An enhancer near *ISL1* and an ultraconserved *PCBP2* exon are derived from a novel retroposon

SUPPORTING MATERIAL

Gill Bejerano[1,*], Craig Lowe[1], Nadav Ahituv[2,3], Bryan King[1], Adam Siepel[1],
Sofie Salama[1,4], Edward M. Rubin[2,3], W. James Kent[1], David Haussler[1,4]

[1]Center for Biomolecular Science and Engineering, University of California Santa Cruz,
Santa Cruz, CA 95064, USA, [2]DOE Joint Genome Institute, Walnut Creek, CA 94598, USA,
[3]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA,
[4]Howard Hughes Medical Institute

24/02/06

Contents

Electronic Supplement

http://www.cse.ucsc.edu/~jill/sine.html

*Corresponding author: Tel: +1 831 459 5232 Fax: +1 831 459 1809 e-mail: jill@soe.ucsc.edu (G.B.)

S1 Coelacanth Retroposon

We used Blastn (Altschul *et al.*, 1997) to search the NCBI nucleotide database for sequences similar to the ultraconserved element uc.338 (Bejerano *et al.*, 2004b), which contains an alternatively spliced exon of PCBP2. This search returned significant hits to eight BACs from the Indonesian coelacanth, *Latimeria menadoensis*. We used Blastz (Schwartz *et al.*, 2003) to assemble five overlapping BACs. The resulting 1Mb of non-overlapping genomic sequence was compared to itself. This revealed that the repeat family, containing a sequence similar to that of uc.338, had 59 members, with pairwise identities between 40–87% (Table S1). A density of 60 copies/Mb is comparable with active SINE families found in other, fully sequenced vertebrate genomes (Kazazian, 2004), and would extrapolate to about 10^5 copies with current estimates of the coelacanth genome size at 2-3Gb (Danke *et al.*, 2004). The longest self-alignments overlapping our sequence of interest never exceeded 480bp, which suggested the length of the repeat may be close to this number. The proposed length was further supported by one family member being inserted into an open reading frame of a LINE. This insertion left a similar sized gap in the alignment when Blastx (Altschul *et al.*, 1997) was used to map a known reverse transcriptase onto the LINE.

We used MUSCLE (Edgar, 2004) to construct a multiple alignment of all occurrences of the repeat in *Latimeria menadoensis*. Columns with a high percentage of gaps were removed, as these columns most likely do not represent the position of an ancestral base. We ran PhyloFit (Siepel *et al.*, 2005) on the trimmed multiple alignment using a neighbor-joining guide tree obtained from MUSCLE to find the most likely ancestral sequence (Fig. S2).

We oriented the reconstructed sequence by sequence similarity to a vertebrate Serine tRNA at one end. This end conserves the sequence and spacing of the A and B boxes that act as a Pol III promoter (Gilbert and Labuda, 1999). The other end has a distinct poly-A tail (Fig. 1). This organization is typical of SINEs, as detailed in the main text.

We used a local alignment tool (Pearson *et al.*, 1997) to search the regions immediately flanking the 59 coelacanth instances for evidence of target site duplication (TSD). Indeed, the two instances closest to the coelacanth consensus exhibit clear candidates: AAACACAGCAAAGTCA, a perfectly conserved 16-mer on both flanks of the most similar instance, and AAAATTTAAAGAAA[C/T]ATG, a near perfectly conserved 18-mer on both flanks of the second instance. Putative decayed TSDs were detected flanking other instances. The two long matches exactly flanking the top instances suggest that these sequences underwent target primed reverse transcription, by hijacking the machinery of a LINE, similar to the relationship between Alus and L1s in the human genome (Ostertag and Kazazian, 2001). It also reaffirms our confidence that the characterized sequence is indeed a SINE, that we have it at its accurate length, and that this SINE has been recently active in coelacanth.

Using Blastz we determined that the reconstructed coelacanth SINE has no significant similarity to any known SINE, or other sequence found in Repbase (Jurka, 2000). We could find no similarity to the 3' end of any characterized LINE in Repbase or to any putative LINE-like repeat present in the 1Mb of sequence from *Latimeria menadoensis*. Mapping the coelacanth SINE to the human genome (below) we found minimal overlap with the 49% of the genome masked by RepeatMasker (Smit and Green, 2005) as known genomic repeats. We conclude that this is a novel repeat, and term it the LF-SINE, for lobe-finned fishes (or "living

fossil") SINE.

Once copies of the LF-SINE were detected in tetrapods, we tested the stability of our reconstruction by using all copies of the LF-SINE in the Indonesian coelacanth as well as outgroup sequences from frog, chicken, possum, mouse, rat, dog, chimp, and human. The addition of outgroups and the use of a maximum likelihood tree topology from PHYLIP (Felsenstein, 1995) for the coelacanth instances changed the identity of only 15 bases (Fig. S14), attesting to the remarkable stability of LF-SINE instances over vertebrate evolution. This suggests that our reconstruction is very close to the sequence of the actual progenitor coelacanth SINE. The maximum-likelihood tree of Fig. 1c was generated using SEMPHY (Friedman *et al.*, 2002) with default parameters for model and transitions/transversions, and an unconstrained topology.

The taxonomy of ancient SINEs is still largely unknown. It is still undecided whether ancient SINE families were each generated anew from different fortuitous tRNA retropositions (Ogiwara *et al.*, 2002), or if they all share some deep common ancestry (Gilbert and Labuda, 1999). Previous work has defined a 65bp "core" region of similarity, adjacent to the 5' tRNA derived region shared by many SINE families. As little as 50% sequence identity over 44/65 core bases was defined sufficient to deem all SINEs passing this threshold as descendants of a common proto CORE-SINE (Gilbert and Labuda, 1999). We note that the CORE-SINEs as defined (Gilbert and Labuda, 1999, 2000), the V-SINEs (Ogiwara *et al.*, 2002), and our LF-SINE, all share this weak signal (Fig. S13), apparently missing from the recently discovered DeuSINEs (Nishihara *et al.*, 2006). However, we contend that more work is required, before this weak signal, detected prior to the availability of any vertebrate genome sequence, can be taken as proof of common descent.

S2   Genomic Distribution

Genomic assemblies of 19 different organisms were searched for instances of the LF-SINE (listed in Table S2). The same assemblies are subsequently used throughout this work. Blastz was used to search the reconstructed sequence in Fig. S2 against the unmasked genome draft sequences. We used a stringent score threshold of 5,500, which, given a thousand different shuffles of the input sequence, retrieved no random hits to the unmasked human genome. Where the genomic draft is already at the chromosome level, we discarded all matches outside the golden path assembly of the chromosomes, to avoid contamination. See the electronic supplement at `http://www.cse.ucsc.edu/~jill/sine.html` for instance lists directly linked to the UCSC genome browser (Hinrichs *et al.*, 2006).

We also searched for the coelacanth sequence in the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces), using megablast (McGinnis and Madden, 2004), and in the EMBL nucleotide database (Kanz *et al.*, 2005) using WU-BLAST (Lopez *et al.*, 2003).

S3   Neutral evolution

To simulate neutral decay of LF-SINE instances we used a reversible rate matrix estimated from placental mammalian ancestral repeats (Siepel *et al.*, 2005), and microindels of up to 10 bases at a rate of 5% of the point substitution rate (Cooper *et al.*, 2004), estimated from genome-wide human, mouse and rat alignments. We used these models to generate neutrally evolving LF-SINE instances at increasing branch lengths between

3

0 and 2 substitutions per site, at increments of 0.01. At each branch length we generated 10^6 decayed sequences, and then attempted to detect them using Blastz, parameterized as in Text S2. Compared to a species tree (Fig. S4) estimated from four-fold degenerate codon positions between available genome drafts (Margulies *et al.*, 2005), the results shown in Fig. S5 predict that, ignoring larger scale rearrangements and repeat expansion, single-copy neutrally evolving human genomic DNA can often be matched to its ortholog in current mammalian genome drafts up to 0.6 subst./site away, but it can seldom be traced in chicken, and never in frog or beyond. These predictions are in good correspondence with current genome-wide comparisons of these organisms (International Mouse Genome Sequencing Consortium, 2002; International Rat Genome Sequencing Consortium, 2004; International Chicken Genome Sequencing Consortium, 2004; International Tetraodon Genome Sequencing Consortium, 2004).

To examine whether the actual human instances we observe have evolved neutrally, we used the liftOver utility at the UCSC genome browser to align orthologous LF-SINE instances from chimp, macaque, mouse, rat, dog, possum, chicken and frog, where available, to each human instance. Rare cases where two or more human instances were matched to the same region in a second species were resolved manually. We then used phyloP from the phast package (Siepel *et al.*, 2005), with the same reversible rate matrix and species tree of Fig. S4 to estimate how well the observed multiple alignments match our neutral evolution model. Even when treating the alignments conservatively, and setting a strict threshold of p <0.05 after Bonferroni correction, more than half the alignment blocks (137/245) clearly evolve slower than neutral evolution dictates.

S4   Human Instances

We performed several *in silico* tests in search of cellular role hypotheses for the human LF-SINE instances:

Energy based mfold prediction (Zuker, 2003) suggests a stable RNA secondary structure for the coelacanth SINE (Fig. S3) and a few of the human instances (Table S9). However, comparison of these instances failed to detect structure preserving substitutions in the human genome. According to the UCSC genome browser, only 10/163 intergenic instances overlap, even partially, any mature RNA transcript (EST/cDNA), mostly at or immediately downstream of the 3' end of a longer, spliced transcript. Additional 19 intergenic instances overlap the recent 5nt resolution transcriptional maps of 10 human chromosomes in 8 cell lines (Cheng *et al.*, 2005). None of these overlap inferred transcripts matching the expected retroposon length. Human LF-SINE instances do not overlap known small RNAs, micro-RNA precursors, or putative RNA editing events (Blow *et al.*, 2004; Kim *et al.*, 2004; Levanon *et al.*, 2004). In particular, intronic instances are evenly distributed as sense (35) and antisense (33) with the host gene, none of which overlap a known antisense transcript to the host gene.

However, human LF-SINE instances tend to co-cluster (Fig. S6). An instance is almost three times as likely to appear within 1Mb of another LF-SINE instance as if the instances were randomly distributed within each chromosome (p $<10^{-5}$), suggesting the kind of synergy often observed between enhancers, distal cis-regulatory regions found up to 1Mb away from their target genes (Levine and Tjian, 2003). The instances also tend to cluster within 1Mb of ultraconserved elements (Fig. S7). Almost 10% of LF-SINEs in the human genome are within 200kb of an ultraconserved element and close to 30% are within 1Mb (p $<10^{-5}$). To simulate a random distribution, bases in the human genome were randomly picked to be the first base of

an element. If a selection fell within a gap of the hg17 assembly, another base was randomly selected. We ran two variants of this simulation. The first variant placed instances randomly in the genome. The second placed instances randomly in each chromosome so that the number of instances per chromosome in the simulation would equal that of the LF-SINE (Table S5). The quantitative results in both scenarios were similar, and we discuss hereafter only the latter. For the self-clustering we calculated the distance from the first base of an LF-SINE instance to the first base of the next LF-SINE in the genome. This yields $n-1$ distances for n occurrences on any given chromosome. Similar distances were recorded from each LF-SINE instance to the nearest ultraconserved (UC) element on the same chromosome. This yields n distances for n occurrences of the LF-SINE, except on chr. 21 which has 3 LF-SINE instances but no UC elements. The hg17 genomic coordinates of the UC elements were obtained from the electronic supplement of Bejerano *et al.* (2004b), at `http://www.soe.ucsc.edu/~jill/ultra.html`.

About a dozen ultraconserved elements have already been confirmed, and many more are suspected of overlapping distal enhancers of transcription factors involved in development (Nobrega *et al.*, 2003; Bejerano *et al.*, 2004b; Woolfe *et al.*, 2005; Ellingsen *et al.*, 2005). Indeed, we find non-exonic LF-SINE instances overlapping or within 1Mb of 35 genes annotated for transcription factor activity ($p = 2.8 \cdot 10^{-6}$). Additional instances are found near 24 genes involved in cell adhesion ($p = 7.3 \cdot 10^{-7}$) and overlapping or near 15 proteins all carrying the fibronectin type III repeat region, common among cell surface binding and receptor proteins ($p = 3.3 \cdot 10^{-6}$). These findings suggest that LF-SINE instances may be involved in neuronal development. The annotation enrichment p-values are computed against a null model comparing the GO (Ashburner *et al.*, 2000) and InterPro (Mulder *et al.*, 2005) annotations of a given set of genes, with their genome-wide distribution. To perform these tests we collected all GO and InterPro annotations available at the UCSC genome browser, and computed a hyper geometric p-value for each attribute in the set vs. its genome-wide distribution in all annotated gene loci in the UCSC known genes track, as of June 2005. All p-values we reported above remain highly significant ($p < 0.005$) after Bonferroni correction.

S5 Exonized Instances

All exonization events use the canonical GT/AG splice signals (Table S7, Fig. S10). However, they insert in all three possible reading frames, and use several different splice sites compared to the coelacanth SINE (Fig. S9). Bases 381 and 289 are the most common 3' and 5' splice site combination in the known human cases (Table S8). This combination, in which the consensus splice site bases are already found in the coelacanth SINE, is also shown in Fig. 1 with respect to PCBP2. The region around the preferable 3' splice site at base 381 is highly similar to its equivalent in exonized instances of the Alu-J primate specific SINE subfamily (Lev-Maor *et al.*, 2003). Both have a run of pyrimidines followed by proximal and distal AG splice sites, which will be in competition and may aid in alternative splicing of the exon. In particular, the sequences surrounding the potential splice sites for both SINEs differ by a single base insertion:

```
LF-SINE: AGAACAGG (bases 383-376)
Alu-J  : AGA-CAGG (bases 281-275)
          ||    ||
       proximal  distal
```

In the Alu-J family, this splice site regulatory sequence is often preceded by a G (at pos. 282), coupled with a strong preference for the distal AG (Lev-Maor *et al.*, 2003). In contrast, in exapted LF-SINE instances the sequence is preceded by a C (pos. 384), showing preference for the proximal AG (Fig. S9, page 20). There is no apparent similarity to the 5' splice site in the Alus beyond the canonical bases.

Interestingly, in the coelacanth sequence of Fig. S2 (as well as in its Fig. S14 variant) the 93 bases spanned by the most common 381/289 splice site combination (Table S8) is free of stop codons in all three reading frames of the SINE reverse strand (which is invariably used in all exonizations we detected). Supporting this observation we find that all six exons that do not insert a stop codon into the reading frame (with the exception of the 5' splice site of KIAA1409) use this splice site combination. In contrast only two exons that use this splice site combination do introduce a stop codon (Table S7, S8). This capacity may make the LF-SINE prone to rapid exonization, following integration in a favourable genomic location, generating a novel coding exon. This exon may then be subjected to positive selection, followed by negative selection, as it acquired a cellular role (Long *et al.*, 2003). To assess the significance of this finding we generated 10^7 sequences of 481bp from a first-order Markov model trained on the 1Mb of available coelacanth sequence. In each of these sequences we searched for a stretch of 93bp or longer, stop codon free in all three frames, and flanked by the GGT/AG splice sites. The simulation indicates that this event is rare, $p \sim 0.002$, in random coelacanth-like sequence.

A Blastx (Altschul *et al.*, 1997) of this region against the NCBI protein database (nr) retrieved two additional candidate exonized instances: An alternatively-spliced exon of human PTGER3, and an unannotated exon of chicken PRDM5 homolog (weak *E*-value is strengthened in both cases by the fact that we match a discrete exon while comparing in protein space). Expanding the Blastx to search all frames of the entire LF-SINE we find a single additional, strong hit to a predicted exon of chicken EMSY homolog, which is out of the golden path in the current chicken genome draft, and was thus excluded from our screen (Fig. S11).

Nonsense-mediated decay (Table S7) was predicted in 11/17 coding instances where the insertion of an LF-SINE exon causes a stop codon to appear more than 50 nucleotides upstream of the final splice junction. These transcripts are almost always quickly degraded in eukaryotes (Lewis *et al.*, 2003).

We found three human single nucleotide polymorphisms (SNPs) in exonized instances of the LF-SINE in dbSNP (Sherry *et al.*, 2001). Two are in the alternatively-spliced exon of PCBP2 (rs11556936, rs11556932). However, these SNPs have not been validated and the sequence reported as flanking the polymorphism does not align well to the human genome. The third, and only validated SNP (rs4905082), in KIAA1409, is a silent wobble position transition that does not change the identity of the encoded amino acid.

PCBP2 encodes a poly-C binding protein, containing three instances of the RNA binding KH domain. Members of this family are involved in many biological processes including mRNA stabilization, translational silencing and enhancement, and various aspects of mRNA expression (Makeyev and Liebhaber, 2002). PCBP2 was shown in mice to have distinct intercellular expression and intracellular accumulation patterns (Hidai *et al.*, 2003; McKee *et al.*, 2005). It expresses two major isoforms, differing only by including or excluding the LF-SINE-derived exon, which is located in an unannotated region between the second and third KH domains. Recent *in vitro* work has shown that the isoform including the LF-SINE-derived exon is localized to the nucleus, while the isoform excluding it is found both in the nucleus and cytoplasm, suggesting that the exon may harbor a nuclear retention signal (Chkheidze and Liebhaber, 2003). This exon may also

be involved in dimerization with itself and other hnRNP proteins (Kim *et al.*, 2000), and post-transcriptional auto-regulation (Waggoner and Liebhaber, 2003).

It was recently shown (Gunduz *et al.*, 2005) that SMARCA4 (also called BRG1), a member of the SWI/SNF complex and a candidate tumor suppressor gene in several human cancers, is generally up-regulated in oral cancer, but the isoform containing the LF-SINE-derived exon is extremely decreased or lost in most tumor samples. In another recent study, of LRP1B (Li *et al.*, 2005), a low density lipoprotein (LDL) receptor-related protein and also a candidate tumor suppressor gene, the isoforms including and omitting the LF-SINE-derived exon were shown to express differently in human and mouse.

LRP1B is one of three LF-SINE instances most conserved between mammals, bird and amphibian. The other two being the novel ISL1 enhancer, and an instance which according to mouse ESTs is embedded in the 3' UTR of HHIP, or hedgehog interacting protein. Interestingly all three proteins may be involved in hedgehog signalling and/or pancreatic development. HHIP is a known member of the hedgehog signalling network. LRP2 (also megalin), which is closely related to LRP1B (May and Herz, 2003), is also known, and ISL1 is suspected of being involved in this network too (Cohen, 2003). ISL1 is also an important factor in pancreatic development (Habener *et al.*, 2005), as are HHIP (Kawahira *et al.*, 2003) and LRP5 (Fujino *et al.*, 2003), which is more distantly related to LRP1B (May and Herz, 2003).

S6   Similar Families

Previously, we have shown that thousands of paralog families exist within the conserved non coding portion of the human genome (Bejerano *et al.*, 2004a, `http://www.cse.ucsc.edu/~jill/dark.html`). That screen however combined paralogs found in much larger matching genomic contexts (e.g., conserved non coding regions that duplicated along with new members of a gene family) with paralogous matches with little or no flanking context (such as would be the case for retroposition exaptation events).

To estimate the frequency of detectable SINE-like exaptation events in the human genome, we performed two screens, in search of genomic elements, known or suspected as functional, that have multiple short paralogous matches elsewhere within the human genome. In the first screen we attempted to find multiple short paralogous matches that overlap and extend into the intronic region flanking short exons (similar to Fig. S1). Of the retrieved set we estimate that dozens, possibly more, exons are similar to short genomic regions unrelated to family members, recent duplications, and retrogenes of the respective host gene (e.g., Fig. S12). The second screen focused on the top 2% multi-species conserved elements in the human genome. From these we removed all known exons, RNA genes, retrogenes, repeats and elements lying in recent segmental duplication regions. In the remainder of highly conserved uncharacterized elements we estimate that hundreds, or more, elements have multiple short matches elsewhere in the genome. Work is in progress to refine these screens to produce putative exapted candidates for further analysis.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids*

Res, **25**, 3389–3402.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–29.

Bejerano, G., Haussler, D. and Blanchette, M. (2004a) Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics*, **20**, i40–i48.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. and Haussler, D. (2004b) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.

Blow, M., Futreal, P. A., Wooster, R. and Stratton, M. R. (2004) A survey of RNA editing in human brain. *Genome Res*, **14**, 2379–2387.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S. and Gingeras, T. R. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.

