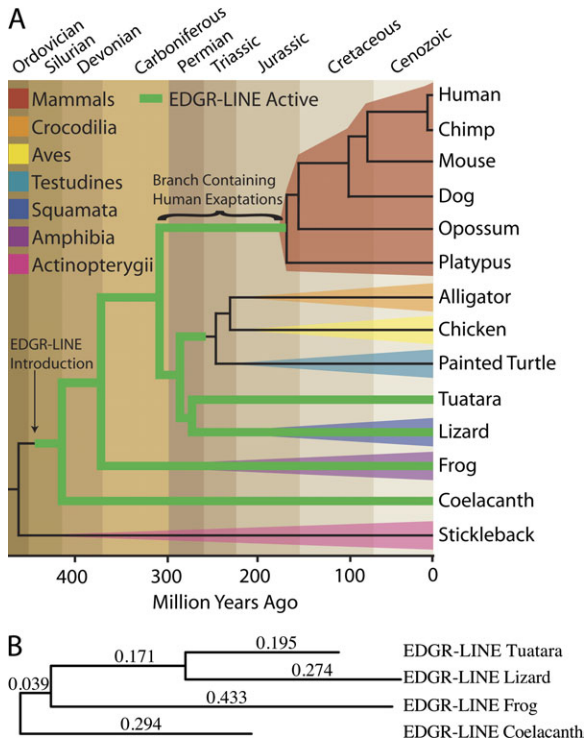

Figure 1. Evolution of the EDGR-LINE in vertebrates. (**A**) Parsimony suggests that the EDGR-LINE was introduced in the common ancestor of tetrapods and lobe-finned fish, and lineages where the LINE was active are shown with green. The LINE is not noticeable in mammals, crocodylia, aves, or testudines, so it has already been inactivated at least twice in evolution. The branch from the amniote ancestor to mammals is where we believe the human exaptations occurred because the exaptations are at least as old as the therian ancestor but show no sign of orthologous instances in nonmammalian amniotes. Subtrees are labeled for lineages that have extant species branching from the lineage to emphasize that the tuatara and coelacanth are thought to be on long branches with no other species. Divergence dates and topology are from previous studies (Graur and Li 2000; Shedlock and Edwards 2009). (**B**) The maximum likelihood branch lengths for the EDGR-LINE consensus sequences, when fit to the species tree, show the consensus to be evolving slower in the tuatara and the coelacanth than the green lizard or frog. Branches are scaled to and labeled with the number of substitutions per site. (This figure appears in color in the online version of *Journal of Heredity*.)

highly conserved NMD-causing exons. *ASXL3* is a homolog of the Additional sex combs (Asx) gene in Drosophila that is involved in histone methylation, which establishes regions of active and repressed chromatin during development (Baskind et al. 2009). From existing messenger RNA and expressed sequence tag data in humans, *ASXL3* appears to be active in the developing brain, adult brain, and testis (Kuhn et al. 2009).

In a limited survey of RNA samples from various stages of mouse development as well as human and mouse cell lines, we found expression of *ASXL3* transcripts, both containing and lacking the EDGR-LINE–derived exon, during embryonic development and in neural cell lines (data not shown). To verify that the exapted exon added a level of posttranscriptional regulation to the *ASXL3* gene during the development of therian mammals, we performed an experiment to demonstrate that the transcripts including this exon are targeted by NMD. We differentiated mouse embryonic stem cells into neural precursors and assayed for the relative amounts of *Asxl3* transcripts that include, and exclude, the exapted exon in both control experiments and when the cells were treated with emetine, which inhibits translation and therefore blocks NMD (see Materials and methods). We detected a 2-fold increase in the quantity of the NMD-causing transcript after treatment with emetine (Figure 3). These results are consistent with the exaptation of the EDGR-LINE having enabled our therian ancestor to more finely control the production of the *ASXL3* protein during brain development.

Importance of the Tuatara Genome

Seven of the human exaptations identified by the sphenodon consensus are also identified by one of the other EDGR-LINE consensus sequences; however, the alignment from *Sphenodon* is in all cases more significant and in 10 cases the sole means by which we currently understand the origins of these conserved, and by extension functional, elements in the human genome (Figure 4 and Supplementary Table S2 online). This observation agrees with the maximum likelihood branch lengths from the EDGR-LINE tree, which illustrates the consensus evolving 28% slower in tuatara than in the green lizard since their speciation event (Figure 1B and Supplementary Text online).

With the current amount of genomic sequence data available for the tree of life, it is not possible to prove that the tuatara is the only species harboring a near-ancestral version of the EDGR-LINE; however, the currently available data are consistent with this hypothesis. The well-sequenced mammalian tree lacks signs of the EDGR-LINE aside from the previously unapparent exaptations. Therefore, sauropsida (nonmammalian amniotes) is the next closest group to the exaptation events (Figure 1A). Crocodylia, aves, and testudines show no clear sign of the EDGR-LINE in their available sequence, implying that their ancestor inactivated the EDGR-LINE. Hence, squamata (lizards and snakes) is the most likely location for an extant genome to harbor a version of the EDGR-LINE that is as close to the mammalian ancestral version as the consensus from tuatara. More basal vertebrate groups (i.e., amphibians) are less likely to harbor an EDGR-LINE closer to the mammalian ancestral version than the tuatara because the divergence between the amniote and tetrapod ancestral versions appears to be substantial when compared with the divergence of the tuatara consensus since the amniote ancestor (Figure 1B). Currently, the only assembled genome in squamata is *A. carolinensis*, which contains an EDGR-LINE that