Chkheidze, A. N. and Liebhaber, S. A. (2003) A novel set of nuclear localization signals determine distributions of the alphaCP RNA-binding proteins. *Mol Cell Biol*, **23**, 8405–8415.

Cohen, M. M., Jr. (2003) The hedgehog signaling network. *Am J Med Genet A*, **123**, 5–28.

Cooper, G. M., Brudno, M., Stone, E. A., Dubchak, I., Batzoglou, S. and Sidow, A. (2004) Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res*, **14**, 539–548.

Crooks, G. E., Hon, G., Chandonia, J.-M. and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188–1190.

Danke, J., Miyake, T., Powers, T., Schein, J., Shin, H., Bosdet, I., Erdmann, M., Caldwell, R. and Amemiya, C. T. (2004) Genome resource for the Indonesian coelacanth, Latimeria menadoensis. *J Exp Zoolog A Comp Exp Biol*, **301**, 228–234.

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–1797.

Ellingsen, S., Laplante, M. A., Konig, M., Kikuta, H., Furmanek, T., Hoivik, E. A. and Becker, T. S. (2005) Large-scale enhancer detection in the zebrafish genome. *Development*, **132**, 3799–3811.

Felsenstein, J. (1995) http://evolution.genetics.washington.edu/phylip.html. Version 3.5c.

Friedman, N., Ninio, M., Pe'er, I. and Pupko, T. (2002) A structural EM algorithm for phylogenetic inference. *J Comput Biol*, **9**, 331–353.

Fujino, T., Asaba, H., Kang, M.-J., Ikeda, Y., Sone, H., Takada, S., Kim, D.-H., Ioka, R. X., Ono, M., Tomoyori, H., Okubo, M., Murase, T., Kamataki, A., Yamamoto, J., Magoori, K., Takahashi, S., Miyamoto, Y., Oishi, H., Nose, M., Okazaki, M., Usui, S., Imaizumi, K., Yanagisawa, M., Sakai, J. and Yamamoto, T. T. (2003) Low-density lipoprotein receptor-related protein 5 (LRP5) is essential for normal cholesterol metabolism and glucose-induced insulin secretion. *Proc Natl Acad Sci U S A*, **100**, 229–234.

Gilbert, N. and Labuda, D. (1999) CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. *Proc Natl Acad Sci U S A*, **96**, 2869–2874.

Gilbert, N. and Labuda, D. (2000) Evolutionary inventions and continuity of CORE-SINEs in mammals. *J Mol Biol*, **298**, 365–377.

Gunduz, E., Gunduz, M., Ouchida, M., Nagatsuka, H., Beder, L., Tsujigiwa, H., Fukushima, K., Nishizaki, K., Shimizu, K. and Nagai, N. (2005) Genetic and epigenetic alterations of BRG1 promote oral cancer development. *Int J Oncol*, **26**, 201–210.

Habener, J. F., Kemp, D. M. and Thomas, M. K. (2005) Minireview: transcriptional regulation in pancreatic development. *Endocrinology*, **146**, 1025–1034.

Hidai, C., Ohno, M., Kazama, H., Ikeda, H., Kasanuki, H., Kokubun, S. and Kawana, M. (2003) Specific regulation of nucleocytoplasmic distribution of poly(C)-binding protein gene mRNA in mouse development. *Biochem Biophys Res Commun*, **309**, 339–343.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., Sultan-Qurraie, A., Thomas, D. J., Trumbower, H., Weber, R. J., Weirauch, M., Zweig, A. S., Haussler, D. and Kent, W. J. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*, **34**, 590–598.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

International Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

International Rat Genome Sequencing Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

International Tetraodon Genome Sequencing Consortium (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–957.

Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, **16**, 418–420.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F. G., Harte, N.,

Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M. A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*, **33**, 29–33.

Kawahira, H., Ma, N. H., Tzanakakis, E. S., McMahon, A. P., Chuang, P.-T. and Hebrok, M. (2003) Combined activities of hedgehog signaling inhibitors regulate pancreas development. *Development*, **130**, 4871–4879.

Kazazian, H. H., Jr. (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, **100**, 11484–11489.

Kim, D. D. Y., Kim, T. T. Y., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S. and Gabriel, A. (2004) Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res*, **14**, 1719–1725.

Kim, J. H., Hahm, B., Kim, Y. K., Choi, M. and Jang, S. K. (2000) Protein-protein interaction among hnRNPs shuttling between nucleus and cytoplasm. *J Mol Biol*, **298**, 395–405.

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

Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Sztybel, D., Olshansky, M., Rechavi, G. and Jantsch, M. F. (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol*, **22**, 1001–1005.

Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

Lewis, B. P., Green, R. E. and Brenner, S. E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A*, **100**, 189–192.

Li, Y., Lu, W. and Bu, G. (2005) Striking differences of LDL receptor-related protein 1B expression in mouse and human. *Biochem Biophys Res Commun*, **333**, 868–873.

Long, M., Betran, E., Thornton, K. and Wang, W. (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet*, **4**, 865–875.

Lopez, R., Silventoinen, V., Robinson, S., Kibria, A. and Gish, W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res*, **31**, 3795–3798.

Makeyev, A. V. and Liebhaber, S. A. (2002) The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms. *RNA*, **8**, 265–278.

Margulies, E. H., Maduro, V. V. B., Thomas, P. J., Tomkins, J. P., Amemiya, C. T., Luo, M. and Green, E. D. (2005) Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc Natl Acad Sci U S A*, **102**, 3354–3359.

May, P. and Herz, J. (2003) LDL receptor-related proteins in neurodevelopment. *Traffic*, **4**, 291–301.

McGinnis, S. and Madden, T. L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, **32**, 20–25.

McKee, A., Minet, E., Stern, C., Riahi, S., Stiles, C. and Silver, P. (2005) A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC Dev Biol*, **5**, 14.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Pagni, M., Ponting, C. P., Quevillon, E., Selengut, J., Sigrist, C. J. A., Silventoinen, V., Studholme, D. J., Vaughan, R. and Wu, C. H. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*, **33**, 201–205.

Nishihara, H., Smit, A. and Okada, N. (2006) Functional non-coding sequences derived from sines in the human genome. *Genome Res,*, p. Submitted.

Nobrega, M. A., Ovcharenko, I., Afzal, V. and Rubin, E. M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.

Ogiwara, I., Miya, M., Ohshima, K. and Okada, N. (2002) V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. *Genome Res*, **12**, 316–324.

Ostertag, E. M. and Kazazian, H. H., Jr. (2001) Biology of mammalian L1 retrotransposons. *Annu Rev Genet*, **35**, 501–538.

Pearson, W. R., Wood, T., Zhang, Z. and Miller, W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res*, **13**, 103–107.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308–311.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W. and Haussler, D. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034–1050.

Smit, A. and Green, P. (2005) http://ftp.genome.washington.edu/RM/RepeatMasker.html.

Uemura, O., Okada, Y., Ando, H., Guedj, M., Higashijima, S.-I., Shimazaki, T., Chino, N., Okano, H. and Okamoto, H. (2005) Comparative functional genomics revealed conservation and diversification of three enhancers of the *isl1* gene for motor and sensory neuron-specific expression. *Dev Biol*, **278**, 587–606.

Waggoner, S. A. and Liebhaber, S. A. (2003) Identification of mRNAs associated with alphaCP2-containing RNP complexes. *Mol Cell Biol*, **23**, 7055–7067.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E. and Elgar, G. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, **3**, e7.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406–3415.
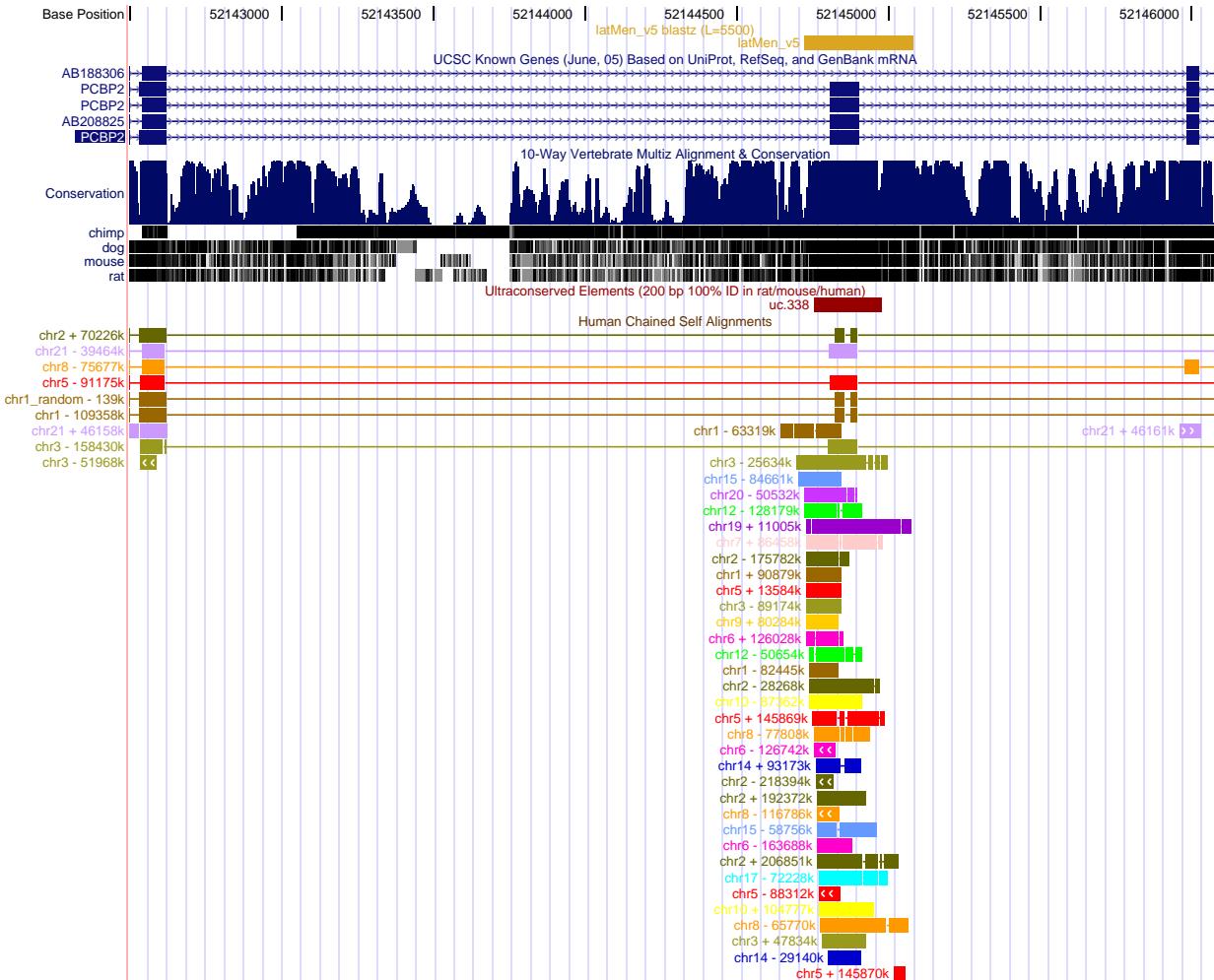
Figure S1: **Distinctive accumulation of short human paralogs to the PCBP2 exonized instance containing uc.338.** A UCSC genome browser shot (http://genome.ucsc.edu) of the PCBP2 exonized instance and the two exons flanking it (3.6kb region). Tracks (top to bottom) show: The region conserved with coelacanth; PCBP2 (whole and fragmented) isoforms, showing the alternatively-spliced nature of the exonization event; Multi-species conservation track (Siepel *et al.*, 2005); Location of uc.338 within the exapted SINE; Chained (Kent *et al.*, 2003) human paralogs to this genomic region. The top seven paralogs conserve, alongside the exonized instance, other portions of PCBP2. All are PCBP2 retro-genes, of which the top (on chr. 2) is PCBP1, a functional retroposed copy of PCBP2 (Makeyev and Liebhaber, 2002). All other paralogs, similar to the exonized SINE alone, are other human instances of the LF-SINE.

```
>LF-SINE
GGGGACTGGATGGCTCAGTGGAATTGGTAATGGGATATGGAGCCTTTCAC
CTCTAGGTCACTGGGTTCAAATCCAGCCCAGGTCAGTAGTGACCGAAAGT
CATTACCATCTGATGGCTGTTCAGTGGCCTATGTGAAATGAGTTGGTGGT
CTCAGTCCAGTTCCTAGTGGACAGGTGTCCACATCACAAAACCACCATCA
CAATTGGCACTAATTGGCACCCTTGTTGGCAGTCTCAGCAGAGAGGCCAA
GGATTGAATGGGCATGGAGACTGAACTACCCTCTCAACCCTGTAGAGGTG
GTCCCTCCAGGGCAGGGTTGAGGCACATTGGCAGGGCAATGTGGGGAAGC
CTGCACTGCTGCTGCCCATGCTGTACCTGTTCTGTGGATAAATAGAGGAC
TTCAGTCTCTGGTGCTATCAATCTAGCACCTTTCACGAGCACTAAATTCA
CACAAAAAAATTTAAAAAAAAAAAAAAAAAAA
```

Figure S2: **Reconstructed Coelacanth LF-SINE.** The LF-SINE was reconstructed from 59 available instances in the Indonesian coelacanth, based on similarity to ultraconserved element uc.338. The SINE is tRNA derived and believed to be recently active in coelacanth, and long-inactivated in tetrapods. See also Fig. 1 and Fig. S14.
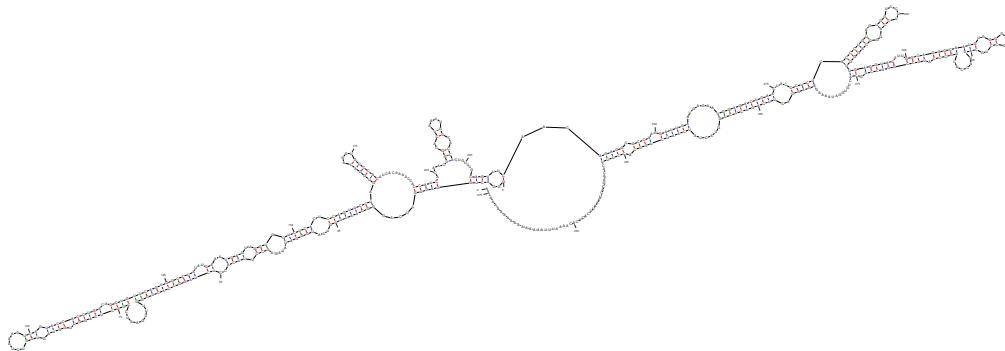

Figure S3: **RNA Secondary structure potential of the LF-SINE.** An energy based mfold (Zuker, 2003) prediction for the closest coelacanth instance to the reconstructed SINE. Secondary structure prediction for the PCBP2 exonized instance containing uc.338 is qualitatively similar (not shown).

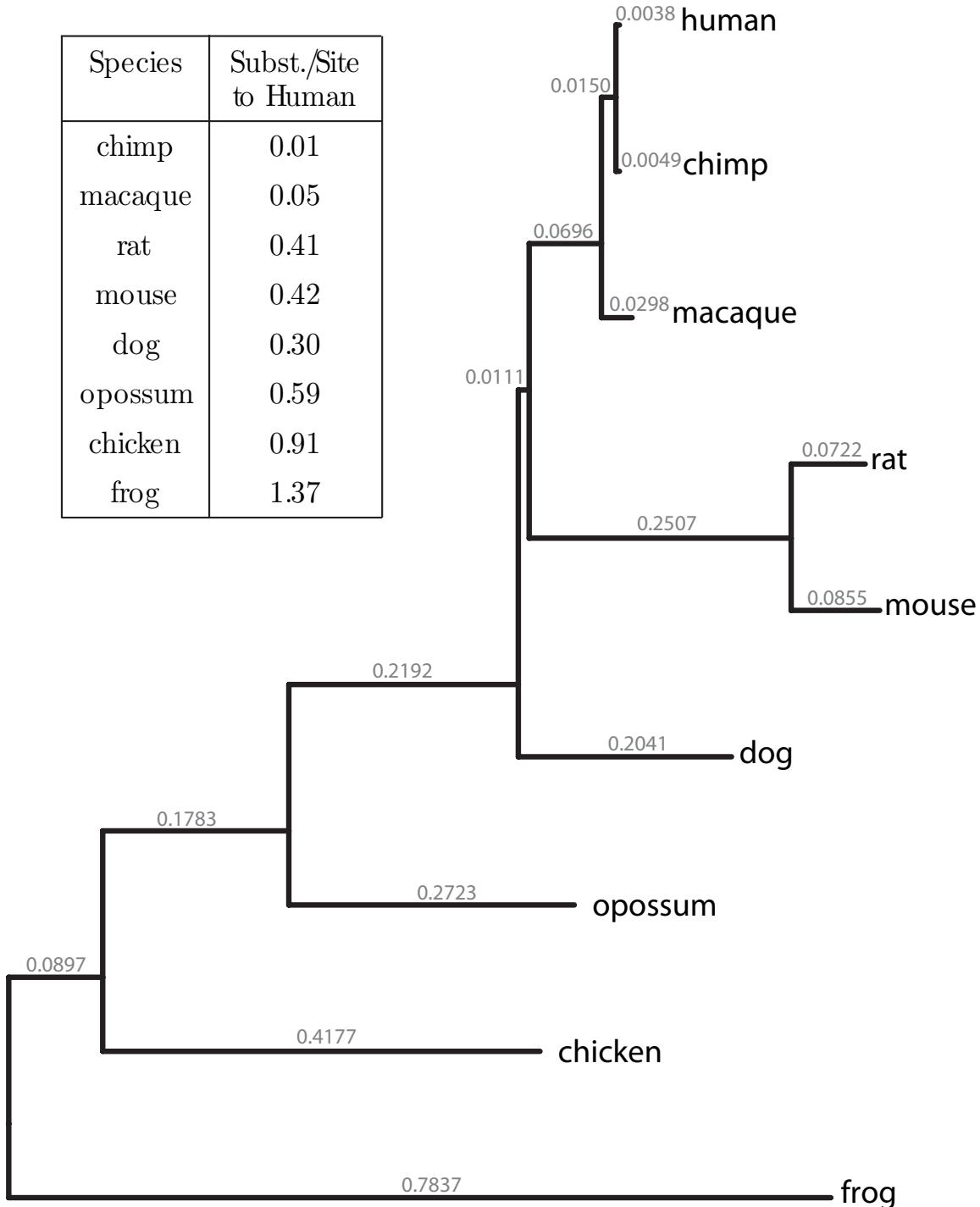| Species	Subst./Site to Human
chimp	0.01
macaque	0.05
rat	0.41
mouse	0.42
dog	0.30
opossum	0.59
chicken	0.91
frog	1.37