Table 1 Exaptations of the EDGR-LINE in the human genome

Name	Conservation *P* value	Putative function	Nearby genes	Historical appearance
edgrExap0	2.3×10^{-132}	As NMD exon	ASXL3	Theria
edgrExap1	2.6×10^{-41}	Gene regulation	FOXP2-PPP1R3A	Theria[a]
edgrExap2	3.9×10^{-145}	Gene regulation	IFIH1-FAP	Mammalia
edgrExap3	6.9×10^{-28}	Gene regulation	PITX2-C4orf32	Mammalia
edgrExap4	3.8×10^{-178}	Gene regulation	ZFPM2-OXR1	Mammalia
edgrExap5	2.6×10^{-46}	Gene regulation	PCDH9-PCDH20	Mammalia
edgrExap6	5.6×10^{-36}	Gene regulation	SLIT2-LCORL	Mammalia
edgrExap7	2.9×10^{-49}	Gene regulation	RANBP17-TLX3	Mammalia
edgrExap8	3.4×10^{-4}	Gene regulation	OFCC1-SLC35B3	Mammalia
edgrExap9	7.3×10^{-7}	Gene regulation	GRIN2B-C12orf36	Theria[a]
edgrExap10	6.9×10^{-31}	Gene regulation[b]	PDE7B-FAM54A	Mammalia
edgrExap11	2.8×10^{-28}	Gene regulation	RELN-ORC5L	Theria
edgrExap12	6.0×10^{-29}	Gene regulation	RIMS2-WDSOF1	Mammalia
edgrExap13	1.5×10^{-16}	Gene regulation	ODZ3-AGA	Mammalia
edgrExap14	9.0×10^{-28}	Gene regulation	LMO4-PKN2	Mammalia
edgrExap15	9.5×10^{-6}	Gene regulation	DAB1-C8B	Theria[a]
edgrExap16	4.2×10^{-6}	Gene regulation	PIK3R1-SLC30A5	Theria[a]
edgrExap17	4.5×10^{-26}	Gene regulation	MGMT-MKI67	Mammalia

The *P* values for conservation were calculated using PHYLOP (Siepel et al. 2006) on a 28-way vertebrate alignment (Miller et al. 2007). One exaptation functions as an alternatively spliced exon that regulates the *ASXL3* gene by targeting the transcript for NMD. For exaptations not appearing in mature transcripts, which we believe are under selection for a gene regulatory role, we list the 2 neighboring transcription start sites. All exaptations appear in the opossum genome (and therefore the therian ancestor), but some do not appear in the platypus genome (and therefore the mammalian ancestor). We believe that all the exaptations may well predate the mammalian ancestor because many of the elements that do not appear in platypus are in a region that overlaps a sequencing gap or are between contigs.

[a] Between platypus contigs or contained within a sequencing gap.

[b] This exaptation overlaps a noncoding RNA exon (BC038719) in human, but this region does not appear to be transcribed in any other species, so the conservation across mammals is likely due to a function at the DNA level.

evolved much faster than the homologous LINE in tuatara and therefore cannot identify many of the human exaptations (Figure 2). There are currently no matches in squamata to the tuatara LINE with higher scores than those found in *A. carolinensis*. This is despite 60 Mb of sequence, including more than 4.3 Mb of sequence from *Thamnophis sirtalis*, which with *Anolis* defines the most basal split in squamata (Albert et al. 2009).

Other Endangered Species Hold Clues to Human Evolution

The tuatara is not the only endangered species that may hold clues to humans' genetic past. In 2 earlier studies, the coelacanth was shown to harbor ancient SINEs in relatively unchanged forms, the living fossil (LF)-SINE (Bejerano et al. 2006) and the DeuSINE (Nishihara et al. 2006). In both cases, most or all other species with sequence data had either inactivated the repeat family or harbored a relatively diverged SINE family. This left the genome of the coelacanth as an important catalyst in understanding the history of human functional elements exapted from insertions of these SINEs.

Both the tuatara and coelacanth are more likely to be indispensable for particular evolutionary analyses because both occupy long branches with no extant species closer than 275 and 420 million years (My), respectively. In dense areas of the tree of life, a closely related species can often substitute for another when trying to understand an ancient state of the genome. If 1 member in a group of closely related species harbors an ancient retroposon, the others will also have at least identifiable remnants of the mobile element. This redundancy prevents any individual species in the group from being indispensable in understanding the consensus sequence of the mobile element shared among the group. However, to concentrate only on species with no close relatives may be too simplistic because extinction events do not act in a random and uniform manner over the tree of life; species closely related to endangered species are often endangered themselves (Purvis et al. 2000). Information about our genetic past may be precariously stored in areas of the tree of life where all species are threatened with extinction. An example of this may be the desert tortoise (*Gopherus agassizii*), which is listed as vulnerable by the International Union for Conservation of Nature (IUCN), as is the rest of its genus: the bolson tortoise (*G. flavomarginatus*), the gopher tortoise (*G. polyphemus*), which are listed as vulnerable, and the Texas tortoise (*G. berlandieri*), which is listed as threatened by the state of Texas (Campbell 2003; IUCN 2009). The only other extant species that branch from the desert tortoise lineage after the testudinidae ancestor are the impressed tortoise (*Manouria impressa*) and the Asian forest tortoise (*Manouria emys*), which are listed as vulnerable and endangered by the IUCN (Le et al. 2006; IUCN 2009).

Supporting the claim that an endangered vertebrate genus may be similarly indispensable in interpreting the human genome, we also found that the desert tortoise, like the tuatara, harbors an ancient CR1-like LINE that has not

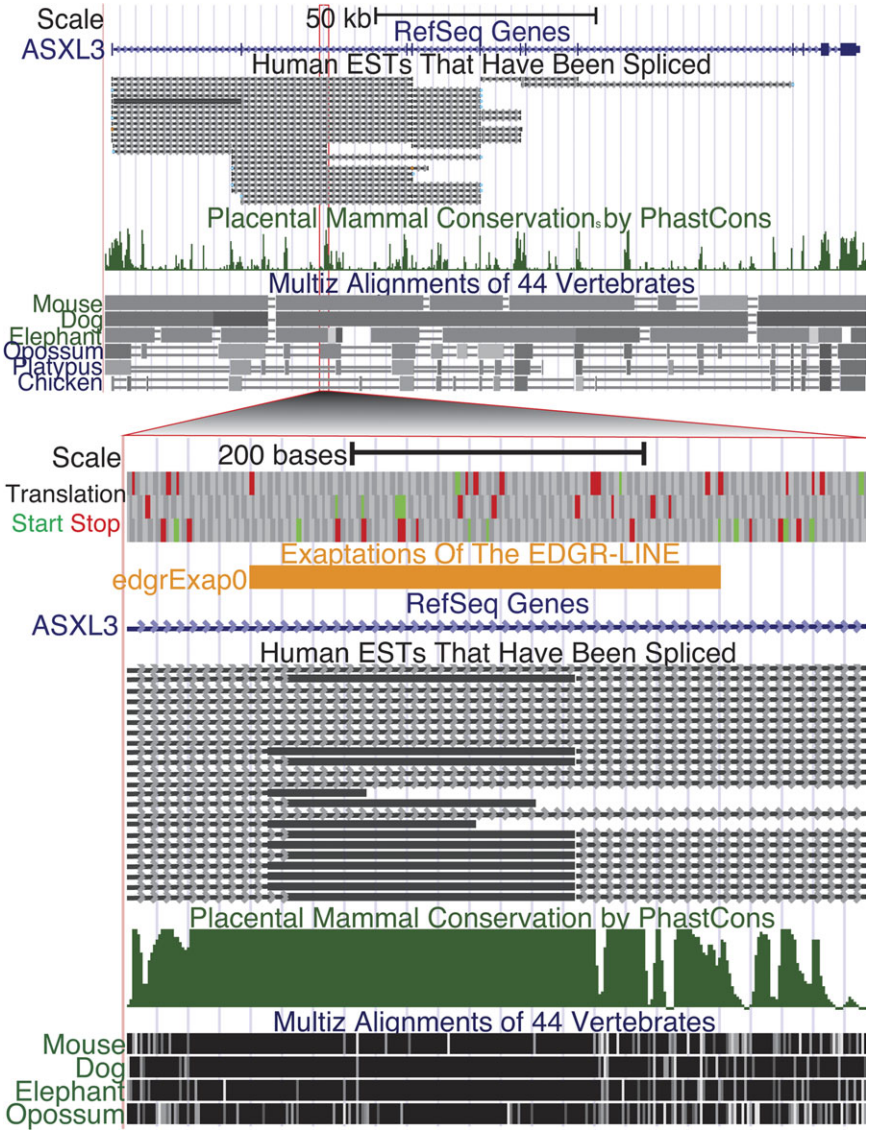

Figure 2. The top view shows the entire *ASXL3* gene, which contains an EDGR-LINE exapted to act as an alternatively spliced exon 2a between exons 2 and 3. The bottom view shows a detailed view of the immediate area of the genome around the exaptation, which is shown in orange. There are stop codons in all 3 frames of the exon, which causes NMD of the transcript. Therefore, this exaptation gave the host additional control over production of the *ASXL3* protein. (This figure appears in color in the online version of *Journal of Heredity*.)

previously been identified (Supplementary Text online). Based on analysis of the ORF2 protein, translated from the consensus, we believe that this LINE is part of a lineage encompassing the already discovered PsCR1 LINE (Kajikawa et al. 1997) from the black spine-neck swamp turtle (*Platemys spixii*) and a novel consensus from the American alligator (*Alligator mississippiensis*) (Supplementary Text online). The desert tortoise consensus enables 12 exaptations from this lineage of LINEs to be identified in the human genome that have not already been identified by either existing repeat annotation in the University of California, Santa Cruz genome browser (Kuhn et al. 2009; Smit et al. 2009) or the consensus sequences from alligator or the swamp turtle (Supplementary Tables S4 and S5 online). This suggests that the desert tortoise and its closely related species, which are all in danger of extinction, may hold information in their genomes that could aid in uncovering the molecular history of functional elements in the human genome.