Figure S4: **A tetrapod phylogenetic tree.** This tree, adapted from (Margulies *et al.*, 2005), was used in estimating the conservation p-values for the human LF-SINE instances and their orthologs.
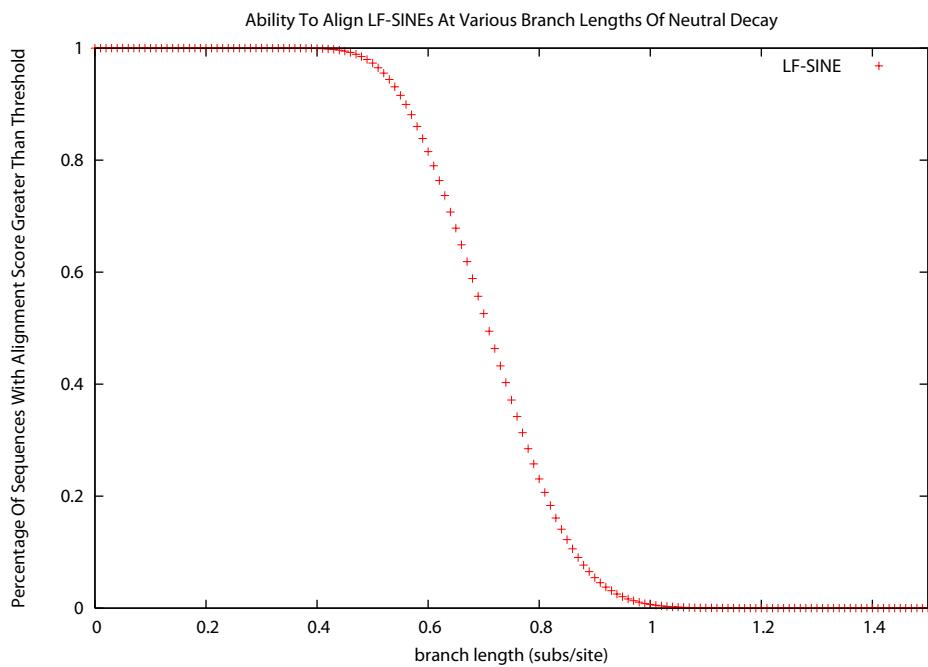
Figure S5: **Detecting homology using pairwise sequence similarity as a function of divergence.** The LF-SINE was placed under an estimated model of neutral decay, at branch lengths of 0-2 substitutions per site, in increments of 0.01. At each sample point we generated 10^6 sequences, and attempted to recover them using Blastz. We plot the fraction of recovered instances above our scoring threshold.
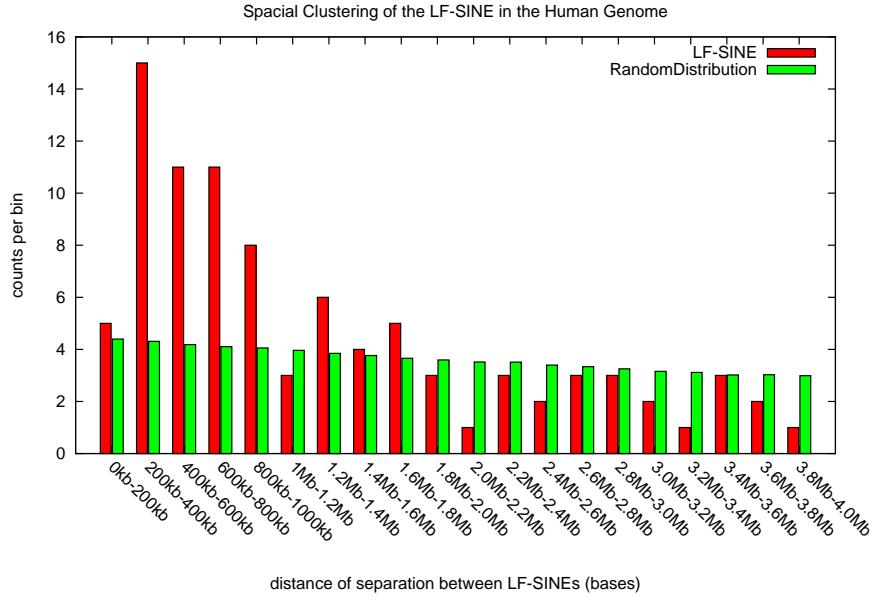
Figure S6: **LF-SINE instances tend to cluster in the human genome.** The histogram shows a distinct preference of the LF-SINE to be within 1MB of another copy. The random distribution has the same number of instances per chromosome, randomly placed avoiding actual assembly gaps in hg17. Based on 10^5 iterations.
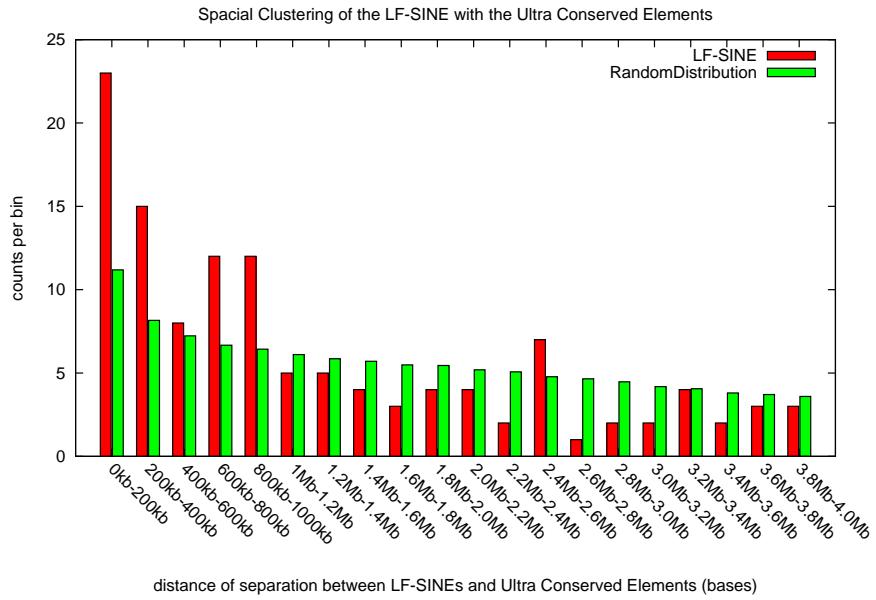

Figure S7: **Human LF-SINE instances are enriched in proximity to ultraconserved elements.** The histogram shows an enrichment for LF-SINE instances within 1MB of human ultraconserved elements. Random distribution as in Fig. S6. Based on 10^5 iterations.
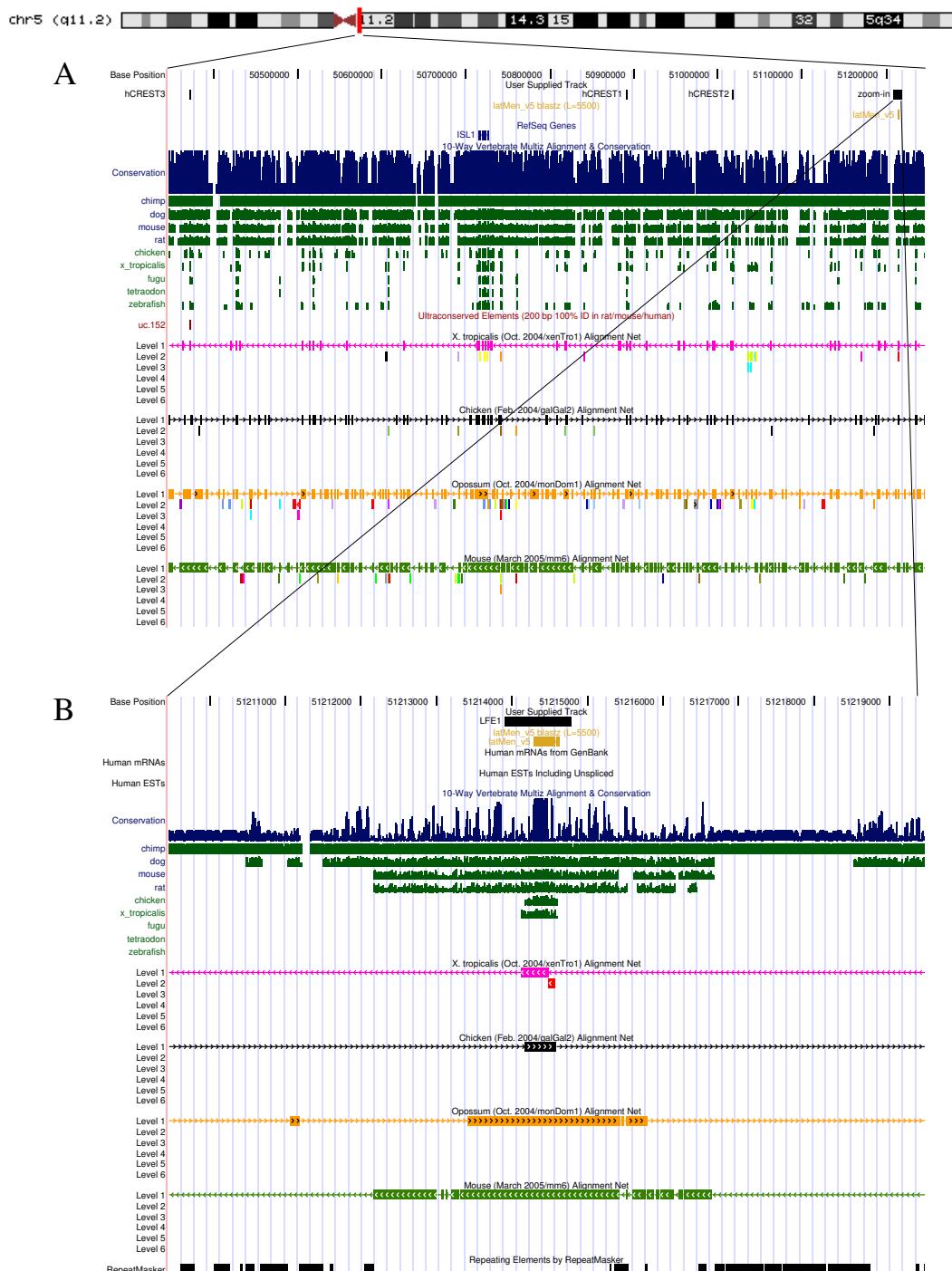
Figure S8: **Genomic context of the novel ISL1 enhancer.** ISL1 resides in a pericentromeric region of chr. 5q (top). (A) blow up of 900kb Containing ISL1, three known distal enhancers, hCREST1-3 (Uemura *et al.*, 2005), the latter overlapping ultraconserved element uc.152, and the LF-SINE instance ("LatMen_v5"). Frog, chicken, opossum and mouse alignment nets (Kent *et al.*, 2003) show that these functional units have conserved their relative ordering since the amphibian-amniote split. (B) Further blow up of 10kb containing our validated enhancer ("LFE1" for living-fossil enhancer), showcasing its multi-species conservation and repeat free immediate vicinity.

```
                      481                                 450
SINE-RevComp          TTTTTTTTTTTTTTTTTTTAAATTTTTTTGTGTGAATTT-AGTGCTCGTGAAAGGTGCTAG-ATTGATAGCACCAG

hg-RORA_f2            ---------------------------------------------------------------------------

mm-RORA_f2           ---------------------------------------------------------------------------

hg-EEF1B2_f1         ---------------------------------------------------------------------------

hg-PTDSR_f1          ---------------------------------------------------------------------------

mm-PTDSR_f1          ---------------------------------------------------------------------------

hg-KIAA1409_f2       ---------------------------------------------------------------------------

mm-KIAA1409_f2       ---------------------------------------------------------------------------

gg-KIAA1409_f2       ---------------------------------------------------------------------------

hg-PCBP2_f0          ---------------------------------------------------------------------------

mm-PCBP2_f0          ---------------------------------------------------------------------------

hg-SMARCA4_f0        ---------------------------------------------------------------------------

mm-SMARCA4_f0        ---------------------------------------------------------------------------

hg-FLJ22833_f1       ---------------------------------------------------------------------------

hg-ARHGAP6_f1        ---------------------------------------------------------------------------

mm-ARHGAP6_f1        ---------------------------------------------------------------------------

hg-LRP1B_f0          ---------------------------------------------------------------------------

hg-DHX30_f2          ---------------------------------------------------------------------------

gg-PPP2R2C_f2        ---------------------------------------------------------------------------

xt-MBNL1_f0          ---------------------------------------------------------------------------
                                       |  E   F   S   A   S   E   R   C   H   I   A   N   L
hg-NT5C2_f1          --------------------TTTTCTAAAGtgaaTTT-agtGCTagtGAGagaTGCcac-ATTgccAACctgAG
                                       |  E   F   S   A   S   E   R   C   H   I   A   S   L
mm-NT5C2_f1          --------------------TTTTCTAAAGtgaaTTT-agtGCTagtGAGagaTGCcac-ATTgccAGCctgAG
                                       |  *   I   Y   C   S   *   K   V   P   D   *   Q   L   W
hg-ATF2_f0           --------------------TTATTTAAAGtgaATAt-atTGCtcgTGAaagGTGccaG-ATtgaCAGcttTGG
                                       |  E   F   *   C   S   *   K   V   S   D   *   K   P   W
gg-DMTF1_f1          --------------------GTTTTTTTAGtgaaTTTttagTGCtcgTGAaagGTGtcaG-ATtgaAAAcccTGG
                                       |  *   F   *   C   S   *   T   V   P   D   *   Q   N   W
JGI-49280_f0         --------------------ATTTCTCTAGtgaTTTt-aaTGCtcgTGAacgGTTccaG-ATtgaCAGaacTGG
                               |   *   F   S   E   R   C   E   S   E   S   L   *   K   M   L   G   L   A   A   P
hg-TCERG1_f1         --ATTTTTATAGttaaTTTagtGAAagaTGTgaaAGTg-aaAGCttgTGAaagATGctgGGAttgGCAgccCCGc
                               |   *   F   S   E   R   C   E   S   E   S   L   *   K   M   L   G   L   A   A   P
mm-TCERG1_f1         --ATTTTTATAGttaaTTTagtGAAagaTGTgaaAGTg-aaAGCttgTGAaagATGctgGGAttgGCAgccCCGc
```

Figure S9: **Alignment of tetrapod exonized instances to the coelacanth SINE.** In all known cases where an LF-SINE instance has undergone exonization, the LF-SINE is found on the opposite strand relative to the gene. The reverse complement of the LF-SINE sequence from Fig. S2 is displayed in the top row. All known instances of exonization are presented at both the amino acid and nucleic acid level. Pipes are placed on the first bases outside of the exons and the case of coding bases is used to show the codon structure. The fN suffix to the gene name indicates the entry reading frame.

19

```
                             400                                                            350
SINE-RevComp    AGACTGAAGTCCTCTATTTATCCACAGAACAGGTAC---AGCATGGGCAGCAGCAGTGC----AG---GCT----
                                                    |  Q  A  W  Y       R  A
hg-RORA_f2      -------------------------------ATAT---CCGCAGggcagGCAtggTAC----agaGCGct----
                                                    |  Q  A  W  Y       R  E
mm-RORA_f2      -------------------------------ATAT---CCGCAGggcagGCAtggTAC----agaGAGct----
                                              |  *     N  M  D  S  S  N       S  T
hg-EEF1B2_f1    ---------------------CACAGAACAGgtag---AACatgGACagcAGCaatTC----Aac--tTT----
                                              |  Y     S  T  G  S  G  S       A  S
hg-PTDSR_f1     ---------------------CACAGAACAGatac---AGCacgGGCagcGGCagtGC----Aa---gcC----
                                              |  Y     S  T  G  S  G  S       A  S
mm-PTDSR_f1     ---------------------CACAGAACAGatac---AGCacgGGCagcGGCagtGC----Aa---gcC----
                                        |  Q  I     Q  P  G  K  R  Q  C
hg-KIAA1409_f2  -----------------TTATCCACAGaacagATAc---agCCTgggAAAcgcCAGtgt----AA----------
                                        |  Q  I     Q  P  G  K  R  Q  C
mm-KIAA1409_f2  -----------------TTATCCACAGaacagATAc---agCCCgggAAGcggCAGtgt----AA----------
                                        |  Q  I     Q  H  G  K  R  L  C
gg-KIAA1409_f2  -----------------TTATTCACAGagcagATAc---agCATggaAAGcggCTAtgc----A-----------
                                        |  D  R  Y     S  T  G  S  D  S       A  S
hg-PCBP2_f0     -----------------TTATCCACAGgacAGGtac---AGCacaGGCagcGACagtGC----Ga---gcT----
                                        |  D  R  Y     S  T  G  S  D  S       A  S
mm-PCBP2_f0     -----------------TTATCCACAGgacAGGtac---AGCacaGGCagcGACagtGC----Ga---gcT----
                                        |  S  R  H  C  S  T  G  S  G  S       A  S
hg-SMARCA4_f0   -----------------TTATCCACAGagcAGAcacTGCagcACGggcAGCggcAGTgc----cA---GCt----
                                        |  S  R  H  C  S  T  G  S  G  S       A  S
mm-SMARCA4_f0   -----------------TTATCCACAGagcAGAcacTGCagcACGggcAGCggcAGTgc----cA---GCt----
                                        |  L  G     A  V  Q  A  A  A  M       R     D
hg-FLJ22833_f1  -----------------TTATCCACAGtctaGGTgc---gGTGcagGCAgccGCAatgC----GA---gat----
                                        |  T  S     S  V  L  P  A  A  V       Q     A
hg-ARHGAP6_f1   -----------------TTGTCCCCAGaacaAGTtc---tGTGcttCCGgctGCAgtgC----AG---gcc----
                                        |  T  S     S  V  L  P  A  A  G       Q     A
mm-ARHGAP6_f1   -----------------TTATTCACAGaaccAGTtc---tGTAttgCCGgccGCAgggC----AG---gct----
                                        |  A  R  Y     I  G  G  G  P  S  A  F       K
hg-LRP1B_f0     -----------------TTATATACAGgccAGGtac---ATAgggGGAggaCCCagtGCTttcAA---Gct----
                                        |  R  A     Q  H  R  Q  R  Q  C           K
hg-DHX30_f2     -----------------CTTTCCCCAGatcggGCCc---agCACaggCAGcgtCAGtgc----AA---Act----
                                        |  E  V     *  Q  G  Q  Q  Q  C           K
gg-PPP2R2C_f2   -----------------TTCTCCACAGgagagGTAt---agCAAgggCAGcagCAAtgc----AA---Gct----
                                        |  A  S  Y     N  Q  A  A  A  V       R     A
xt-MBNL1_f0     -----------------TTATCCACAGgccAGCtac---AAT-cagGCAgctGCAgtgC----GA---gct----
                 R  D                                 G     T  A  P  A  A  P  C           R
hg-NT5C2_f1     Agat------------------------GGGac---aGCAccaGCAgctCCA-tgc----AG---Gct----
                 R  D                                 G     Q  R  Q  Q  L  H       A     G  S
mm-NT5C2_f1     Agat------------------------GGAc---agCGCcagCAGctcCA-Tgc----aG---GCtctCA
                    K  L  K  D  S  N  Y  P  Q  T  |
hg-ATF2_f0      aaaCTGaagGACtctAATtatCCAcagACCagGTAC---AATATG-----------------------------
                    R  L  K  S  S     L  F  H  R  T  |
gg-DMTF1_f1     agaCTGaagTCTtct-TTAttcCACagaACagGTAC---AGCACA-----------------------------
                    K  L  Q  F  S  N  Y  P  Q  N  |
JGI-49280_f0    aaaCTGcagTTCtccAATtatCCAcagAACagGTAA---AGCATG-----------------------------
                 Q  T  E  V  L  C  S  S  T  E  Q|
hg-TCERG1_f1    agACTgaaGTCctcTGTtcaTCCacaGAAcagGTAC---AGCACG-----------------------------
                 Q  T  E  V  L  C  S  S  T  E  Q|
mm-TCERG1_f1    agACTgaaGTCCctcTGTtcaTCCacaGAAcagGTAC---AGCACG-----------------------------
```

Figure S9: **Alignment of tetrapod exonized instances to the coelacanth SINE (contd.)**

```
                                                               300
SINE-RevComp    -T-C-CCCAC---ATTGCC--C-------T---GCCAATGTGCCTCAACCCTGCCCT-GGAGGGACCA-CC-TCT
                L   P   C     T  A            L  P  M   S   S  S  L   A   R  R  D    P  L
hg-RORA_f2      -t-C-CTtgc---ACAgcc--T-------T---GcccATGag-cTCGtccCTGgccC-GGaggGACcc-aC-TCt
                L   P   C     T  A            L  P  M   S   S  S  L   A   W  R  D    P  L
mm-RORA_f2      -t-C-CTtgc---ACAgcc--T-------T---GcccATGagc-TCGtccCTGgccT-GGaggGACcc-aC-TCt
                F   P   H     T  A            L  P  M   C   L  D  L   D   L  G  G    P  P
hg-EEF1B2_f1    -T-c-ccCAC---actGCC--c-------t---gCCAatgTGCctcGACcttGACct-gGGGggcCCA-cc-tCT
                  H   P     Q  P    P         R    A  T  V  S  Q  P  *  P   G  G  T     S  S
hg-PTDSR_f1     -A-C-ccaCA---GccaCC--Ccg-----t---GCCactGTGtccCAAcccTGAcct-GGAgggACCa-gc-TCT
                  H   P     Q  P    P         R    A  T  V  S  Q  P  *  P   G  G  T     S  S
mm-PTDSR_f1     -A-C-cccCA---GccaCC--Ccg-----t---GCCactGTGtccCAAcccTGAcct-GGAgggACCa-gc-TCT
                                N                V   P  T  C  L  N  P  D   L  E  G  Q     P
hg-KIAA1409_f2  ------------------C--g-------t---gCCAacgTGCctaAACcctGACct-gGAGggaCAG-cc-aTT
                                N                V   P  M  C  L  N  P  D   L  E  G  Q     P
mm-KIAA1409_f2  ------------------C--g-------t---gCCAatgTGCctaAACcctGACct-gGAGggaCAG-cc-aTT
                             N  V          P          V   C  L  N  P  D   L  E  G  P     P
gg-KIAA1409_f2  -------------ATgtcC--C-------T---g----ttTGCctaAACcctGACtt-gGAGggaCCA-cc-aCT
                  F   P   H     T  T           P   S  M  C  L  N  P  D   L  E  G  P     P
hg-PCBP2_f0     -TTc-ccCAC---accACC--c-------c---gTCCatgTGCctcAACcctGACct-gGAGggaCCA-cc-tCT
                  F   P   H     T  T           P   S  M  C  L  N  P  D   L  E  G  P     P
mm-PCBP2_f0     -TTc-ccCAC---accACC--c-------c---gTCCatgTGCctcAACcctGACct-gGAGggaCCA-cc-tCT
                  F   A   H     T  A           P  P  P  A  G  V  N  P  D   L  E  E  P     P
hg-SMARCA4_f0   -t-cGCCcac---ACTgcc--C-------CTccgCCAgcgGGCgtcAACcccGACtt-gGAGgagCCA-cc-tCT
                  F   A   H     T  A           P  P  P  A  G  V  N  P  D   L  E  E  P     P
mm-SMARCA4_f0   -t-cGCCcac---ACTgcc--C-------CTccgCCAgcgGGCgtcAACcccGACtt-gGAGgagCCA-cc-tCT
                  S   I   H  Y  Y              P     G  N  D  L  H  P  D   L  E  E  P  S  S
hg-FLJ22833_f1  -T-C-CataCACtacTACc--c-------t---GGTaatG-ATctcCACcctGACct-gGAGgagCCAtcc-TCT
                  C   P     Q  Y               P     A  S  M  F  T  P  *  P   G  G  T     S
hg-ARHGAP6_f1   -T-G-CccaC---AGtacC--C-------T---gccAGCatgTTTacgCCCtgaCCT-ggaGGGaccA-GC-t--
                  C   S     Q  S               P     A  S  D  F  T  P  *  P   G  E  T     S
mm-ARHGAP6_f1   -T-G-CtcaC---AGagcC--C-------T---gccAGCgacTTTacaCCCtgaCCT-ggaGAGaccA-GC-t--
                  L   P   H     T  A           P  P  I  Y  L  N  S  D   L  K  G  P     L  T
hg-LRP1B_f0     -t-C-CAcac---ACAgcg--C-------C---GcccATCtacCTAaacTCTgatTT-GaaaGGAcca-CTAact
                  L   P   P     P  R  L        P  P  M  C  V  N  P  T  P   G  G  T     I  S
hg-DHX30_f2     -tCC-CccaC---CCcgcCTTc-------c---aCCCatgTGTgtcAACcctACCccaGGAgggACCa-tc-TCT
                  L   P   H     I  V           L  P  M  F  L  N  P  V   R  K  R  S       L
gg-PPP2R2C_f2   -t-C-CCcac---ATTgtc--C-------T---GccaATGttcCTCaacCCAgttCG-GaagAGGtca-CT-Tct
                  S   P     R  C               P     T  N  D  L  S  P  D   V  E  G  A  P  S
xt-MBNL1_f0     -T-C-CccgC---GTtgcC--C-------A---accAATg-acCTCagcCCCgatGT-GgagGGAgcaCCC-tct
                  L   S   P     P  T           R  P  L   S  L  P  V  C   S  H   H   F
hg-NT5C2_f1     -c-T-CAcct---CCCaca--C-------G---GcccCTCt-ccTTGccaGTGtgcT-GCagcCACca-cT-TCt
                  H   L   P     H  G     P  S  P     C  Q  C  A  A  A  M  F   G  R  D   H  L
mm-NT5C2_f1     Cc-t-cCCAc---acGGCc--ccTCTcctT---GCcagTGTgctGCAgccATGtttG-GAaggGACca-cT-TGt

hg-ATF2_f0      --------------------------------------------------------------------------

gg-DMTF1_f1     --------------------------------------------------------------------------

JGI-49280_f0    --------------------------------------------------------------------------

hg-TCERG1_f1    --------------------------------------------------------------------------

mm-TCERG1_f1    --------------------------------------------------------------------------
```

Figure S9: **Alignment of tetrapod exonized instances to the coelacanth SINE (contd.)**

```
                                                          250
SINE-RevComp   ACA--GGGTTGAGAGGGTAGTTCAGTCTCCATGCCCATTCAATCCTTGGCCTCTCTGCTGAGACTGCCAACAAGGGT
                 *  R  |
hg-RORA_f2     agAGGtgGT--AAGTAATT-------------------------------------------------------------
                 *  R  |
mm-RORA_f2     agAGGtgGT--AAGTAATC-------------------------------------------------------------
                   L  G  *  E  G  G  S              F  I  Q  S  L  A  I  L  L  K  L  P  T  R
hg-EEF1B2_f1   -----GgggTGAgagGGTggcTCAtt---------cATTcagTCCttgGCCattCTTctgAAActgCCAacaAGGg-
                   R  |
hg-PTDSR_f1    -cg--gGGTAAGAGAGG----------------------------------------------------------------
                   R  |
mm-PTDSR_f1    -cg--gGGTAAGAGAGG----------------------------------------------------------------
                   L     R  M  R |
hg-KIAA1409_f2 --G--aggATGagaGGTTAGTTGAG---------------------------------------------------------
                   L     R  T  R |
mm-KIAA1409_f2 --G--aggACAagaGGTTGGTTGAG---------------------------------------------------------
                   L     R  I  R |
gg-KIAA1409_f2 --A--agaATCagaGGTTAGTTCAG---------------------------------------------------------
                 L       E|
hg-PCBP2_f0    Aga---gGT-GAGAGGGG----------------------------------------------------------------
                 L       E|
mm-PCBP2_f0    Aga---gGT-GAGAGGGG----------------------------------------------------------------
                 L       K|
hg-SMARCA4_f0  Aaa---gGT-GAGAGGGG----------------------------------------------------------------
                 L       K|
mm-SMARCA4_f0  Aaa---gGT-GAGAGGGG----------------------------------------------------------------
                 L     G  V   * |
hg-FLJ22833_f1 cta--GGGgtg-TAAgGTAATTCAGC--------------------------------------------------------
                 *       |
hg-ARHGAP6_f1  -aa--GGGT-GAGAATGT----------------------------------------------------------------
                 *       |
mm-ARHGAP6_f1  -aa--GGGT-GAGAATGT----------------------------------------------------------------
                 A     G|
hg-LRP1B_f0    GCT-gggGT-GAGAGAAG----------------------------------------------------------------
                   R  |
hg-DHX30_f2    cga---GGTAAGGGAGG-----------------------------------------------------------------
                 L       E|
gg-PPP2R2C_f2  gGA---GGTGAGTGGGT-----------------------------------------------------------------
                 L     W|
xt-MBNL1_f0    TTA-tggGT-GAGAGGTA----------------------------------------------------------------
                               *  L  A                  S  P  G  N         E  D  G  Q  Q  R
hg-NT5C2_f1    ------------------gaTTGgctTC------------TcctGGGaat------GAAgacGGCcagCAGaggT
                 *    L  G  L  K  T  A  S  R  G  A  N  C  D  N  *  V  G|
mm-NT5C2_f1    gaC--TGggaTTGaagACGgccAGCagaGGTgctAATtgtGACaacTGAgtgGGAgGTTTGTGTGA-----------

hg-ATF2_f0     ---------------------------------------------------------------------------------

gg-DMTF1_f1    ---------------------------------------------------------------------------------

JGI-49280_f0   ---------------------------------------------------------------------------------

hg-TCERG1_f1   ---------------------------------------------------------------------------------

mm-TCERG1_f1   ---------------------------------------------------------------------------------
```

Figure S9: **Alignment of tetrapod exonized instances to the coelacanth SINE (contd.)**

```
                                 200                                                             150
SINE-RevComp    GCCAATTAGTGCCAATTGTGATGGTGGTTTTGTGATGTGGACACCTGTCCACTAGGAACTGGACTGAGA-CCACC

hg-RORA_f2      -----------------------------------------------------------------------------

mm-RORA_f2      -----------------------------------------------------------------------------
                                     V   E   V   W   G   S   G   L   Q   F   T   K   N   G   T   E    H   T
hg-EEF1B2_f1    --------------------tgGAAgttTGGgaaAGTggaCTCcagTTCactAAGaatGGAactGAGc-atACT

hg-PTDSR_f1     -----------------------------------------------------------------------------

mm-PTDSR_f1     -----------------------------------------------------------------------------

hg-KIAA1409_f2  -----------------------------------------------------------------------------

mm-KIAA1409_f2  -----------------------------------------------------------------------------

gg-KIAA1409_f2  -----------------------------------------------------------------------------

hg-PCBP2_f0     -----------------------------------------------------------------------------

mm-PCBP2_f0     -----------------------------------------------------------------------------

hg-SMARCA4_f0   -----------------------------------------------------------------------------

mm-SMARCA4_f0   -----------------------------------------------------------------------------

hg-FLJ22833_f1  -----------------------------------------------------------------------------

hg-ARHGAP6_f1   -----------------------------------------------------------------------------

mm-ARHGAP6_f1   -----------------------------------------------------------------------------

hg-LRP1B_f0     -----------------------------------------------------------------------------

hg-DHX30_f2     -----------------------------------------------------------------------------

gg-PPP2R2C_f2   -----------------------------------------------------------------------------

xt-MBNL1_f0     -----------------------------------------------------------------------------
                  C   *   L   *   Q   L   S   G   R   F   V   *   *   L   S   V   Q   A   E   L   H   *   N   P
hg-NT5C2_f1     GCtaaTT-GtgaCAActgAG-TgggAGGtttGTGtgaTGAttaTCTgtcCAAgcaGAAttaCACtaaAACcccCC

mm-NT5C2_f1     -----------------------------------------------------------------------------

hg-ATF2_f0      -----------------------------------------------------------------------------

gg-DMTF1_f1     -----------------------------------------------------------------------------

JGI-49280_f0    -----------------------------------------------------------------------------

hg-TCERG1_f1    -----------------------------------------------------------------------------

mm-TCERG1_f1    -----------------------------------------------------------------------------
```

Figure S9: **Alignment of tetrapod exonized instances to the coelacanth SINE (contd.)**

```
                                                                     100
SINE-RevComp    AACTCATTTCACATAGGCCACTGAACAGCCATC---AGATGGTAATGACTTTCGGTCACTACTGACCTGGGCTGG

hg-RORA_f2      --------------------------------------------------------------------------

mm-RORA_f2      --------------------------------------------------------------------------
                  *  *  N  L  K  Q  K  C  S  Y  *  K  S  K  S  |
hg-EEF1B2_f1    taaTGAaatCTCaaaCAGaaaTGTtcaTACtgaAAGagtAAAtcaTAGTGAGTATTG-----------------

hg-PTDSR_f1     --------------------------------------------------------------------------

mm-PTDSR_f1     --------------------------------------------------------------------------

hg-KIAA1409_f2  --------------------------------------------------------------------------

mm-KIAA1409_f2  --------------------------------------------------------------------------

gg-KIAA1409_f2  --------------------------------------------------------------------------

hg-PCBP2_f0     --------------------------------------------------------------------------

mm-PCBP2_f0     --------------------------------------------------------------------------

hg-SMARCA4_f0   --------------------------------------------------------------------------

mm-SMARCA4_f0   --------------------------------------------------------------------------

hg-FLJ22833_f1  --------------------------------------------------------------------------

hg-ARHGAP6_f1   --------------------------------------------------------------------------

mm-ARHGAP6_f1   --------------------------------------------------------------------------

hg-LRP1B_f0     --------------------------------------------------------------------------

hg-DHX30_f2     --------------------------------------------------------------------------

gg-PPP2R2C_f2   --------------------------------------------------------------------------

xt-MBNL1_f0     --------------------------------------------------------------------------
                 P  T  H  F  T  A  A  T  Q  E  P  F  R  D  |
hg-NT5C2_f1     AactCATttcACAgcaGCTaccCAGgagCCAttcAGGgatGGTAATTTAAT-----------------------

mm-NT5C2_f1     --------------------------------------------------------------------------

hg-ATF2_f0      --------------------------------------------------------------------------

gg-DMTF1_f1     --------------------------------------------------------------------------

JGI-49280_f0    --------------------------------------------------------------------------

hg-TCERG1_f1    --------------------------------------------------------------------------

mm-TCERG1_f1    --------------------------------------------------------------------------
```

Figure S9: **Alignment of tetrapod exonized instances to the coelacanth SINE (contd.)**

Figure S10: **Sequence logos for human exonized instances.** The observed splice sites donated by the LF-SINE do not show a distinct pattern beyond the AG and GGT that is common for exons in the human genome. Figures generated using WebLogo (Crooks *et al.*, 2004).

```
LF-SINE  AACAGGTACAGCATGGGCAGCAGCAGTGCAGGCTTCCCCACATTGCCCTGCCAATGTGCCTCAACCCTGCCCTGGAGGGACCACCTCTACAGG
Frame0    N  R  Y  S  M  G  S  S  S  A  G  F  P  T  L  P  C  Q  C  A  S  T  L  P  W  R  D  H  L  Y  R
Frame1     T  G  T  A  W  A  A  A  V  Q  A  S  P  H  C  P  A  N  V  P  Q  P  C  P  G  G  T  T  S  T
Frame2      Q  V  Q  H  G  Q  Q  Q  C  R  L  P  H  I  A  L  P  M  C  L  N  P  A  L  E  G  P  P  L  Q


Blastx of this region adds
--------------------------


>gi|50746727|ref|XP_420628.1| Gene info PREDICTED: similar to PR domain containing 5 [Gallus gallus]
         Length=722

 Score = 41.2 bits (95),  Expect = 0.010
 Identities = 16/30 (53%), Positives = 20/30 (66%), Gaps = 0/30 (0%)
 Frame = +3

Query  3    QVQHGQQQCRLPHIALPMCLNPALEGPPLQ  92
            +VQH  Q C LPH    +CLNP LE PP++
Sbjct  459  KVQHRHQHCELPHSVSSICLNPVLEEPPVE  488



>gi|1172071|sp|P43115|PE2R3_HUMAN Gene info Prostaglandin E2 receptor, EP3 subtype
         (Prostanoid EP3 receptor) (PGE receptor, EP3 subtype) (PGE2-R)  Length=390

 Score = 37.7 bits (86),  Expect = 0.11
 Identities = 15/30 (50%), Positives = 20/30 (66%), Gaps = 0/30 (0%)
 Frame = +1

Query  4    RYSMGSSSAGFPTLPCQCASTLPWRDHLYR  93
            RY   + ++   +LPCQC+STL W DHL R
Sbjct  361  RYHTNNYASSSTSLPCQCSSTLMWSDHLER  390



Blastx to entire LF-SINE adds
-----------------------------


>gi|50731480|ref|XP_417282.1| Gene info PREDICTED: similar to EMSY protein [Gallus gallus] Length=1219

 Score = 48.5 bits (114),  Expect = 7e-05
 Identities = 31/70 (44%), Positives = 40/70 (57%), Gaps = 3/70 (4%)
 Frame = -1

Query  412  PETEVLYLSTEQVQH--GQQQCRLPHIALPMCLNPALEGPPLQG*EGSSVSMPIQSLASL  239
            P T+   +T++ ++  G      P  A PMCLNP L+G PL G  G SVSM IQ LASL
Sbjct  270  PSTQTTNTTTQKNRYSTGSGSASFPQTAPPMCLNPDLKG-PLLGMRGDSVSMLIQFLASL  328

Query  238  LRLPTRVPIS  209
            + + T  P S
Sbjct  329  VIIVTTSPSS  338
```

Figure S11: **Additional Blastx exon hits to the LF-SINE.** The stop-codon free region between splice sites 381/289 is shown (top), along with two novel distant exon hits obtained by Blastx of this region (middle), and an additional, strong hit found by Blastx of the entire LF-SINE to a predicted exon which is out of the golden path in the current chicken genome draft and was thus excluded from our analysis (bottom).
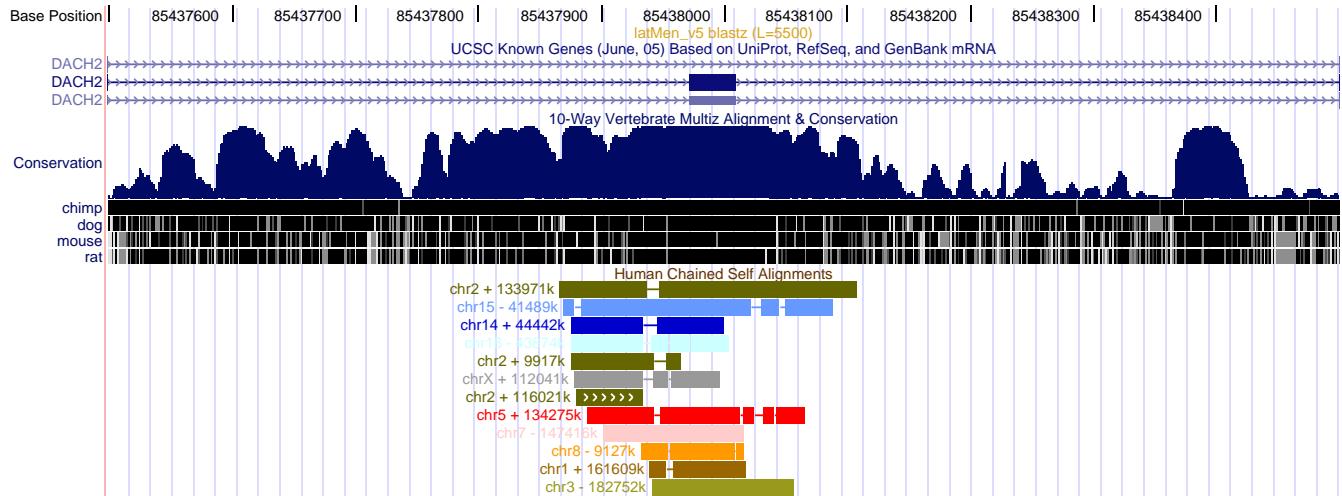
Figure S12: **An exonic candidate for an exaptation event, not related to the LF-SINE.** The second exon of DACH2 displays the characteristics of an exaptation event: it is an alternatively-spliced exon, not found in DACH1. The exon size is divisible by three, and forms part of the ORF in one isoform, is part of the 5' UTR in another, and is skipped in a third known gene isoform. It is found only in mammals, has a weak pairwise similarity to a mammalian-specific interspersed repeat (MIR) element, as well as to additional intronic and intergenic short paralogs in the human genome. A preliminary screen suggests the existence of dozens of exons with similar properties.

```
                        100        110        120        130        140
LF-SINE       GAAAGTCA-TTACCATCTGATGGCTGTTCAGTGGCCTA---TGTGAAATGAG
              : ::::::: ::: : :::       ::   ::::: :::     ::: ::::: :
Core Domain   GCAAGTCACTTAACCTCTCTGAGCC--TCAGTTTCCTCATCTGTAAAATGGG
                        100        110        120        130        140


                        90        100        110        120
V-SINE(shark) CTGTGTGGAGTTTGCACATTCTC----CCCGTGTCTGCGTGGGTTTCCTC
              :::::::    :: :   : :: :    ::  : :  :: :  :::::::::
Core Domain   CTGTGTGACCTTGGGCAAGTCACTTAACCTCTCTGAGCCTCAGTTTCCTC
                        90        100        110        120        130
```
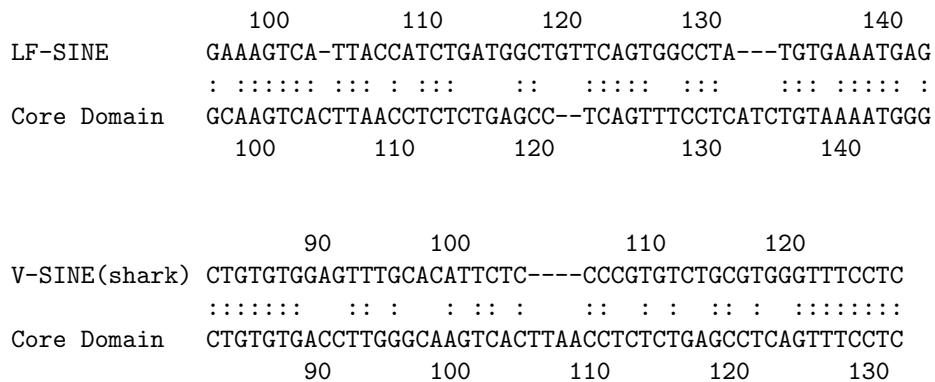
Figure S13: **The CORE domain signature of SINEs.** Both our LF-SINE and the V-SINE (Ogiwara *et al.*, 2002) show similarity to a section of the proposed "core domain" of the CORE-SINE superfamily (Gilbert and Labuda, 1999), although the LF-SINE and the V-SINE themselves do not align well. These similarities, 68% and 64% identity over 50 bases for LF-SINE and V-SINE, respectively, computed as in (Gilbert and Labuda, 1999), are in par with similarities observed in other vertebrate SINE families, classified into the CORE-SINE superfamily on this basis. We find this level of similarity by itself insufficient to determine common descent.