Mobile Elements Are Active Longer in Small Populations

Previous researchers hypothesized a link between transposon accumulation and small population sizes (Jurka et al. 2007), as well as the more general case of a small effective population size (Lynch and Conery 2003). It has even been shown that species threatened with extinction tend to have

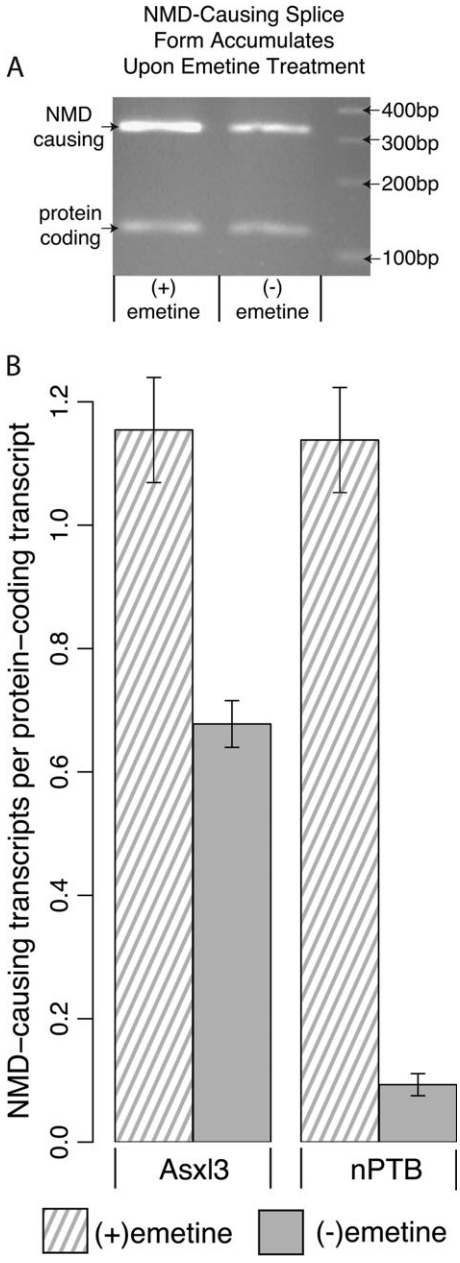

Figure 3. The exon exapted from the EDGR-LINE causes NMD of some transcripts from the *Asxl3* gene during neural differentiation. We incubated neural precursor cells in the presence or absence of the NMD inhibitor emetine (see Materials and methods). (**A**) We visualized the change in splice form abundance by running equal masses of reverse transcription polymerase chain reaction (RT-PCR) products from emetine-treated and control cells on an agarose gel and staining with ethidium bromide. (**B**) We used an Agilent Bioanalyzer to quantify the relative amount of RT-PCR products including or excluding the exon exapted from the EDGR-LINE (see Materials and methods). The bar graph shows the ratio of inclusive to exclusive transcripts for 16 independent wells from 2 independent differentiations (error bars depict 95% confidence intervals). In the 8 wells treated with emetine, we see a doubling of the *Asxl3* transcripts that include the exapted exon. When the results from the 8 control experiments are compared with results from emetine treatment using a 1-sided unpaired *t*-test, a *P* value of 1×10^{-6} validates the conclusion that the inclusive splice form is accumulating when NMD is inhibited. A transcript from neural polypyrimidline tract binding protein known to be targeted by NMD (Boutz et al. 2007) also increases in relative abundance in these experiments. These results are consistent with the exaptation adding an additional layer of control to the expression of the Asxl3 protein during brain development.

larger genomes than other evolutionarily close species, which is most likely due to an increased repeat content (Vinogradov 2004). It is likely that lineages where the effective population size has been historically low will not only have an increased repeat content but also these mobile elements will be enriched for ancient families that have been preserved in a relatively ancestral state. Such species give researchers a unique look into the genomic past.

In large populations that maintain a large effective population size, advantageous alleles fix and deleterious alleles are eliminated; when populations are small, or have a small effective population size, there is also a significant stochastic term determining if an allele becomes fixed. This stochastic term allows deleterious alleles to occasionally fix and advantageous alleles to be eliminated. As effective population size becomes smaller, the stochastic term is weighted more, making deleterious mutations more likely to fix and advantageous mutations less likely to fix (Kimura 1962). Because mobile element insertions tend to be mildly deleterious (Houle and Nuzhdin 2004), additional copies will be more likely to fix in small populations and beneficial mutations within the mobile elements that inactivate them will be less likely to fix (Figure 5A,B and Supplementary Text online). Both of these effects will increase the likelihood of ancient mobile elements remaining active in extant genomes of species with historically small populations. This is because the life span of a mobile element is determined by the rate at which it can fix new progenitor copies in the population and the rate at which the host can inactivate existing active (progenitor) copies.

In order to quantify the effect of population size on the life span of mobile elements, we have created a simulation that begins with diploid populations of 10, 30, 300, or 10 000 having 80 active repeat copies per haploid genome (Supplementary Text online). In 100 000 separate trials, the population was allowed to evolve at a constant size until the repeat family had no remaining active copies in the population. Measuring the mean number of active copies in a haploid genome at each generation (averaging over all members of the population and all trials) shows that as the population size decreases, active copies stay in the population for a longer time, allowing the mobile element to still be identified further in the future (Figure 5C). The trend of smaller populations to harbor

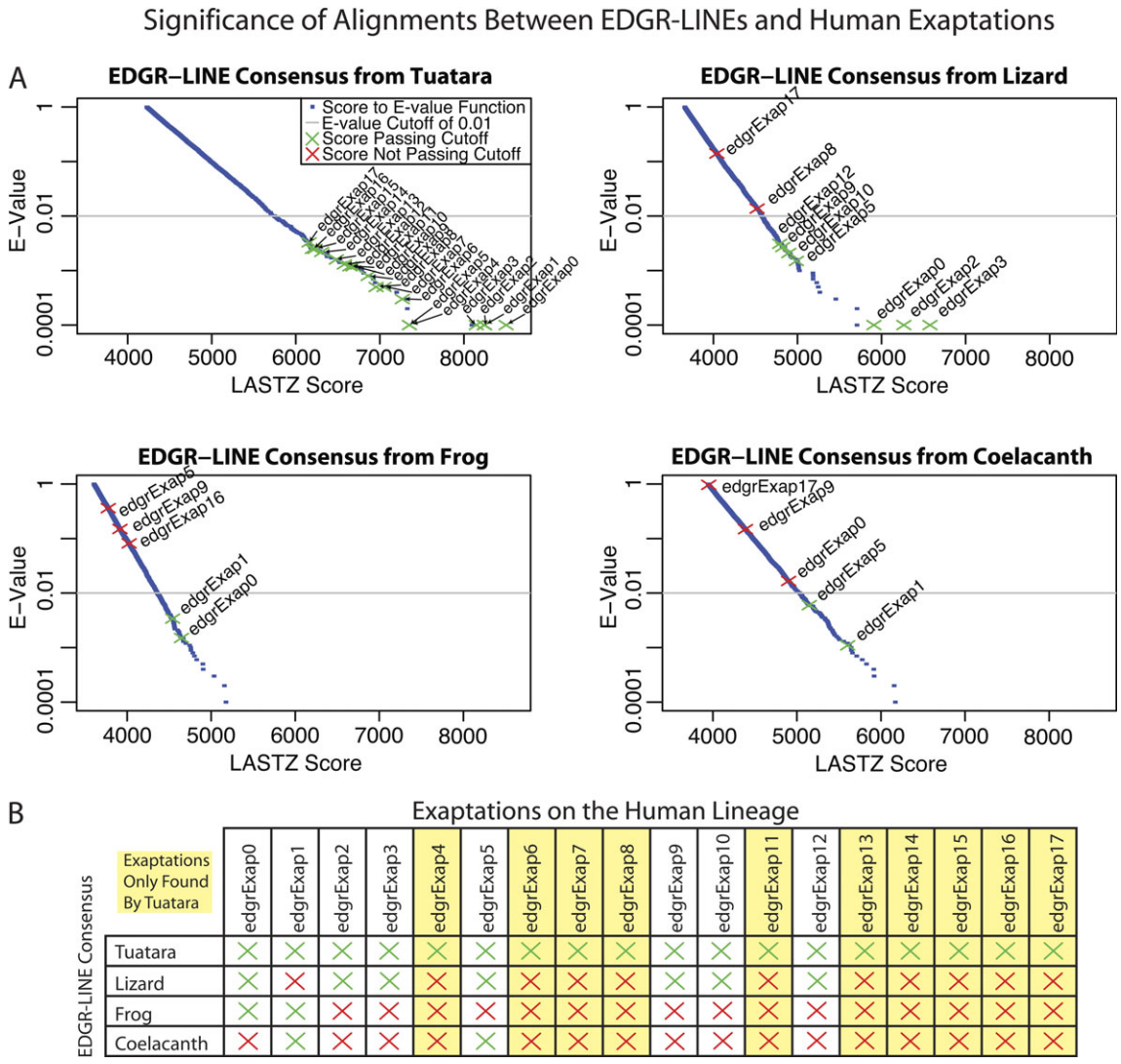

Figure 4. Significance of alignments between EDGR-LINE consensus sequences and human exaptations. (**A**) To quantify the significance of alignments between each consensus and the human genome, we computed *E* values by simulation. We shuffled each consensus 10 000 times and searched each shuffled sequence against the human genome. We now know how many biologically insignificant hits we expect for a given score, down to an *E* value of 0.0001. All 18 exaptations give LASTZ scores with *E* values less than our *E* value cutoff of 0.01 (green "X"), but the other consensus sequences find only a subset of the exaptations at this level of significance. Alignments with *E* values between 0.01 and 1 are marked with a red X, and alignments with *E* values greater than 1 are not shown. (**B**) Ten exaptations are only found with the consensus from the tuatara. (This figure appears in color in the online version of *Journal of Heredity*.)

active repeats in their genomes for a greater amount of time can also be seen by viewing the distribution of how many generations elapsed for each of the 100 000 trials to inactivate the repeat. Smaller populations inactivated the repeat family slower and with more variance than larger populations (Figure 5D).