27

```
LF-SINE       GGGGACTGGATGGCTCAGTGGAATTGGTAATGGG..ATATGGAG.CCTTT....C.A.CCTCTAGGTCACTGG..GTTCAAATC.
LF-SINE2      GGGGACTGGATGGCTCAGGGGAATTGGTCATGGG..ATATAGAG.CCTTT....C.A.CCTCTAGGTCACTGG..GTTCAAATC.
Coelacanth0   .GGGGACTGGATGGCTCAGGGGAATTGGTCATGGG..ATATAGAG.CCTTT....C.A.CCTCTAGGTCACTGG..GTTTGAATC.
Coelacanth1   ..GGACTAGATGGCTCAGCGGAATTGGTCATGGG..ATAAAGAG.CCTTT....C.A.CCTCTGGGCACTGG..GTTTGAATC.
Coelacanth2   .GAAACTGGATGGCTCAGCGGAATTGGTCATGGG..ATACAGAG.CCTTT....C.A.CCTCTAGGCCGCTGG..GTTTGAATC.
Coelacanth3   ..GGACTGGAGCACTCAGCGGAATTGCTCATGGCGGATATAAAG.CCTTT....C.A.CCTCTAGATCACTGG..GTTTGAATC.
Coelacanth4   ................................ATGGG..ATACAGAG.CCTTT....C.A.CCTCAAGGTCGCTGG..GTTCAAATC.
Coelacanth5   ..........TCAAGGGAATTGGTAATGGG..ATATAGAT.CCTTT....C.A.CCTCTAGGTCACTAG..GTTAAAATC.
Coelacanth6   .GGGGATCGGGTGGCTCAGAGGAAATGTCATTGGG..ATATGGAG.CCTTT....C.A.TTTTGAGGTCACTGG..GTTTGAATC.
Coelacanth7   GGGGACAGGATAGCTCAGCGGAATTGGTCAGAGG..ATATAGAG.CCTTT....C.A.CCTCTAAGTCACTGGCTGATCAAATC.
Coelacanth8   GGGGACTGGATGGCTATGGGGAATTGGTCATGGG..ATACAGAG.CCTTT....C.A.GTTCTAGGTCACTGG..GTTGCAATCA
Coelacanth9   ..GGATTGGATGACTCAGTGGAATTGGCACTGGG..ATTGGAAC.CCTTT....C.A.GCTCTAGGTCACTGG..GTTTGAATC.
Coelacanth10  .GGGGACTGGATGGTTTAGGCGAAATTGGTCATGGG..ATACAGAG.CCTTT....C.A.CCTCTAGGTCA.....
Coelacanth11  GAAAACTGGATGGCTCAGCGGAATTGGTCCTGGG..ATACAGAG.CCCCT....T.A.CTTCTAGGTTACTGT..GTTCAAATC.
Coelacanth12  GGGAACTGGATGGCTAATGGGAGCTGTCACTGGG..ATATGGAG.CCTTT....C.A.CCTCTAAGACACTGG..GTTCAAATT.
Coelacanth13  GGGGACCAGATAGGTCAGTGGA.TTGGTACTGGG..ATACGGAG.CCTTT....C.A.TCTCCAAGTCACTGG..ATTTGAACT.
Coelacanth14  GAGGACTGAATGGCTCAGTAGG.TTGGAATTGGG..ATATAGAG.CCTTT....T.A.TATCTAGGTGATTGG..TTTAAATC.
Coelacanth15  GGAAACTAGATAGCTCCGTGGAATTGGCCAATGGG..ACACAAAT.CCTTT....C.A.CCTCCAGATCACTGG..GTTTGAATC.
Coelacanth16  ...............................CCTTT....C.A.C.TCTATATCACTGG..GTTTGAATC.
Coelacanth17  ...GATGGGATGGATCAGTGGA.TTGGCACTGGG..ATACAGAG.CCTTT....C.A.CG...........A..ATTTGAATC.
Coelacanth18  GGGGATGGGATGGCTCAGTGGAATTGGTAGTAGG..ATATAAAG.CCTTT....C.A.CCTCCAGGCCACTGG..GTTCAAATC.
Coelacanth19  .GGCTACTGGATGGCCCAGTGGA.TTGGCAATGAG..ATACAGAG.CTTTTGGATC.A.GCTCTAGGTCA.TGG..GTTAAAATC.
Coelacanth20  ..........................................................................
Coelacanth21  ..........................................................................
Coelacanth22  GGGGACTTGATGGCTCAGTGGAATTGAGGGTGAG..ACAC..AG.TTTTT....C.A.CCTTTAAGTCAATGG..GTTCAAAAC.
Coelacanth23  .GAGAGTAGGTGACTCTGTGGA.CTGGCATTGGG..GTATGAAT.CTATT....C.A.CCTCTATGTCGCTGG..TTCAAGTT.
Coelacanth24  .....................TTTGGCAGTAGG..AT..GGCA.CCTTT....T.A.CCGCTAGGTTACCAG..ATTCAAATC.
Coelacanth25  GAGGACTGGATGGCTCAGTAAAAATTGGCAATGGG..TTACAGAA.CGCTT....C.A.GTTCTAGATCACTGG..GTTCAAATT.
Coelacanth26  .GAGACTGACTGACTCAGAGGA.CTGATATTGGA..ATATGAAA.CCTTT....C.A.GTTCTAGGTCA.TGG..ATTCAGATA.
Coelacanth27  ..........................................................................
Coelacanth28  ..............CAGGGGAATGGCAATTAAG..ATAAAATAAACCTTT....C.A.CCTCTAAGTTACTGG..TTCAAAAC.
Coelacanth29  ..................................................AGATCATGGG..AGGAGAATC.
Coelacanth30  ....................................................................GTTCAAAGT.
Coelacanth31  ..........................................................................
Coelacanth32  ..GAAGTGGATGGTTGTGTAGCCT.........GG..TGATGGAG.TCTCT....C.AACTTTTGGGTCACTGA..TTCATATC.
Coelacanth33  ..........................................................................
Coelacanth34  .GGGGACTAT.TGATTCAGTAGA.CTGGTATTGGG..ATATGGAG.TTCCTT....T.G...TCATGGCCACCGA..TTTGAACA.
Coelacanth35  ...........................................................CTCA..GTTAACATG.
Coelacanth36  ..........................................................................
Coelacanth37  .....................CAG..ATACAGAA.CCTTA....C.A.CCACT.........G..ACTCAAATC.
Coelacanth38  ..........................................................................
Coelacanth39  ..........................................................................
Coelacanth40  GAAGACTGAATGGATCAG.GGGACTGGTACTGGG..ATATGGAG.CTTTT....C.A.AGT.GAAGTGACTGG..ATTTGAACC.
Coelacanth41  .GGGGACTGGAGGGCCCAGCGGAATTGGTCATGAG..AAACAGAG.CCTTT....C.A.CCTCTAGGTCACTGG..GTTTGAATC.
Coelacanth42  ..........................................................................
Coelacanth43  .GGGATTGGTTGGCAC.CGGGATTGCTCACTGGG..ATATGGAG.CCTTT....CCA.CTTTTATGTCACTGT..CTTCAGATC.
Coelacanth44  .......GGATGGCTCACTTGGGTTGGTACTGGG..ATATGCCA.CCTTT....A.A.GCTCTAGGTCACTGA..TTTGAAGCC.
Coelacanth45  ..........................................................................
Coelacanth46  ..........................................................................
Coelacanth47  .........................................TTT....C.A.TTTCTAGGGTATTGG..GT..GACTG.
Coelacanth48  ..........................................................................
Coelacanth49  ..........................................................................
Coelacanth50  ..........................................................................
Coelacanth51  .......................................C.TTTTT....A.A.TTTTTGAGTACTGG..GTTAAAATA.
Coelacanth52  ..........................................................................
Coelacanth53  ..........................................................................
Coelacanth54  .......................................C.TTTTT....A.A.TTTTCTGAGTACCGG..GTTAAAATA.
Coelacanth55  ..........................................................................
Coelacanth56  .....................................TTT....T.A.CTTTT..GTGGCTTG..ATTTGAGTC.
Coelacanth57  ..........................................................................
Coelacanth58  ..................ATGCAGAT.CCATC....C.A...TCCAAGTCGCTGG..CTT.GAGTC.
OutGroup      GGGGACTGGATGGCTCAG.GGAATTGGTCATGGG..ATATAGAG.CCTTT....C.A.CCTCTAGGTCACTGG..TTCAAATC.
PCBP2-Human   ..........................................................................
PCBP2-Chimp   ..........................................................................
PCBP2-Dog     ..........................................................................
PCBP2-Mouse   ..........................................................................
PCBP2-Rat     ..........................................................................
PCBP2-Opossum ..........................................................................
ISL1-Human    ..........................................................................
ISL1-Chimp    ..........................................................................
ISL1-Dog      ..........................................................................
ISL1-Mouse    ..........................................................................
ISL1-Rat      ..........................................................................
ISL1-Opossum  ..........................................................................
ISL1-Chicken  ..........................................................................
ISL1-Frog     ..........................................................................
DACH1-Human   GGGAATTGGACGGTTCAG.GGAGTTGCAAGTAGG..ATAAAGAG.CAGTT....C.T.CCTCTAGGTCACTAG...TTCAAACT.
DACH1-Chimp   GGGAATTGGACGGTTCAG.GGAGTTGCAAGTAGG..ATAAAGAG.CAGTT....C.T.CCTCTAGGTCACTAG...TTCAAACT.
DACH1-Dog     GGGAATCGGACGGTTCAG.GGAGTTGCAAGCCGG..ATAAAGAG.CAGTT....C.T.CCTCTAGGTCACTAG...TTCAAACT.
DACH1-Mouse   GGGAATCGGACGGTTCAG.GGAGTTGCAAGTGGG..ATAAAGAG.CAGTT....C...TCTCTAGGTCACTAG...TTCAAACT.
DACH1-Rat     GGGAATTGGACGGTTCAG.GGAGTTAGAAGTGGG..ATAAAGAG.CAGTT....C...TCTCTAGGTCACTAG...TTCAAACT.
DACH1-Opossum GGGAACTGGATGGTTCAG.GGAGTTGCGGAGTCAG..GGAGAGAG.CATTT....C.T.CTTCTAGGTCACTAG...TTCAAACT.
```

Figure S14: **LF-SINE reconstruction.** The LF-SINE sequence is a reconstruction of 59 repeats in the Indonesian coelacanth that contain a region similar to the the ultraconserved exon in PCBP2. The LF-SINE sequence was used to locate similar instances in tetrapods. We reconstruct the LF-SINE a second time using all coelacanth sequences as well as outgroup instances from available tetrapods related to ISL1 and PCBP2 (discussed in the main text), and a conserved human instance that spans the LF-SINE length, near DACH1. This second reconstruction is labeled LF-SINE2. Only 15 bases differ between it and the first reconstruction, a testimony to the remarkable conservation of the LF-SINE during vertebrate evolution.

```
LF-SINE        .CA.GCCC.AGGT....CAG.T.AGTGACCGAAAG....TCATTACCATCTG....ATGGCT.GTT.......CAGTGG....
LF-SINE2       .CG.GCTC.AGGT....CGG.T.AGTGACCGAAAG....TCAGTACCATCTG....ACAGCT.GTT.......CAGTGG....
Coelacanth0    .CG.GCTC.AGGT....TGG.T.AGTGGTCTAAAG....TCAGTACCACTG.....ACAGCT.GTT.......CAGTGG....
Coelacanth1    .TC.GCTC.AGGT....TGG.T.AGTGACCGAAAG....TTAGTACCATCTG....ACAGCC.GTT.......CAGTGA....
Coelacanth2    .CA.GCTC.AGGT....CAG.T.AGTGACCGAAAG....TTAGTACCATCTG....ATAGCT.GTT.......CAGTGG....
Coelacanth3    .CG.GCTC.AGGT....CGG.T.AGTGACCGAAAG....TTGGTACCATCTG....ACAGCC.ATT.......CAGTGG....
Coelacanth4    .TG.GCTC.AGGT....CGG.C.AGTGACCGAAAG....TCACTACCATCTG....AAAGCC.GTT.......CAGTGG....
Coelacanth5    .AG.GCTC.AGGT....TCA.T.AGTGATCAAAAG....TTGGTACCATCTA....CTGGGT.GTT.......CAGTGG....
Coelacanth6    .CA.GCCC.GGGT....CAG.T.AATGAC..AAAG....TAGTTACCACCTG....TTGGTT.GTT.......CTGTGA....
Coelacanth7    .TG.GTTC.AGGT....CAG.T.AGTGATCGAAAG....TCAGTACCATCTG....AAGGCC.GTT.......CAGTGG....
Coelacanth8    GCT.CCTC.AAGT....CGG.T.AACCAGAGAAAG....TGGGTACCATCTG....ACTGCC.ATT.......CAGTG.....
Coelacanth9    .AA.GCCC.AGGT....CAG.T.TCTGATTGAAAC....TTGTTATCATAAG....ATAGCT.ATT.......CAGTGG....
Coelacanth10   ................A.T.AGTGACCAAAAC....TCGGTACCATCTG....AGAGCT.GTT.......CAGTGG....
Coelacanth11   .CA.GCTC.AGGT....CAG.C.AGTGACCAAAAG....TCAGTACCATCTG....ACAACC.ATT.......TAGTGG....
Coelacanth12   .CA.GCCC.AAGT....CTG.T.AGTGATGGAGAG....TCATTGCCATGTG....ATGGCT.GTT.......TAGTGA....
Coelacanth13   .CA.GTCC.AGGT....TGG.T.AGTGACCAAAAG....CTGTTACAACTG.....ATGGCT.GTT.......CAGTGG....
Coelacanth14   .CA.GATA.AGGT....TGG.T.AATGACCCAAAG....TCATTACCATCTG....ATGGTT.GCT.......CGATGG....
Coelacanth15   .C......C.AGTT....TGA.C.AGTGACCAAAAG....TTGTTCCCATCTG....ATGGCT.GTT.......TGGTGG....
Coelacanth16   .CG.GCTC.AGGT....GGG.T.AGTGACCCAAAG....TTGGTACCATCTG....ACAGCC.ATT.......CAGTGG....
Coelacanth17   .CA.GCCT.AAGT....TGG.T.TGTGGTTGAAAG....TTATCACC...TG....GCAGCT.ATT.......CTGTGG....
Coelacanth18   .TA.GCTC.AGGT....CAG.T.AGTTCCTGAAAG....TCTTTACCATCTG....ACACCT.GTT.......CTGTGG....
Coelacanth19   .CA.GCCC.AGGT....CAG.T.AGTGACTGAAAA....TTGTTACCACCTG....GCAAGT.ATT.......TGGTGG....
Coelacanth20   ..................................TA....GTAGCT.GTTAAAACACCAGATA......
Coelacanth21   .............................................................................
Coelacanth22   .CA.ATTC.AGGT....CAG.TTAGTGATCAAAAG....TCACTTCCGTCCA....ACAGCT.GTT.......CATTGG....
Coelacanth23   ..G.GCAC.AGCTGCTACGG.T.AGCGATAAAAAG....CCATTACTACTTG....ACCACT.GTT.......TAGTGG....
Coelacanth24   .TA.GCTC.AGGC....TGT.T.AATGACCAAAAG....TTGGTACTTTCTG....ATAGCTTATA.......AAGTGG....
Coelacanth25   .CATGTGC.AGGT....TGG.T.AGTGACCAAAAG....TTGTTGC.ATCTG....ATTGCA.GTT.......TGGTGA....
Coelacanth26   .TG.GCTC.ATGT....CAC.T.AGGAATTGGAAA....TTATGACC............TT.GTT.......TGGTGG....
Coelacanth27   ...............................................T.GTT.......CAATGG....
Coelacanth28   .TG.ATCT.AGGT....TGGGT.AGTGACTG........TTACAATCTG....ATGATT.GTT.......CAGTGA....
Coelacanth29   .TA.CCCC.ATGT....TGT.T.TGTGACTGAAAG....TTATCATCACCTG....ATGGCT.TTT.......CAATCA....
Coelacanth30   .GG.GCCT.AGGT....CAG.T.AGCAACTGAGAG....TTGTTGCCATCTG....CTGGAT.GTT.......TCCTGA....
Coelacanth31   ......................................GTCT.TAT.......TGGCGG....
Coelacanth32   .CA.TCCCCAGGT....CAG.T.AGTGATTGAAAA....TCATTACCATGCG....ATGTCT.GCT.......TACGAG....
Coelacanth33   ...........................TCATTATATG....ATGGGC.GTT.......TGGTGG....
Coelacanth34   .TG.GTCC.AGGT....CAG.T.ACTGATCAAAAG....TCATTAC...TT....GTAGCT.GCT.......TCG.GG....
Coelacanth35   .AA.CCCC.A.......CTG.T.AGGGACTACAAA....TGATCACCACCTG....ATGTGA.TTT.......CATT....
Coelacanth36   .............................................................................
Coelacanth37   .TG.GTCC.AGTT....TGG.T.AGTGCCTGAAAG....TTGTTACCATCAG....AAAGCT.AGC.......TGGTGG....
Coelacanth38   ...........................TCATTTCCACTAA....GCAACT.TTT.......CAATGG....
Coelacanth39   .............................................................................
Coelacanth40   .CA.GTCC.ATGT....CCA.T.AGGGACTAAAAT....TCATTATCATACG....ATGGT..........GGTGT....
Coelacanth41   .TA.GCTC.AGGT....TGG.T.AGTGACCAAAAG....TTGGTACCATTTG....ACAGCC.GTT.......CAGTGG....
Coelacanth42   .............................................................................
Coelacanth43   .CA.GCCC.AGGT....TGG.T.AGTGGCCAAAAA....TCATTACCACCTG....GTGGCT.GTT.......CTGTGG....
Coelacanth44   .TC.T...........T.AGTTGTAGAAAG....TCATTACCATCTA....ATAAGG.GTT.......TGATGA....
Coelacanth45   .............................................................................
Coelacanth46   ...........................CAGCACCATTTGTAGAATGGTT.GTA.......TTTCAG......
Coelacanth47   .CA.GTCG.TGGT....CTG...AATCACCAACAA....TCATT..CAGATG....AGGCAT.CTG.......CAATCA....AT
Coelacanth48   .............................................................................
Coelacanth49   .............................................................................
Coelacanth50   .............................................................................
Coelacanth51   .TA.GCCC.CAGC....TA..C.AGTGACAAAAAAAA....TCATTCAGGTGGG....AAACCT.GCA.......TTGAGG...T..
Coelacanth52   .............................................................................
Coelacanth53   .............................................................................
Coelacanth54   .TA.GCCC.CAGC....TTG...AGTAACAAAAAAAAAATCATTCAGGTGGG....AAACCT.GCA.......TTGAGG....
Coelacanth55   .....CC.ACGT....CAG.T.GGTGACCTAAAG....TCAT..CATTTG....ACAGAT.GCC.......TCCC......
Coelacanth56   .CA.TCCC.CAGC....CTG...AGTAACTGAAAA....TCAGCCCAGTGAG....AAAACT.ATC.......AACAGG....
Coelacanth57   .............................................................................
Coelacanth58   .TG.ACCC.TGGT....CAG.T.AGTGACTGAAGG....GCATTGAAATCTG....ACATCT.GTT.......CAGTGA....
OutGroup       .CG.GCTC.AGGT....CGG.T.AGTGACCGAAAG....TCATTACCATCTG....ATGGCT.GTT.......CAGTGG....
PCBP2-Human    ...................................TTACCATCTG....GTGG...GCT.......ACGCGG....
PCBP2-Chimp    ...................................TTACCATCTG....GTGG...GCT.......ACGCGG....
PCBP2-Dog      ...................................TTACCATCTG....GTGG...GCT.......ACGCGG....
PCBP2-Mouse    ...................................TTACCATCTG....GTGG...GCT.......ACGCGG....
PCBP2-Rat      ...................................TTACCATCTG....GTGG...GCT.......ACGCGG....
PCBP2-Opossum  ...................................TTACCATCTG....ATGG...GCT.......ACGCGG....
ISL1-Human     .............................................................................
ISL1-Chimp     .............................................................................
ISL1-Dog       ...................................................................GA......
ISL1-Mouse     .............................................................................
ISL1-Rat       .............................................................................
ISL1-Opossum   .............................................................................
ISL1-Chicken   ...........................CTAAAG....TTCTGGTGTTCAT....TTGGGT.GTT.......CCTTGGAAA...
ISL1-Frog      .............................................................................
DACH1-Human    .TC.GCCC.ATAT....TGG.T.AATGACTGACAG....TCATTAATATATG....ATGGCT.GTT.......TGATGG....
DACH1-Chimp    .TC.GCCC.ATAT....TGG.T.GATGACTGACAG....TCATTAATATATG....ATGGC...............
DACH1-Dog      .TG.GCCC.GTAT....TGG.T.AATGACGGACAG....TCATTAATATGTG....ATGGCT.GTT.......TGATGG....
DACH1-Mouse    .TC.GCCA.ATAT....TGG.T.AATGACTGACAG....TCATTAATATATG....ATGGCT.GTT.......TGATGG....
DACH1-Rat      .TC.GCCA.ATAT....CGG.T.AATGACTGACAG....TCATTAATATATG....ATGGCT.GTT.......TGATGG....
DACH1-Opossum  .TA.GCCG.ATAT....CGG.T.AATGACTGACAG....TCATTAATATATG....ATGGCT.GTT.......TGATGG....
```