Mobile Elements Evolve Slower in Small Populations

The EDGR-LINE consensus appears to be evolving at a slower rate in the tuatara than in the squamata (Figure 1B). In a previous study where a similar set of exaptation events were examined, the LF-SINE was also seen to have a surprisingly slow rate of evolution in the coelacanth (Bejerano et al. 2006). This species is also currently threatened with extinction and may have had a historically low population size (IUCN 2009). We believe that repeat consensus evolution is the result of an arms race between the mobile element and the host genome that escalates slowly when the effective population size of the host organism is small, as in most endangered species. In an asymmetric arms race between a host and parasite, the parasite may enjoy "periods of grace" after each new adaptation to avoid the host's defense mechanisms (Dawkins and Krebs 1979). This period of grace will continue until the host adapts a new

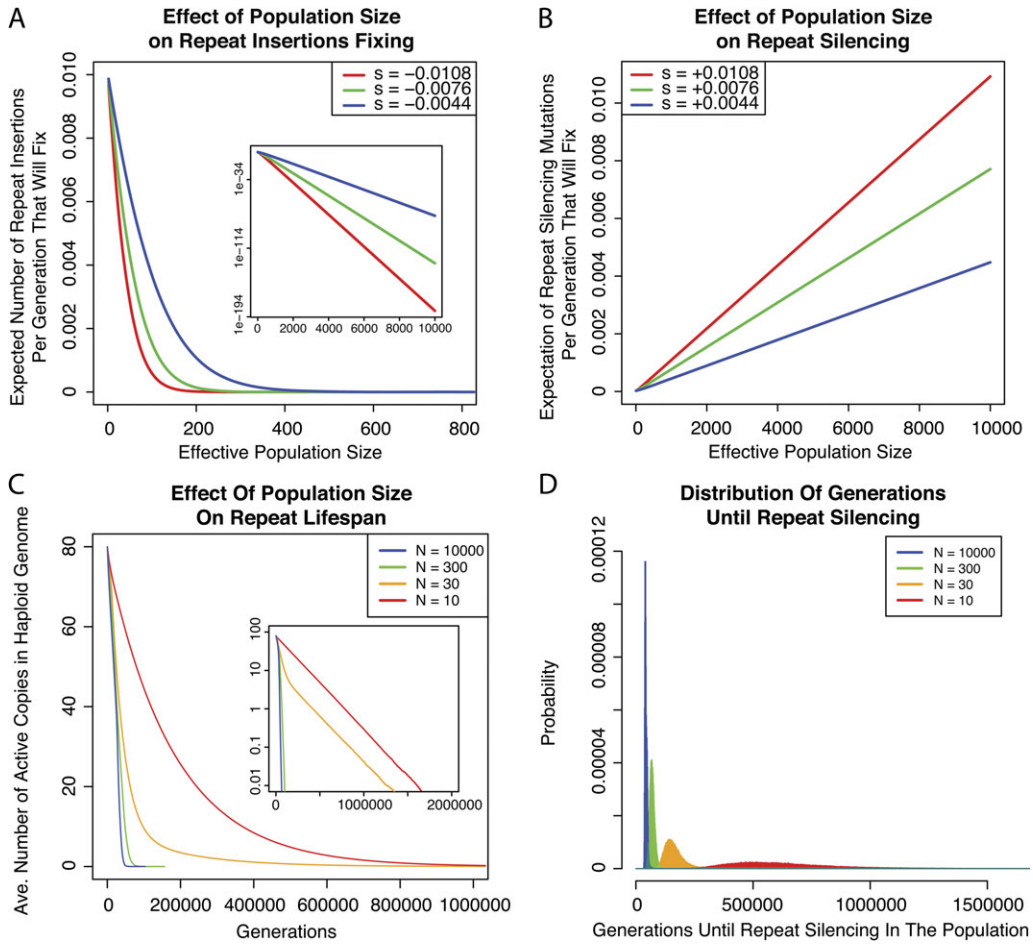

Figure 5. Effect of population size on repeat propagation and silencing. (**A**) We use 3 different values of the coefficient of selection for a mobile element insertion that represent a realistic range for a LINE insertion (Houle and Nuzhdin 2004). From the coefficient of selection and an estimate of an active LINE having a probability of 0.01 to retrotranspose each generation (Ostertag et al. 2002), we calculated the expected number of repeat insertions that will happen each generation and eventually become fixed in the population. As the population size increases, less repeat insertions will be expected to fix in the genome. The inset plot shows the same data, but on a log scale and up to a population size of 10 000. (**B**) Mutations that silence repeat instances will be slightly beneficial to the host. We calculate that more of these repeat silencing mutations that will eventually fix in the population are expected as the population size increases. Together (**A** and **B**) demonstrate that species with low effective population sizes are expected to accumulate more repeat insertions and less mechanisms to silence or remove them. (**C**) We simulated a genome evolving with a mobile element family for 4 different population sizes and plotted the average number of active copies (copies that can make new instances of the mobile element) present in a haploid genome. Source copies are present in the genomes of small populations for longer periods of time and in higher quantities. The inset figure shows the same data on a log scale and up to 2 000 000 generations. (**D**) In our simulation, a repeat family dies when it no longer has any active copies in any member of the population, meaning that it will not be able to make additional insertions. This graph shows the probability density of generations until complete silencing of the repeat family for 4 population sizes. As populations become smaller, the time until silencing becomes longer and more stochastic. (This figure appears in color in the online version of *Journal of Heredity*.)

defense mechanism, at which time the parasite will again be under strong selective pressure to change. Because species with small population sizes will not fix these new defensive mechanisms as often as a larger population size (Figure 5B), the retroposon may enjoy, on average, longer periods of grace and therefore will be placed under strong selective pressure to change fewer times, leading to a retroposon consensus that is slower evolving and more ancestral.