Figure S14: **LF-SINE reconstruction (contd).**

```
LF-SINE       .C.C.TATGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....GTGG
LF-SINE2      .C.C.TATGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....GTGG
Coelacanth0   .C.C.TATATGAAATG.A......G.T......T............GGT......C.G.TCTCACTCCAGTTCCT.A....GTGG
Coelacanth1   .C.C.CACGTGAAATG.A......T.T......T............GGT......G.G.TATCAGTCCAGTTCCT.A....GTGG
Coelacanth2   .C.C.CATGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGTCCA.............G
Coelacanth3   .C.C.AATATGAAATG.A......G.T......T............GGT......C.G.TCTCTGTGCAGTTCCT.A....GTGG
Coelacanth4   .C.C.TATGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCA.A....GTGG
Coelacanth5   .C.CCTATGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGTCCAATTACT.A....GTGG
Coelacanth6   .CAC.TATATGAAATG.A......G.T......T............GGT......C.G.TTTCAGTCCATTTCCT.A....GGGG
Coelacanth7   .C.C.TATGCGGAACG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....GTAG
Coelacanth8   ..........AAATG.A......A.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....TTGG
Coelacanth9   .C.C.TATGTGAAATG.A......G.T......T............................CCT.A....TTGG
Coelacanth10  .C.C.TATATGAAATG.A......G.T......T............GGT......C.A.TCTCAGTTCAGTTCCT.A....GTGG
Coelacanth11  .C.C.TATGTGAAATA.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....GTGG
Coelacanth12  .T.C.TGTGTGAAATG.A......A.T......T............GGT......C.G.TCTCAGTTCAGTTCCT.A....GTGG
Coelacanth13  .T.C.TATGTGAAATA.A......G.T......T............GGT......C.C.TCTCAGCCCAGTT........GCAC
Coelacanth14  .C.C.AATGTGAAATG.A......G.C......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....GTGG
Coelacanth15  .T.C.TGTGTAAAATG.A......A.T......T............GGT......C.G.TTTCGTTCCATTCCT.A....GTGG
Coelacanth16  .C.C.AATATGAAATG.A......G.T......T............GGT......C.G.TCTCTGTGCAGTTCCT.A....GTGG
Coelacanth17  .C.C.TATGTAAAATG.A......G.T......T............GAT......C.G.TTTCAATCCATTCCT.A....GTGG
Coelacanth18  .C.T.TATGTGAAATG.A......G.T......T............ACTA.....C.G.TCCCAGTCTAATTCCT.A....GTGG
Coelacanth19  .C.T.TAAGTGAAATG.A......G.T......T............GGT......A.G.TCTCAGTCCAGCTCCTAA.A....GTGG
Coelacanth20  .C.C.TATGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....GTGG
Coelacanth21  ..........TG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....GTGG
Coelacanth22  .T.C.CATGTGACATG.A......C.T......T............GGC......C...TCTCAGTCCAGTTCCT.A....GTAG
Coelacanth23  .C.C.TATGTGAAATG.G......G.T......T............GGT......C.G.TCTTGGTTCAGTTCCC.A....GTGG
Coelacanth24  .C.C.TGTGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....ATAG
Coelacanth25  .C.C.AGTGTGAAACA.A......G.T......T............GGT......C.GACCTCAGCCCAGCTTTT.A....CGAG
Coelacanth26  .C.T.GGTGTGAAATG.A......G.T......T............GTT......C.G.CATCAGTCTAGTTTCT.C....ATGG
Coelacanth27  .C.C.TATGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGCCCAATTCTT.A....GTGG
Coelacanth28  .C.C.TATGTGAAATG.T......G.T......T............GGT......C.G.TTTCAGTCTAGTTCCT.A....GTGG
Coelacanth29  .C.C.TATGTGAAATA.A......G.T......T............GAT......T.T.TCTCAGTCCAGTTCCT.A....GTCA
Coelacanth30  .C.C.T....GAAATG.A......G.T......T............CCT......C.G.CCCTAGACCAATTACT.A....GTGG
Coelacanth31  .C.C.TATGTGAAATG.G......G.C......T............GGC......C.G.TCATAATCCAGTTTCA.T....GTGG
Coelacanth32  .AAT.TATCTGAAATA.T......A.T......T............GGT......T.G.TCTTAATCCACTTTCT.G....TTAG
Coelacanth33  .C.C.TATGTGCAAAG.A......A.T......T............GGT......C.G.TC......CAGTTCCG.A....GGGG
Coelacanth34  .T.C.TACGTAGAATG.T......G.C......T............GGT......C.G.GCTCAGTCTAGGTTCT.A....GTGG
Coelacanth35  ...C.TATTTAAAATG.A......G.T......T............GGT......C.T.TCTCAGTTCAATTCCT.A....GTGG
Coelacanth36  ...........................................................................
Coelacanth37  .C.C.TAT......TA.A......G.T......T............GCT......C.A.CTGCAGTCCAGTTCTA.A....GTGG
Coelacanth38  .T.C.AATGTGAAATG.A......G.C......T............GGT......C.A.TCTGAATCCAGTTCC.......GTGG
Coelacanth39  .....ATAAGAAATC.A......G.T......T............GGT......C.G.TCTCAGTCCACTTCTT.A....TTAG
Coelacanth40  .C.C.TCTGTGAAATG.A......G.T......T............GGT......C.G.TCTCAATCCAGTTCCT.G....ATGG
Coelacanth41  .T.T.TATGTGAAATA.A......G.T......T............GGT......CAC.TTTCAGTCCAATTCCT.A....GTGG
Coelacanth42  ...........................................................................
Coelacanth43  .C.C.TATGTGAAATG.A......G.C......T............T......C.G.TCCCAGTCCACTTCCT.A.........
Coelacanth44  .C.C.TACGTGAAGTG.A......G.C......T............TGT......C.G.TTTCAGTACAGCTACT.A....GTAG
Coelacanth45  ...........................................................................
Coelacanth46  .C.C.TAAGTGAAAAC.AC......T.T......T............TTT......C.G.TTTCAGTGCAATTCTT.A....TTAG
Coelacanth47  TC.A.AAGTGATATTTA......T.T......T............GGA......C.G.TCTAAGCTCAGCTACT.T....TTGA
Coelacanth48  .........GAAACG.G......G.T......T............CAA......C.T.AGCCAGTCCAGTTCTT.AAATTGTGG
Coelacanth49  ...........................................................................
Coelacanth50  ...........................................................................
Coelacanth51  .C.C.AAGTGAAATT.T......C.T......T............TACA.....C.G.TCACAGCTCAACTAGT.T....TTCA
Coelacanth52  ...........................................................................
Coelacanth53  ...........................................................................
Coelacanth54  .T.C.CA....AAGTG.A......A.T......T............TTTTTTACAC.G.TCACAGCTCAACTAGT.T....TTCA
Coelacanth55  .C.C.TATGTAAAATG.G......A.T......T.CT.........GAC......C.G.TATCAATCTAGCTCAT.........
Coelacanth56  .TAT.TAAGTGAAGTG.A......G.T......T..T.........GGT......C.G.TCTCAGCTTCATTTAT.T....GTTG
Coelacanth57  ....TGTGTGAAATA.A......G.T......T............GGT......C.A.TTTCAGTCCATTTCCT.A.........
Coelacanth58  .A.A.T......GATG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTACCT.G....CTAG
OutGroup      .C.C.TATGTGAAATG.A......G.T......T............GGT......C.G.TCTCAGTCCAGTTCCT.A....GTGG
PCBP2-Human   .C.C.TGGGTGAAATG.A......GGT......T............GGG......C.G.GGTCAGTCCAGGTCCC.C....GTGG
PCBP2-Chimp   .C.C.TGGGTGAAATG.A......GGT......T............GGG......C.G.GGTCAGTCCAGGTCCC.C....GTGG
PCBP2-Dog     .C.C.TGTGTGAAATG.A......T.G......T............GGG......C.G.GGTCAGTCCAGGTCCC.C....GTGG
PCBP2-Mouse   .C.C.TGTGTGAAATG.A......G.G......T............GGG......C.A.GGTCAGTCCAGGTCCC.C....GTGG
PCBP2-Rat     .C.C.TCTGTGAAATG.A......G.G......T............GGG......C.A.GGTCAGTCCAGGTCCC.C....GTGG
PCBP2-Opossum .C.C.TGTGTGAAATG.A......G.G......T...TGGGGGAGGGG......C.G.GGTCAGTCCAGGTCCC.C....GTGG
ISL1-Human    ...................T......T............GGC......A.A.TCTCAGCTCAGTCCCC.A....GTAA
ISL1-Chimp    ...................T......T............GGC......A.A.TCTCAGCTCAGTCCCC.A....GTAA
ISL1-Dog      .C.T.TAAATGAATGG.T......G.T......T............GGT......C.A.TCTCAGCTCAGTCCCC.A....GTAA
ISL1-Mouse    ..........GAAATG.A.CGGCGT.T......T............GCT......C.C.TCTCATCTCAGTCCCC.A....GTAA
ISL1-Rat      ..............................................TCTCATCTCAGTCCCC.A....GTAA
ISL1-Opossum  ..........TGAAATG.A......A.TGGCGTTT..........GGT......C.A.TCTCAGCTCAGTTCCC.A....GT..
ISL1-Chicken  .C.G.TAAATGAGTAG.T......G.T......TT............GGT......C.A.TCTCAGTTCAGTTTCT.A....GTAA
ISL1-Frog     ..............................................GT......C.G.TCTGAGCAGAGTTGCT.A....GT..
DACH1-Human   .C.C.TATATGAG.....................................................................
DACH1-Chimp   ...C.TATATGAG.....................................................................
DACH1-Dog     .C.C.TC.....GATG.A......G.T......................................................
DACH1-Mouse   .C.C.TATATGAG.....................................................................
DACH1-Rat     .C.C.TATATGA......................................................................
DACH1-Opossum .C.C.TA.....GATG.A......G.T......................................................
```

Figure S14: **LF-SINE reconstruction (contd).**

```
LF-SINE        A.........CAGG...T....G...TCC...AC.ATCAC.A..AAA.C..CA.CCAT.CA.......C..AA.TTGGC....
LF-SINE2       A.........CAGG...T....G...TCC...AC.ATCAC.A..AAAAC.CA.CCAT.CA.......C..AA.TTGGC....
Coelacanth0    A.........CAGG...T....G...TCC...AC.ATCAC.A..AAA.C..CA.CCAT.CA.......C..AA.CTGGC....
Coelacanth1    A.........CAGG...T....G...TCC...AC.ATCAT.G.A....A..CA.CCAT.CA.......C..AA.TTGGC....
Coelacanth2    G.........CAGG...T....G...TCC...AC.ATCAC.A..AAAAC.CA.CTGT.CA.......C..AA.TTGGT....
Coelacanth3    A.........CAGG...T....G...TCC...AC.ATCAC.A..AAAAA..CA.CCAT.CA.......C..AA.TTAGC....
Coelacanth4    A.........CAGG...T....A...TCC...AC.ATCAC...............................................
Coelacanth5    A.........CAAC...G....G...TAC...AC.ATCAC.A..AAA.C..CA.CCAT.C.......................
Coelacanth6    A.........CAGG...T....G...TCC...AT.ATTGC.A..AAA.C..CA.CCAT.CA.......C..AG.TTGGC....
Coelacanth7    A.........CGAC...A....G...TCC...AC.ATCAC.A..AAAAC.CA.CCAT.CA.......C..AA.............
Coelacanth8    A.........CAAC...T....G...TCT...AC.ATCAC.A..AAA.A..AA.AAAT.CA.......C..AA.TTAGC....
Coelacanth9    G.........CAGG...T....G...TCC...AC.ATCAC.A..AAA.C..CA.CCAT.CA.......C..AA.TTGGC....
Coelacanth10   A.........CAAC...T....G...TCC...AC.ATCAC.A..AAAAT..CA.CCAT.CA.......C..AA.TTGGC....
Coelacanth11   A.........CAGGTAGA...T....G...TCC...AC.ATCAC.............................................
Coelacanth12   A.........TGGG...T....G...TCC...TC.ATCAT.T..AAA.C..CG.TCAT.CA.......T..AA.TTATC....
Coelacanth13   A.........TAGA...T....G...TCC...AT.GTCGC.A..AAC.C..CA.CCAT.CA.......T..AA.TTGAC....
Coelacanth14   A.........T..G...T....G...TCC...AC.ATCAC.A..AAA.C..CA.CCAT.CA.......C..AA.TTGGC....
Coelacanth15   A.........CAGG...T....G...TCC...AC.ATCAT.A..AAA.T..CA.CCAT.CA.......C..AG.TTAGC....
Coelacanth16   A.........CAAC...G....G...TAC...AC.ATCAC.A..AAA.G..CA.CCAT.C......................
Coelacanth17   A.........CAGG...T....G...TCC...AC.ATCAC.A..A..................................TTAAC
Coelacanth18   A.........CAGG...T....G...TCC...AA.ACCAC.A..AAA.C..CT.CCAT.CA.......C..AA.CTAGC....
Coelacanth19   A.........CAGG...T....T...CCC...AC.GTGAC..........................................
Coelacanth20   A.........GTGG...T....G...TCC...AC.ATCGC.A..AAA.C..TA.CCAT.CA.......T..AA.TTGGC....
Coelacanth21   A.........CAGG...T....G...TCC...AC.ATCAC.A..AAA.CACCA.CCAT.CA.......C..AA.TTGGT....
Coelacanth22   A.........CGGG...T....G...TCC...AC.ATCAC.A..A.......................................
Coelacanth23   A.........CAGA...T....G...TCT...GC.ATCAC.A..AAGCC..CA.CCAT.CA.......C..AG.TTGGT....
Coelacanth24   A.........CAGG...T....G...TCC...AC.ATCAC.A..A.......................................
Coelacanth25   A.........CAGG...T....G...TCC...AG.ATCAC.A..AAA.C..CA.CCAT.C......................
Coelacanth26   A.........CAGG...T....G...TCC...AC.ATCAC.ATAGAA.A..CA.CCAT.CA.......C..AG.TTGGC....
Coelacanth27   A.........GAGG...T....G...TCC...AC.AACAT.A..ACA.C..CA.CCAT.CA.......C..AG.TTGGC....
Coelacanth28   A.........CTAC...T....G...TCC...AC.ATTAC.A..AAA.C..CA.TCAT.CG.....................
Coelacanth29   A.........GAAC...T....G...TCC...AT.ATCAG.A..AAA.G..CA.CCAT.TA.....................
Coelacanth30   G.........CAAA...T....T...TTC...AC.ATCAT.A..AAA.T..CA.CCAT.AA.......C..AG.TTGGC....
Coelacanth31   A.........CAAA...T....A...TCC...AC.ACAAC.A..AAA.C..CA.T.......................AA.TTG....
Coelacanth32   A.........TATG...TACATG...TCC...AC.ATCCC.A..AAA.C..CT.CCAT.TA.......T..AA.ATTGC....
Coelacanth33   A.........CAGA...T....G...CCC...AC.ATCAC.A..AAA.C..CA.GCAT.CA.......C..AA.ATGGC....
Coelacanth34   A.........TGGA...A....G...G.....GC.ATCAC.A..AAA.C..CA.CTGT.CA.....................
Coelacanth35   A.........CAGG...T....A...TCC...AC.ATCAT.A..GAAGC..CA.CCAT.CA.......C..AA.CTGGC....
Coelacanth36   ...................................................................................
Coelacanth37   C.........TAGG...T....G...TCT...AC.ATGAC.A..AAA.C..CA.CCAT.TA.......C..AA.CTGGG....
Coelacanth38   .................................AC.ATCAC.C..AAA.C..CG.TCAT.CA....................
Coelacanth39   A.........CAGG...T....G...TCT...AC.ATCAC.G..AAA.C..CA.CCAT.CA.......C..AA.TAGGC....
Coelacanth40   A.........CAAC...T....G...TCC...A..................CAT.CA.......T..AA.TTGGC....
Coelacanth41   G.........CAGG...T....G...T.C...AC.ATCAC...............................................
Coelacanth42   ...................................AC.A..GAA.C..CAT.CCAT.CCATAAAATC..CA.TTAGC....
Coelacanth43   ......................G....TCT...AC.ATCAT.A..AAA.........A.........G..GG.TGGTA.....
Coelacanth44   ATAGGGTTGCCAGA....T....G...CCCTTTAT.TTTAA.G..GGA.T..GT.CCCT.CA.......T..TT.TAGGCTAAAA
Coelacanth45   ...................................................................................
Coelacanth46   A.........CTAC...A....G...TCC...AC.................................C..AA.TCAGC....
Coelacanth47   A.........CTGA...T....G...TCC...AC.ACCAC.A..AAA.G..TA.CACA.CA.......C...............
Coelacanth48   A.........C......T....G...TCC...AC.GTCA...............................................
Coelacanth49   ...................................................................................
Coelacanth50   ...................................................................................
Coelacanth51   A.........CTTT...T....G...GCT...AC.ATCAG.A..CAA.A..AA...AA.TG.......C..AT.TTGGC....
Coelacanth52   ...............T....G...CAC...AT.ATCAC.A..AAG.T..CA.CCAT.CA.......C..AG.CTGAC....
Coelacanth53   ...................................................................................
Coelacanth54   A.........CTTT...T....G...GCT...AC.ATCAG.A..CAA.A..AA...AA.TG.......C..AT.TTGGC....
Coelacanth55   ...................TTG...AT.ATTAT.A..AAC.C..CA.CCAT.TG.......C..AA.ATGAC....
Coelacanth56   A.........CTGC...A....T...TCC...AGTATCTG.C..CAA.T..CA.ACAC.CA.....................
Coelacanth57   ...............G....T....G...TCC...AC.AGCGC.A..AAACG..CA.CCAT.GG.......C..AG.CTAGT....
Coelacanth58   ...................................................................................
OutGroup       A.........CAGG...T....G...TCC...AC.ATCAC.A..AAAAC..CA.CCAT.CA.......C..AA.TTGGC....
PCBP2-Human    A.........CAGG...T....G...GGTGTCC...AC.AGCACCA..AAAAT..CA.CCAG.CG.......C..CA.CTGGC....
PCBP2-Chimp    A.........CAGG...T....G...GGTGTCC...AC.AGCACCA..AAAAT..CA.CCAG.CG.......C..CA.CTGGC....
PCBP2-Dog      A.........CAGG...T....G...GGTGTCC...AC.AGCAC.A..AAAAT..CA.CCAG.CG.......C..CA.CTGGC....
PCBP2-Mouse    A.........CAGG...T....G...GGTGTCC...AC.AGCAC.A..AAAAT..CA.CCAG.CG.......C..CA.CTCGC....
PCBP2-Rat      A.........CAGG...T....G...GGTGTCC...AC.AGCAC.A..AAAAT..CA.CCAG.CG.......C..CA.CTCGC....
PCBP2-Opossum  A.........CAGG...T....G...GGTGTCC...AC.AGCAC.A..AAAAT..CA.CCAG.CG.......C..CA.CTGGC....
ISL1-Human     A.............T....G...TCC...AC.ATCAC.A..AAA.C..CA.CAAC.CC.......TT.AA.TTTTC....
ISL1-Chimp     A.............T....G...TCC...AC.ATCAC.A..AAA.C..CA.CAAC.CC.......TT.AA.TTTTC....
ISL1-Dog       A.............T....G...TCC...AC.ATCAC.A..AAA.C..CA.C.AT.CA.......C....TTAAT....
ISL1-Mouse     A.............T....G...TCC...AC.ATCAC.A..GAG.C..CA.CAACACT.......T..AA.TTGTA....
ISL1-Rat       A.............T....G...TCC...AC.ATCAC.A..AAG.C..CA.CAACACT.......T..AA.TTGTA....
ISL1-Opossum   ..........AAA...T....G...TCC...AC.ATCAC.A..AAA.C..CA.CTAT.........AC.TTAAT....
ISL1-Chicken   A.............T....G...TCC...AC.ATCAC.A..AAA.C..CA.CTGT...............AC
ISL1-Frog      ..........AAA...T....G...TCT...AC.ATCAT.A..AAA.C..AA.CT.T.CA.......CTTAA.TTTCC....
DACH1-Human    ..........TAGG...T....G...CCC...AT.ATCAG.A..AAAGT..CA.CCAT.TG.......T..AA.TTGGC....
DACH1-Chimp    ..........TAGG...T....G...CCC...AT.ATCAG.A..AAAGT..CA.CCAT.TG.......T..AA.TTGGC....
DACH1-Dog      ..........GGG...T....G...CCC...AT.ATCAG.G..AAAGT..CA.CCAT.TG.......T..AA.TTGGC....
DACH1-Mouse    ..........TAGG...T....G...CCC...AT.ATCAG.A..AAAGT..CA.CCAT.TG.......T..AA.TTGGC....
DACH1-Rat      G.........TAGG...T....G...CCC...AT.ATCAG.A..AAAGT..CA.CCAT.TG.......T..AA.TTGGC....
DACH1-Opossum  ..........AGG...T....A...CCC...AT.ATCAG.A..AAAGT..CA.CCAT.TG.......T..AA.TTGGC....
```