Discussion

The tuatara has been on the brink of extinction for most of the past 100 years, but in order to explain a mobile element being preserved for hundreds of millions of years in the genome, there must be an explanation that holds over a similar amount of time. One would be that the tuatara has had a historically low population size and been battling

extinction for at least tens of millions of years. What may be more probable is that the tuatara has had a small population size at times, such as during the oligocene flood that nearly submerged New Zealand 35–22 million years ago (Jones et al. 2009), but has had a historically low effective population size.

Two factors that can reduce the effective population size are social dominance limiting reproduction to a few successful males and an uneven sex ratio (Wright 1938). Tuatara are known to be highly territorial, which is thought to restrict mating opportunities to a small percentage of the males in the population. This effect is so severe that it reduces the effective population size by a factor of 2 in the 1 population examined (Moore et al. 2008). The tuatara's temperature-dependent sex determination, long generation time, and small island habitat that prevents migration (Mitchell et al. 2008) when coupled with the periodic temperature cycles of the earth, with a period of 40 000–100 000 years (Augustin et al. 2004), may often result in an unbalanced sex ratio, further reducing the effective population size of the population. These may be long-term attributes of the tuatara that have caused a historically low effective population size over the last 200 My, compared with other species.

It may be decades until we have sampled enough of the tree of life to be certain that the tuatara is the only extant genome that can elucidate the origins of these particular functional elements in the human genome. However, today, the tuatara's genome appears to be indispensable. When the tuatara was fully protected in 1895, people may not have realized that over 100 years later the scientific community would be using the species to understand the genetics of human evolution in ways that may otherwise have been impossible. We may be equally naive now in our ability to understand what information about our own genetic past we will be able to learn in 100 years from the tuatara, or other endangered species, that will not be recoverable without them.

Two new projects have been proposed to generate high-quality genomic sequence data from diverse vertebrate species. We are excited by a consortium of scientists who are currently proposing to sequence recently extinct species along with those that may become extinct in the near future because this project will capture diversity that is quickly vanishing from our planet (Nicholls 2009). The Genome 10K Community of Scientists, which plans to sequence 10 000 vertebrate genomes (Genome 10K Community of Scientists 2009), will provide a wealth of evolutionary information. It is a testament to continuing conservation efforts that an undertaking of this breadth and depth is possible. The approach described in this study will enable these new sequencing efforts to shed light on human evolution in addition to understanding vertebrate diversity. Conservationists have not only been protecting biodiversity but also our ability to understand genetic aspects of human evolution.

Supplementary Material

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

Funding

National Human Genome Research Institute (1U01HG004695-01 to D.H.); Human Frontier Science Program (RGY81/2008 to G.B.); Howard Hughes Medical Institute to D.H.

Acknowledgments

We wish to thank Manny Ares, Mark Diekhans, Dave Feldheim, Grant Pogson, Brian Raney, and Jeremy Sanford for insightful discussions. We also wish to thank Manny Ares, Jason Underwood, Bryan King, Sol Katzman, Courtney Onodera, Dave Greenberg, Bob Sellers, and Gayatri Pal for reagents and technical advice.

References

Albert EM, San Mauro D, García-París M, Rüber L, Zardoya R. 2009. Effect of taxon sampling on recovering the phylogeny of squamate reptiles based on complete mitochondrial genome and nuclear gene sequence data. Gene. 441:12–21.

Augustin L, Barbante C, Barnes PRF, Barnola JM, Bigler M, Castellano E, Cattani O, Chappellaz J, Dahl-Jensen D, Delmonte B, et al. 2004. Eight glacial cycles from an Antarctic ice core. Nature. 429:623–628.

Baskind HA, Na L, Ma Q, Patel MP, Geenen DL, Wang QT. 2009. Functional conservation of asxl2, a murine homolog for the Drosophila enhancer of trithorax and polycomb group gene asx. PLoS ONE. 4:e4750.

Beard C, Hochedlinger K, Plath K, Wutz A, Jaenisch R. 2006. Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. Genesis. 44:23–28.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 441:87–90.

Boutz PL, Stoilov P, Li Q, Lin CH, Chawla G, Ostrow K, Shiue L, Ares M, Black DL. 2007. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev. 21:1636–1652.

Brosius J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene. 238:115–134.

Campbell L. 2003. Endangered and threatened animals of Texas. Austin (TX): Texas Parks and Wildlife Department.

Chiaromonte F, Yap V, Miller W. 2001. Scoring pairwise genomic sequence alignments. In: Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein T, editors. Pacific Symposium on Biocomputing. *2002 Jan 3–7*. Singapore: World Scientific. p. 115–126.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 10:691–703.

Daugherty CH, Cree A, Hay JM, Thompson MB. 1990. Neglected taxonomy and continuing extinctions of tuatara (Sphenodon). Nature. 347:177–179.