Figure S14: **LF-SINE reconstruction (contd).**

Figure S14: **LF-SINE reconstruction (contd).**

Figure S14: **LF-SINE reconstruction (contd).**

```
LF-SINE        AAC.........TAC.CCT..CT.CAAC...CCT......GTAG...AGGT.....GG.....TCCCT.CC.AGGG.CAG...
LF-SINE2       AAC.........TAC.CCT..CT.CAAC...CCT......GTAG...AGGT.....GG.....TCCCT.CC.AGGG.CAG...
Coelacanth0    AAC.........TAC.CCT..CT.C.AC...CCT......GTA.....GT.....GG.....TCCTT.CC.AGGG.CAG...
Coelacanth1    AAC.........TAC.TCT..C...AC...CCT......GTAG...TGGT.....GG.....TCCCT.CC.AGGG.CAG...
Coelacanth2    AAC.........TAC.CCC..TT.TCTC...CCT......ATAG...AGGT.....GG.....TCCTT.CC.AGGA.CAG...
Coelacanth3    AAC.........TAC.TCT..CT.C.AC...CTT......GTAG...AGGT.....TA.....TCTCT.CT.GGGA.CAA...
Coelacanth4    AAC.........TAC.CCC..TT.TCAC...CCT......CTAG...AGGT.....GA.....TCCCT.CC.AGGG.CAG...
Coelacanth5    AAC.........TAC.CCT..CT.CA.C...CCT......GTAG...AGAT.....TG.....TCCCT.CC.AAGG.CAG...
Coelacanth6    AAT.........TAT.TAT..CC.C.AC...CCT......CTAG...AGAT.....GG.....TTACT.CC.ATGG.TAG...
Coelacanth7    AAC.........TAC.CCT..CC.C.AC...CCT......GTAG...AGGT.....AG.....TGCCT.CC.AGGG.TAG...
Coelacanth8    AAT.........TAT.CCC..TC.TCAC...CCT......ATAG...AGAT.....GG.....TCTCT.CC.AGGG.CAA...
Coelacanth9    AAC.........TAC.CCT..CC.........AG......AGGT.....GG.....TCCCT.TT.AGGA.AAG...
Coelacanth10   AAC.....TAC.CCC..CC.TCCC...CCT......ATAA...CAGT.....GG.....TCCCT.CC.CGGG.CAG...
Coelacanth11   AAC.........TAC.CCC..TC.TCAC...CC.......AG...CAGT.....GG.....TCCCT.CC.AGGG.CAG...
Coelacanth12   AAC.........T.........AC...TCT......ATAG...AGGT.....GG.....TCCCT.CC.AAAG.CAG...
Coelacanth13   AAC.........TAC.CCT..CT.C.AC...CTG......CTGG...AGGT.....GG.....TCCCT.GC.A.GG.CAA...
Coelacanth14   AAC.........TAC.CCC..TT.TCAC...TAT......CTAG...ATGT.....GG.....TTGTT.CC.AGAG.CAG...
Coelacanth15   AAC.........TAC.CCT..C......CTT......CCAG...AAAT.....AA.....TTCCT.CC.AGAG.CAG...
Coelacanth16   AAC.........TAC.CCT..CT.C.AC...CCT......GTAT...AGAT.....TG.....TCCCT.CC.AAGG.CAG...
Coelacanth17   AAC.........TAC.TCT..C.........CAG......AGGT.....GG.....TCCCT.CC.AGGG.CAT...
Coelacanth18   AAC.........TAT.CCT..CT.C.AC...CCT......GCAG...AGGT.....GG.....CTACT.CC.AGGG.GAG...
Coelacanth19   AAC.........TTC.A.............CCCT.CC.AGGG.TAG...
Coelacanth20   AAC.........TAC.CCT..CT.C.AC...CCT......GTAA...TGGT.....AG.....TATTT.CC.AAGG.AAA...
Coelacanth21   AAC.........TAT.CCC..TC.TCAC...CCT......ATAG...TGGT.....GG.....TCCCT.CC.AGGG.CAG...
Coelacanth22   AAC.........TAC.CTT..CT.C.AC...CCT......GTAA...AGGT.....GG.....TCCCT.GC.AGGA.CA....
Coelacanth23   AAG.........TA....T..TT.CATC...CC.......TAG...AGGT.....GG.....TCCCT.CC.AGGG.CAG...
Coelacanth24   AAC.........TAC.CAT..CT.CA.C...CCT......GCAA...AGAT.....GG.....TCCCT.CT.AGGG.CAA...
Coelacanth25   AAC.........TAC.TCT..CT.C.AC...TCT......CCTG...AGGT.....GG.....TTTCC.CC.CAG.......
Coelacanth26   AAC.....TTCAGTAT.CCT..CT..AAC...CCT......CTAG...AGGT.....GG.....TCCTT.TC.AGAG.CAG...
Coelacanth27   AAC.........TAT.CCC..TT.TCAC...CCT......GTAG...AGGT.....GG.....TCCCTACC.AGGG.CTG...
Coelacanth28   AAC.........CAC.CCT..CT..GT...CTT......CTGA...AGGT.....GG.....TCCCT.CC.AGAG.CAA...
Coelacanth29   AAG.........TAC.TCT..CT.CA.............CCCT.CC.AGGG.TAG...
Coelacanth30   ACC.........TAT.CCT..GT.CACA...T........G...AGGT.....AG.....TCCCT.CC.ACAA.TAG...
Coelacanth31   AAA.........CAT..CT.GT.TG.C...CCT......CTAG...AGAT.....GG.....TCCCT.CC.AGGG.CAG...
Coelacanth32   AAC.........TCC.TCT..TT.CA....CTT......GTAG...AGGT.....GG.....TCCCT.CC.AGGG.CAG...
Coelacanth33   AAC.........TAC.ACT..GT.CACC...TCTCATCTAATAA...TGGT.....GT.....CCACT.CTGAACA.AAT...
Coelacanth34   AAC.........TAC.CCA..CT.TACC...TCC......CAG...AGGT.....TG.....ATCCT.CT.GTAG.CA...
Coelacanth35   AAC.........TAC.CCT..TT.CATG...TCC......TTGA...GGGC.....TG.....TCCTT.CC.AGGT.CAG...
Coelacanth36   AAC.........TTCTCCT..CT.CAA....CCT......GTAG...AGGT.....GG.....TCTCT.CC.AGGG.CAG...
Coelacanth37   AAC.........TAT.TCT..C.........TAG......AGGA.....AG.....TCCTT.CC.AAGG.AAG...
Coelacanth38   AAC.........TGC.AAT..CT.CAAC...T........T.CT.AAGG.CTG...
Coelacanth39   AGC.........TAC.CCT..CT.CACC...CAC......AGAG...GGT.....GG.....TTTCT.CC.AAGG.CAA...
Coelacanth40   AAG.........TAC.TTC..CT.CATC...CTC......TTGA...AGTT.....TG.....TCCCT.CC.AGGG.TAG...
Coelacanth41   AGC.........TGG.GTG..GG.ATAC...CTA......AAAA...AGGC.....GC.....CACCT.GA.GGCG.AGG...
Coelacanth42   AAC.........TAC.CCT..C.........TAG......AGGT.....AG.....TCCTT.............
Coelacanth43   ATC.........TAC.CC...AG.AAAC...CCT......GGAC...AAGT.....TC.....TATCT.GC.......
Coelacanth44   AGT.........CAC.CAT..CA.CAGCTGACAC......TCAT...GGGC.....AC.....CCCCT.TC.TTGG.CAG...C
Coelacanth45   ............................CCT.CT.AGGG.CAC...
Coelacanth46   AAT.........TAC.CCC..CT.TATT...CTA......TTAG...ACTC.....TG.....TTCCT.TT.AGGGTCAG...
Coelacanth47   AAG.........TAT.CCT..CT.TAGT...CAA......TTAG...GGGTT.....GG.....TGCCT.CC.AGGT.CTA...
Coelacanth48   AAC.........TAT.TCT..CT.AATC...CCA......AAGA...TGGT.....GG.....CCTCT.TC.AGGG.CAG...
Coelacanth49   GAC.........TAC.CC...........CT...GTCA.............CCCT.CC.AGGG.CAA...
Coelacanth50   ACC.........TAC.TCT..CT.TATT...CCT......TCAA...GTGT.....GT.....TCTCT.AT.AA..CAG...
Coelacanth51   AAG.........CAA.TTT..CT.TACT...CCA......TTAG...AGGCTGGTATG.....TACCT.CC.ATGT.CTA...
Coelacanth52   AAC.........TAC.CCT..CT.CACC...CTT......AAAG...GCGC.....TG.....TCGCC.C.......
Coelacanth53   ............................GTAG...AGAT.....TT...CAAC.TT.AGGT.TAG...
Coelacanth54   AAC.GAAA.....CAA.TTT..CT.TACT...CCA......TTAG...AGGC.....TGGTATGTACCT.CC.ATCT.CTA...
Coelacanth55   AAC.........TAC.TCTTACC.CTCC...CTT......TTTA...AAAT.....GAATT..TAGTG.CT.ATGG.AAA...
Coelacanth56   TAA.........TAC.GTT..TTGTCTC...CAT......AATA...AGGC.....AG.....CCCAT.CC.AAGT.CAG...
Coelacanth57   ...................................................................................
Coelacanth58   ...................................................................................
OutGroup       AAC.........TAC.CCT..CT.CATC...CCT......GTAG...AGGT.....GG.....TCCCT.CC.AGGG.CAG...
PCBP2-Human    AACA.........TCC.CCT..CT.CA....CCT......CTAG...AGGT.....GG.....TCCCT.CC.AGGT.CAG...
PCBP2-Chimp    AACA.........TCC.CCT..CT.CA....CCT......CTAG...AGGT.....GG.....TCCCT.CC.AGGT.CAG...
PCBP2-Dog      AACA.........TCC.CCT..CT.CA....CCT......CTAG...AGGT.....GG.....TCCCT.CC.AGGT.CAG...
PCBP2-Mouse    AACA.........TCC.CCT..CT.CA....CCT......CTAG...AGGT.....GG.....TCCCT.CC.AGGT.CAG...
PCBP2-Rat      AACA.........TCC.CCT..CT.CA....CCT......CTAG...AGGT.....GG.....TCCCT.CC.AGGT.CAG...
PCBP2-Opossum  AACA.........TCC.CCT..CT.CA....CCT......CTAG...AGGT.....GG.....TCCCT.CC.AGGT.CAG...
ISL1-Human     TAT.........TAC.CCT..C..CATT...CCT......GTGA...GGGT.....GG.....TCCAG.CC.AGGT.CAA...
ISL1-Chimp     TAT.........TAC.CCT..C..CATT...CCT......GTGA...GGGT.....GG.....TCCAG.CC.AGGT.CAA...
ISL1-Dog       TAT.........TAC...T..CT.CCAT...TCT......GTGA...GGGT.....GG.....TCCAG.CC.AGGT.CAA...
ISL1-Mouse     TAT.........TAC.CCT...T.CATT...CCT......ACCA...GGGT.....GG.....TCCAG.CC.AGGT.CAA...
ISL1-Rat       TAT.........TAC.CCT..T.CATT...CCT......GCCA...GGGT.....GG.....TCCAG.CC.AGGT.CAA...
ISL1-Opossum   TAT.........TAC.CTT..C..CATT...CTA......TTGA...GGGT.....GG.....TCCAG.CC.AGGT.CAA...
ISL1-Chicken   TAT.........TAC.C.T..CC.CATT...CTA......GTAA...GGGT.....GG.....TCCAG.CC.AGGT.CAA...
ISL1-Frog      TAT.........TA..CCT..CT.CACA...CTG......CTGGTGAGGGT.....GG.....TCCAG.GT.AGGT.CAA...
DACH1-Human    AAT.........TAT.CCT..CT.CATC...CTG......G..G...AAGC.....GG.....CCTC.CC.AGGG.CAA...
DACH1-Chimp    AAT.........TAT.CCT..CT.CATC...CTG......G..G...AAGC.....GG.....CCTC.CC.AGGG.CAA...
DACH1-Dog      AAT.........TAT.CCT..CT..GAT...CCT......G.AG...AAGT.....GG.....GCCTC.CC.AGGG.CCGCGC.
DACH1-Mouse    AAT.........TAT.CCT..CT.CATC...CT......GAG...AAGT.....GG.....GCCTC.CC.AGGG.CAG...
DACH1-Rat      AAT.........TAT.TCT..CT.CATC...CT......GAG...AAGT.....GG.....GCCTC.CC.AGGG.CAC...
DACH1-Opossum  AAT.........TAT.CCT..CT.CATC...CT......GAA...AAGT.....GG.....GCCTC.CC.AGGG.CAG...
```

Figure S14: **LF-SINE reconstruction (contd).**

```
LF-SINE        ...GGTT..............GA..........GGCACAT.TG......GCAG....GG.....CAATGTG.....GGGAA.GC..CT
LF-SINE2       ...GGTT..............GA..........GGCACAT.TG......GCAG....GG.....CAATGTG.....GGGAA.GC..CT
Coelacanth0    ...GGTT..............GA..........GGTACAT.TG......GCAG....GG.....CAATGTG.....GGGAA.GC..CT
Coelacanth1    ...GGTT..............AA..........GGTACAT.TG......GTGG....GG.....CAAGGTG.....GGGAA.GC..CT
Coelacanth2    ...AGTT..............GA..........GGCACATTATG......GCAG....GG.....CAATGTG.....GGAAA.GC..CT
Coelacanth3    ...AGTT..............GA..........GGTACAC.TG......GCTG....GA.....CAGTGTG.....GGGAA.GC..CT
Coelacanth4    ...GGTT..............GA..........GGCACAT.TG......GCAG....GA.....CAATGTG.....GGGAA.GC..CT
Coelacanth5    ...GCTT..............GA..........GGCACAT.TT......GTGG....GG.....CAATGTG.....GGGAA.GT..CT
Coelacanth6    ...AGTT..............GA..........TGAGCAT.TG......GCAG....GA.....CAGCATG.....GGGAG.GC..CT
Coelacanth7    ...GGGA..............GG..........GGCACAT.TG......GCAA....GG.....CAACATG.....GGGAA.GC..CT
Coelacanth8    ...GGTA..............GA..........GGCACAT.TG......G.GG....GG.....AAATGTG.....GGGAA.GC..CT
Coelacanth9    ...GATT..............GA..........GGCACAC.TG......GTGC....GG.....TAATATG.....GTGAA.GC..CT
Coelacanth10   ...GGCT..............GA..........GGCACAT.TG......GCAG....GG.....CAATGTG.....GGGAA.GC..CT
Coelacanth11   ...GGTC..............AA..........GGCACAT.TG......GCAG....GG.....TAATGTG.....GGGAA.GC..CT
Coelacanth12   ...GGTT..............GA..........AGCACAT......................TGTG.....AGAAA.AG..CT
Coelacanth13   ...GGCT..............GA..........GAGCAT..TA......GAAG....GG.....CAGTATG.....GGATG.GC..AC
Coelacanth14   ...GGTT..............GA..........GACACAT.AG......GTGG....GG.....CAATGTG.....GAGAT.GC..CT
Coelacanth15   ...GATC..............AA..........AGCACAT.AA......GCAG....GG.....CAACGTG.....GGGAA.GT..CC
Coelacanth16   ...GCTT..............GA..........GGCACAT.TT......GTGG....GG.....CAATGTG.....GGGAA.GT..CT
Coelacanth17   ...GGCT..............GA..........GGCACAC.TG......GTAG....GG.....CAATGTG.....GGGAA.GC..CT
Coelacanth18   ...TATT..............AA..........GGCACAT.TG......GT.G....GG.....CAATGTG.....GGGAA.GC..CT
Coelacanth19   ...TGTT..............GA..........GGAAGAT.TG......GCGG....GG.....CAATGTG.....GGGAA.GC..TT
Coelacanth20   ...GGTT..............GA..........GACACAT.TG......GC.G....GA.....AAATGTG.....GGGAA.GC..CT
Coelacanth21   ...GATT..............GA..........GGCACAT.TG......ACAG....GG.....CAGTGTG.....GAGAA.GC..CT
Coelacanth22   ........GGTT.........GA..........TGCACAT.TG......GTAG....GG.....CACTGTG.....GAGAA.GA..CT
Coelacanth23   ...GGTT..............GA..........GGCACAT.TT......GCAA....AA.....CAGTGTG.....GGG.........
Coelacanth24   ...GGTT..............GA..........AGCACAC.TG......GT.G....GG.....CAATGTG.....GGGAA.GC..CT
Coelacanth25   .....................G...........................GCAG....GG.....CAATGGG.....GC....GT..CT
Coelacanth26   ...GGTT..............GA..........GGCACAC.AG......GAGG....GG.....TAATGTG.....TCAGT.TC..TC
Coelacanth27   ...GGTT..............GA..........GGCACAA.TG......GCAG....GG.....CAATGTG.....GGGAA.GC..CC
Coelacanth28   ...AGTT..............GA..........GGCACAT.TG......GCAG....GA.....CAGTATT.....AGGAA.GC..TT
Coelacanth29   ...GGTT..............GA..........GACACAT.AG......GTAG....TA.....CAGTTGG.....AAGAA.GC..TT
Coelacanth30   ...GACT..............GA..........AGCACAT.TG......ACAG....GA.....AAATATG.....GGAAA.GC..TT
Coelacanth31   ...GGTT..............GA..........GGCACAT.TG......GCAATA.AG.....CAATGTG.....GGGAA.GC..CT
Coelacanth32   ...GGTT..............GT..........GTCACAT.TG......GTGG....GCG....CAGTGTA.....GGAGA.GC..TG
Coelacanth33   ...AGTT..............GA..........AACACTT.TG......GTGG....GG.....CACTGTG.....AGGAA.GA..CT
Coelacanth34   ...GGCC..............GA..........GGTACGC.TG.TATGTGTGG....AG.....TGCTGTG.....GGGGA.GC..CT
Coelacanth35   ...GGTT..............GA..........GGCACA.TG......GTGG....GT.....CAGTGTG.....TACTA.GT..TC
Coelacanth36   ...GGCT..............GA..........GGCACAT.TC......AAAG....GG.....CAGTGTG.....GGGAA.GC..TT
Coelacanth37   ...GGAT..............GA..........GGTACAG.TG......GTGG....GC.....CAGTGTGTGTGTGGAGA.GG..GG
Coelacanth38   ...GGTT..............GA..........GGTACCT.AG......TCAGTCAGGG.....TAATGAG.....TAGAA.GC..TT
Coelacanth39   ...AGTT..............GA..........GGCTCAG.TG......GCAG....GG.....TGATATG.....GGGAA.GC..CT
Coelacanth40   ...AGTT..............AA..........AGCCCAT.TG......GC..................CTAGAGG.....TGAAA.GG..CT
Coelacanth41   ...GGTCTGGAGGCTGCCAGA.........GGCCCC.AG......TGGG....GT.....CCAGGGG.....GCAAA.GC..CC
Coelacanth42   .....................CAT.TG......GCAG....GG.....CAATAGA.....GGGAA.GC..AT
Coelacanth43   .....................C........AACACTC.TG......GCAT....GC.....CATCTGG.....GGGCA.GG..CT
Coelacanth44   CACAGCA..............GA..........GGGGCCA.AG......GTTT....GA.....ATGAGCA.....TGGAA.AC..CC
Coelacanth45   ....TT...............GA..........TGC.CAA.TG......GTAA....GG.....CAATGTG.....GGGAA.GC..CT
Coelacanth46   ...GGTT..............G...........................TG.....ACAA....GC.....CCATATC.....AC.AA.GC..AT
Coelacanth47   ...GGTT..............GA..........AAGATAC.TC......AGAG....G......CAATGTG.....GTGAC.AC..TT
Coelacanth48   ...GGTT..............AC..........TGTGCAT.TC......ATGA....GG.....CAGTGTG.....AAGAT.AT..CT
Coelacanth49   ...GGTT..............GA..........GCCACAC.TG......GCAG....GG.....CAA.........GC..CT
Coelacanth50   ...GGTT..............CA..........GGCACTT.TG......GCAA....GG.....AAGTGTG.....GGGAA.GC..CC
Coelacanth51   ...GGTT..............GA..........AGGAT.T.TG......AAAA....AG.....GAAAGAA.....GAGAT.AC..CT
Coelacanth52   ...........................................................................
Coelacanth53   ...GTCT.........GATTGCATTGGCCACAT.TG......GCAG....GG.....CAATGTG.....GGGAA.GC..CT
Coelacanth54   ...GGTT..............GA..........AAGAT.T.TG......AAAA....AG.....AAAAGAA.....GAGAT.AC..CT
Coelacanth55   ...GGT...............................GCCAGAT.TG......ACAG....CC.....CCTGGAG.....AGGAA.G.....
Coelacanth56   ...GGTT..............GA..........GACTTAC.TG......GCGG....AG.................................
Coelacanth57   ...........................................................................
Coelacanth58   ...........................................................................
OutGroup       ...GGTT..............GA..........GGCACAT.TG......GCAG....GG.....CAATGTG.....GGGAA.GC..CT
PCBP2-Human    ...GGTT..............GA..........GGCACAT.GG......ACGG....GG.....TGGTGTG.....GGGAAAGC..TC
PCBP2-Chimp    ...GGTT..............GA..........GGCACAT.GG......ACGG....GG.....TGGTGTG.....GGGAAAGC..TC
PCBP2-Dog      ...GGTT..............GA..........GGCACAT.GG......ACGG....GG.....TGGTGTG.....GGGAAAGC..TC
PCBP2-Mouse    ...GGTT..............GA..........GGCACAT.GG......ACGG....GG.....TGGTGTG.....GGGAAAGC..TC
PCBP2-Rat      ...GGTT..............GA..........GGCACAT.GG......ACGG....GG.....TGGTGTG.....GGGAAAGC..TC
PCBP2-Opossum  ...GGTT..............GA..........GGCACAT.GG......ACGG....GG.....TGGTGTG.....GGGAAAGC..TC
ISL1-Human     ...AGTC..............GA..........GGT.CAA.TA......ATAG....GG.....CA.........GGAA.G.....
ISL1-Chimp     ...AGTC..............GA..........GGT.CAA.TA......ATAG....GG.....CA.........GGAA.G.....
ISL1-Dog       ...AGTC..............GA..........GGT.CAA.TA......ATAG....GG.....CAA........GAA.G.....
ISL1-Mouse     ...AGTC..............GA..........GGT.CAA.TA......ATAG....GG.....CA.........GAAA.G.....
ISL1-Rat       ...AGTC..............GA..........GGT.CAA.TA......ATAG....GG.....CA.........GAAA.G.....
ISL1-Opossum   ...AGTC..............GA..........GGT.CAA.TA......ATAG....GG.....CAA........GAGAA.GC....
ISL1-Chicken   ...AGTC..............GA..........GGT.CAA.TA......GTAG....GA.....CA.........GGGAA.G.....
ISL1-Frog      ...AGCC..............GA..........GGT.CAA.TA......GCAG....GA.....CA...................
DACH1-Human    ...GGAT..............GA..........AGCACTG.TC......ACAG....TG.....CATGGTG.....GGAAG.G.....
DACH1-Chimp    ...GGAT..............GA..........AGCACTG.TC......ACAG....TG.....CATGGTG.....GGAAG.G.....
DACH1-Dog      ...GGTG..............CA..........GGTGCAG..GT......GCAC....GC.TGCACGGTGCA.....CGGTG.CT..CG
DACH1-Mouse    ...AGAT..............GA..........AGCATCC.GGC.....GCGG....TC.....CTGGCG.....GGAAG.G.....
DACH1-Rat      ...AGAT..............GA..........AGCGTCC.AGC.....GCTG....TC.....CCTGGCG.....GGAAG.GCGACTT
DACH1-Opossum  ...AGAT..............GA..........AGCACAT.TA......GCAT....GG.....CATTGTG.....GGAAA.TC..CG
```