Dawkins R, Krebs JR. 1979. Arms races between and within species. Proc R Soc Lond B Biol Sci. 205:489–511.

Deininger PL, Moran JV, Batzer MA, Kazazian HHJ. 2003. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev. 13:651–658.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 9:397–405.

Gish W. 2006 May. Wu-blast 2.0. [Internet]. [cited 2008 Sep 3] Available from: URL http://blast.wustl.edu/.

Genome 10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered. 100:659–674.

Gould SJ, Vrba ES. 1982. Exaptation—a missing term in the science of form. Paleobiology. 8:4–15.

Graur D, Li WH. 2000. Fundamentals of molecular evolution. Sunderland (MA): Sinauer.

Harris R. 2007. Improved pairwise alignment of genomic DNA. State College (PA): Pennsylvania State University.

Houle D, Nuzhdin SV. 2004. Mutation accumulation and the effect of copia insertions in Drosophila melanogaster. Genet Res. 83:7–18.

IUCN. 2009 Nov. IUCN red list of threatened species. Version 2009.1 [Internet]. [cited 2009 Jun 2] Available from: URL http://www.iucnredlist.org/.

Jones MEH, Tennyson AJD, Worthy JP, Evans SE, Worthy TH. 2009. A sphenodontine (Rhynchocephalia) from the Miocene of New Zealand and palaeobiogeography of the tuatara (Sphenodon). Proc R Soc Lond B Biol Sci. 276:1385–1390.

Jurka J, Kapitonov VV, Kohany O, Jurka MV. 2007. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 8:241–259.

Kajikawa M, Ohshima K, Okada N. 1997. Determination of the entire sequence of turtle CR1: the first open reading frame of the turtle CR1 element encodes a protein with a novel zinc finger motif. Mol Biol Evol. 14:1206–1217.

Kapitonov VV, Jurka J. 2003. The esterase and PHD domains in CR1-like non-LTR retrotransposons. Mol Biol Evol. 20:38–46.

Kazazian HH. 2004. Mobile elements: drivers of genome evolution. Science. 303:1626–1632.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics. 47:713–719.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. 2009. The UCSC Genome Browser Database: update 2009. Nucleic Acids Res. 37:D755–D761.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature. 446:926–929.

Le M, Raxworthy CJ, McCord WP, Mertz L. 2006. A molecular phylogeny of tortoises (Testudines: Testudinidae) based on mitochondrial and nuclear genes. Mol Phylogenet Evol. 40:517–531.

Lev-Maor G, Sorek R, Shomron N, Ast G. 2003. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. Science. 300:1288–1291.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci USA. 104:8005–8010.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science. 302:1401–1404.

Maquat LE. 2005. Nonsense-mediated mRNA decay in mammals. J Cell Sci. 118:1773–1776.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 447:167–177.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. 2007. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. 17:1797–1808.

Mitchell NJ, Kearney MR, Nelson NJ, Porter WP. 2008. Predicting the fate of a living fossil: how will global warming affect sex determination and hatching phenology in tuatara? Proc Biol Sci. 275:2185–2193.

Moore JA, Nelson NJ, Keall SN, Daugherty CH. 2008. Implications for social dominance and multiple paternity for the genetic diversity of a captive-bred reptile population (tuatara). Conserv Genet. 9:1243–1251.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. 21:708–718.

Nicholls H. Aug 2009. Time to sequence the red and the dead [Internet]. [cited 2009 Apr 15] Available from: URL http://www.nature.com/news/2009/090414/full/458812a.html.

Nishihara H, Smit AF, Okada N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. 16:864–874.

Olmo E. 1981. Evolution of genome size and DNA base composition in reptiles. Genetica. 57:39–50.

Ostertag EM, DeBerardinis RJ, Goodier JL, Zhang Y, Yang N, Gerton GL, Kazazian HH. 2002. A mouse model of human L1 retrotransposition. Nat Genet. 32:655–660.

Purvis A, Agapow PM, Gittleman JL, Mace GM. 2000. Nonrandom extinction and the loss of evolutionary history. Science. 288:328–330.

Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, Kimura-Yoshida C, Matsuo I, Sumiyama K, Saitou N, et al. 2008. Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci USA. 105:4220–4225.

Shedlock A, Edwards S. 2009. Amniotes (Amniota). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 375–379.

Siepel AC, Pollard KS, Haussler D. 2006. New methods for detecting lineage-specific selection. In: RECOMB. *2006 Apr 2–5*. Berlin: Springer. p. 190–205.

Smit A, Hubley R, Green P. 2009. Repeatmasker open-3.27 [Internet]. [cited 2009 Feb 1] Available from: URL http://www.repeatmasker.org/.

Vinogradov AE. 2004. Genome size and extinction risk in vertebrates. Proc Biol Sci. 271:1701–1705.

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays. 28:913–922.

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci USA. 104:18613–18618.

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 453:175–183.

Wright S. 1938. Size of population and breeding structure in relation to evolution. Science. 87:430–431.

Ying QL, Smith AG. 2003. Defined conditions for neural commitment and differentiation. Methods Enzymol. 365:327–341.

Received October 6, 2009; Revised January 19, 2010; Accepted January 27, 2010

Corresponding editor: William Murphy