Figure S14: **LF-SINE reconstruction (contd).**

LF-SINE
LF-SINE2
Coelacanth0
Coelacanth1
Coelacanth2
Coelacanth3
Coelacanth4
Coelacanth5
Coelacanth6
Coelacanth7
Coelacanth8
Coelacanth9
Coelacanth10
Coelacanth11
Coelacanth12
Coelacanth13
Coelacanth14
Coelacanth15
Coelacanth16
Coelacanth17
Coelacanth18
Coelacanth19
Coelacanth20
Coelacanth21
Coelacanth22
Coelacanth23
Coelacanth24
Coelacanth25
Coelacanth26
Coelacanth27
Coelacanth28
Coelacanth29
Coelacanth30
Coelacanth31
Coelacanth32
Coelacanth33
Coelacanth34
Coelacanth35
Coelacanth36
Coelacanth37
Coelacanth38
Coelacanth39
Coelacanth40
Coelacanth41
Coelacanth42
Coelacanth43
Coelacanth44
Coelacanth45
Coelacanth46
Coelacanth47
Coelacanth48
Coelacanth49
Coelacanth50
Coelacanth51
Coelacanth52
Coelacanth53
Coelacanth54
Coelacanth55
Coelacanth56
Coelacanth57
Coelacanth58
OutGroup
PCBP2-Human
PCBP2-Chimp
PCBP2-Dog
PCBP2-Mouse
PCBP2-Rat
PCBP2-Opossum
ISL1-Human
ISL1-Chimp
ISL1-Dog
ISL1-Mouse
ISL1-Rat
ISL1-Opossum
ISL1-Chicken
ISL1-Frog
DACH1-Human
DACH1-Chimp
DACH1-Dog
DACH1-Mouse
DACH1-Rat
DACH1-Opossum

Figure S14: **LF-SINE reconstruction (contd).**

36

Figure S14: **LF-SINE reconstruction (contd).**

```
LF-SINE       ....TCTA.........GCACCTT.TC...ACGA............GCACT.AAA.T..TC......A.C...ACAA..AA.
LF-SINE2      ....TCTG.........GCACCTT.TC...ACGA............GCACT.AAA.T..TC......A.C...ACAC..AA.
Coelacanth0   ....TCTG.........GCACCTT.TC...ACAA............GCACT.AAA.T..TC......A.C...ACAC..AA.
Coelacanth1   ....TCCA.........GCACCTT.TC...ATGA............GCACT.AAA.T..TC......A.C...ACAC..AA.
Coelacanth2   ....AGCA.........GCACCTT.TT...ACTA............GCACT.AAA.T..T.......T...TAAA..AA.
Coelacanth3   ....TCCA.........GCACCTT.TC...ACGA............GCACT.AAA.T..TT......A.C...ATGC..AA.
Coelacanth4   ....TCTG.........GCACCTT.TC...ACGA............GCACT.AAA.T..TC......A.C...ACAC..AC.
Coelacanth5   ....TCCA.........GCACCTT.TC...ACGA............GCACT.AAA.T..TT......A.C...ATGC..AA.
Coelacanth6   ....TCTG.........GCACCTT.TC...ATGA............GCACT.AAA.T..TC......A.T...ATAA..AA.
Coelacanth7   ....TCTA.........GCACCTT.TC...ACTA............GCACT.AAA.T..TC......A.C...ACAA..AA.
Coelacanth8   ....TCTG.........GCATCTT.TC...ACGA............GCACT.AAA.T..TC......G.T...ACAC..AA.
Coelacanth9   ....TCTG.........GCACCTT.TC...ATGA............GCACT.AAA.T..TC......A.C...AAAA..AA.
Coelacanth10  ....TCTG.........GCAGCTT.TC...ATGA............GTACTGAAA.T..TC......A.C...ACAA..AA.
Coelacanth11  ....TCCA.........GCTGCCT.TC...ACAA............GCACT.AAA.T..TC......A.C...ACAC..AAT
Coelacanth12  ....TCCA.........GCATCTT.TC...ACAA...........................C.C...ACAA..AA.
Coelacanth13  ....GCCA.........GTACTTT.TC...ACAA............GCACA.AAA.T..TC......A.C...TTAA..AA.
Coelacanth14  ......G.........GCACCTT.TT...GTGA............GCACT.GTA.T..TC......A.T...GAAA..AA.
Coelacanth15  ....TCTG.........GCACCTT.TC...ATA.............ACT.ACA.T..TC......A.T...AAAA..AA.
Coelacanth16  ....TCCA.........GCACCTT.TC...ATGA............GCACT.AAA.T..T.....................
Coelacanth17  ....TCTG.........GCACCCT.TC...ACAA............GCACT.AAA.T..TT......A.C...AAAA..AA.
Coelacanth18  ....TTTT.........GC..........................................................
Coelacanth19  ....TCTA.........GCACCTT.TC...AAGA............GTCCT.AAA.T..TC......A.C.....A..AA.
Coelacanth20  ....TCTG.........GCACTTT.TC...ATGT............GCACT.AAA.T..TC......A.C...ACAC..AC.
Coelacanth21  ....TCTG.........GCACCTT.TC...ATGA............GCACT.AAG.T.TCTC.....A.T...ACAA..AA.
Coelacanth22  ....TCCA.........GAACCTT.TC...GCCA............GCACT.AAA.T..CC......A.T...ACACTTTA
Coelacanth23  ....TCTG.........GCACCTC.TC...CATA............GCACT.AAA.T..TC......A.T...TTTA..AA.
Coelacanth24  ....TCTG.........GCACATC.TC...TCAA.............................AA.AA.
Coelacanth25  ....TCCA.........GCACCTT.TC...ACAA............GCATT.AAG.T..TC......G.C...AT........
Coelacanth26  ....TCTG.........ACACCTT.TC...TTGG............GCACT.AAA.T..TC......A.C...TAAA..AG.
Coelacanth27  ....TTCA.........GCACCTT.TC...ATGA............GCATT.AAGCT..TC......A.C...AGAG..GAT
Coelacanth28  ....TCAC.........ACACCTT.TC...ATGA............ACACT.AAA.T..GT......A.C.............
Coelacanth29  ....TCTG.........GCACCCT.TC...ATAA............GTACG.AAG.T..T.......C...ACAA..AA.
Coelacanth30  ....TCTG.........CTATGTT.CA...AAGA............GCACT.AAA.G..TC......A.C...TTAA..AC.
Coelacanth31  ....TCCT.........GAACTTT.TC...ATTA............TCACT.AAG.T..TC......A.C...AAAA..A..
Coelacanth32  ....TCCA.........ACTCCTT.TC...ACGA............ACACC.AAA.T..TC......A.C...TTAA..AA.
Coelacanth33  ...............ACCTT.TG...ATCA............CCATT.TAG.T..TC......C.C...CCC..CC.
Coelacanth34  ....TCTG.........ATACCTT.TT...AAAA............GCACT.AAA.C..TG......A.C..........A
Coelacanth35  ....TCTT.........GTATCTT.TC...ATAA............GCACT.AAG.T..TC......A.C...CAT..AA.
Coelacanth36  ....TCTG.........GCACCTT.TC...ACGA............GCACT.GAA.T..TT......G.C...ACAG..AA.
Coelacanth37  ...........................................................................
Coelacanth38  ....TTCA.........GCACCTT.TC...ACTA............GCACT.AAA.T..TC......A.C...........
Coelacanth39  ....AGT.........GCACCAC.AC...AGCT............GTTTT.AAA.C..TC......ATT...ACAA..AAT
Coelacanth40  ....TGGA.........AAACCT......................................................
Coelacanth41  ....TTAA.........GC..........................................................
Coelacanth42  ....TCTG.........ACACCTT.TC...ATAA............GCACC.AAG.T..TCTCTCAAG.C...CCAA..TG.
Coelacanth43  ATCCCATG.........GTGCCCT.TC...ACAA............AGAGT.A.......................
Coelacanth44  ....TCTA.........GCACCTT.CC...ATGA............ATATT.AAA.T..TC......A.C............
Coelacanth45  ....TCCA.........GCATCTT.TC...ATCA............GCACT.AAA.T..TC......................A.
Coelacanth46  ....TGTC.........TCATTTT.TC...ATGA............GCACT.GAA.T..TC......A.C...CATT..TG.
Coelacanth47  ....TTCC.........TCACCTT.TC...CAGA............GACGT.TAC.C..CC......G.T...AGAC..TT.
Coelacanth48  ....TCCC.........TCATCTT.TG...CCAA............GTCCT.AAA.T..TC......A.T...AACA..AA.
Coelacanth49  ....GCCA.........GCACCTT.TC...ACAA............GCACT.AAA....................
Coelacanth50  ....TCTG.........GAAATTC.TT...AGGC............GTGCT.AAA.T..TT......A.CTATACAA..AT.
Coelacanth51  ....TCCC.........TCACCCT.TC...CAAA............GAAGT.ACA.T..TC......A.T...AGAT..TC.
Coelacanth52  ...........................................................................
Coelacanth53  ....TCTG.........GTGCCTT.TT...ACAA............ATGTT.ATA.T..TC......A...............
Coelacanth54  ....TCCC.........TCACCCT.TC...CAAG............GAAGT.ACA.T..TC......A.C...AGAT..TC.
Coelacanth55  ...........................................................................
Coelacanth56  ....CCTG.........GATCCTT.T...................................................
Coelacanth57  ...........................................................................
Coelacanth58  ...........................................................................
OutGroup      ....TCTG.........GCACCTT.TC...ACGA............GCACT.AAA.T..TC......A.C...ACAA..AA.
PCBP2-Human   ....TCCG.........GCACCTTGTC...ACCA............GCATG.AAA.T..TC......A.C...ACAA..AA.
PCBP2-Chimp   ....TCCG.........GCACCTTGTC...ACCA............GCATG.AAA.T..TC......A.C...ACAA..AA.
PCBP2-Dog     ....TCCG.........GCACCTTGTC...ACCA............GCATG.AAA.T..TC......A.C...ACAA..AA.
PCBP2-Mouse   ....TCCG.........GCACCTTGTC...ACCA............GCATG.AAA.T..TC......A.C...ACCA..AA.
PCBP2-Rat     ....TCCG.........GCACCTTGTC...ATCA............GCATG.AAA.T..TC......A.C...ACAA..AA.
PCBP2-Opossum ....TCCG.........GCACCTTGTC...ACCA............GCATG.AAA.T..TC......A.C...ACAA..AA.
ISL1-Human    ....TTTGCACAAACAGATGTAGTCT.TC...ACTAATGATAAAAGGCTATACT.AAA.C..TC......A.A...ACAG..CA.
ISL1-Chimp    ....TTTGCACAAACAGATGTAGTCT.TC...ACTAATGATAAAAGGCTATACT.AAA.C..TC......A.A...GCAG..CA.
ISL1-Dog      ....TGCAAACAGAC....ATAGTCT.CC...ACTA............ATACT.AAA.......................
ISL1-Mouse    ....CCTG.........GAACAGT.CC...A...............ACT.CAA.C..TC......A.A...ATAG..CA.
ISL1-Rat      ...........................................................................
ISL1-Opossum  ...............TT.TC...ACAA............GCA...............................
ISL1-Chicken  ...........................................................................
ISL1-Frog     ...........................................................................
DACH1-Human   ....TTTG.........GCACCTT.GCTTGATGA............GCCCT.GAA.T..CT......A.C...TTAA..AA.
DACH1-Chimp   ....TTTG.........GCACCTT.GCTTGATGA............GCCCT.GAA.T..CT......A.C...TTAA..AA.
DACH1-Dog     ....TTTG.........GCACCTT.GCTTGACGA............GCCCC.GAA.T..CT......G.C...TTAA..AA.
DACH1-Mouse   ...........................................................................
DACH1-Rat     ...........................................................................
DACH1-Opossum ....CCTG.........GCACCTT.TCTTGATGA............GGACT.GAA.T..CC......A.C...TTAA..AA.
```

Figure S14: **LF-SINE reconstruction (contd).**

38

```
LF-SINE        ..AAA......TTTAAAAAAAA....AAAAAAAAAAA
LF-SINE2       ..AAA......TTTAAAAAAAA....AAAAAAAAAA.
Coelacanth0    ..AAT......TTAAAAAAAA....AAAAAAAAAA.
Coelacanth1    ..AAAATTT..TTTAAAAAAAA....TTTAAAGAAA.
Coelacanth2    ..AAA......TGCAGAGATA.....CAGAA......
Coelacanth3    ..AAT......TTAAAAAGAA....AAAAAAAAGAA
Coelacanth4    ..ACA.......................
Coelacanth5    ..AAT......TTAAAAAGAA....AAAAAAAAGAA
Coelacanth6    ..AAG......TTGATAGAGA.....GTGAGTAAGA.
Coelacanth7    ..AAA......TGTAAAAA.........
Coelacanth8    ..AAA......CTGAGAAAAA.....AATCAATTAAA
Coelacanth9    .GGG......AGCACAAGAA.....TATAGAGA...
Coelacanth10   ..ATT......GTAAAAAAAG.....TGAA.......
Coelacanth11   TTAGA......TTAAAAAAAA....AAAAGAAA...
Coelacanth12   ..ATT......CCTGAAAAAA....GAAAAAAAAA.
Coelacanth13   .ATA......CAAGGAATTA.....TAAAAAGGAGA
Coelacanth14   ..AAA......ATTAAGAGAA...............
Coelacanth15   ..AAA...................
Coelacanth16   ..........TGTAAAAAAA.....AAAAAAAAAAA
Coelacanth17   ..AAA......AATTAAA...........
Coelacanth18   ...........................
Coelacanth19   ..AAA......TGTGAAAAAA.....AAA........
Coelacanth20   .ACA......CATACACAAATTTTTAAAAAAA....
Coelacanth21   ..AAA......TTACAAAAAA....AAAAAAAAAA..
Coelacanth22   ..AAC......CCTAAAACAA.....AGCAA......
Coelacanth23   ..AA........................
Coelacanth24   ..AAT......TATAAAAAG.....AAAAAGAAAGA
Coelacanth25   ..........TTTGAAAAAG.....AAAATAGAA.
Coelacanth26   .AAGATACCTTTTATAATGG.....AAAAGGGAAA.
Coelacanth27   ..ATA......TCTCAGAAAA...........
Coelacanth28   ..........TTTAGTAAAA....AAAAAAAAAA.
Coelacanth29   ..ACA......TTAAGAAAAA.....A.........
Coelacanth30   ..TAA......TAAAAA............
Coelacanth31   ...........................
Coelacanth32   ..ACC......CCTGAA................
Coelacanth33   ..AAA......AAAAAAGAAA.....TAAAAAAAA.
Coelacanth34   ..AAT......TTTAAAAAAA............
Coelacanth35   ..AAA......TTT..............
Coelacanth36   ..AAG......TTTAAAAAAA.....AAA........
Coelacanth37   ...........................
Coelacanth38   ...........................
Coelacanth39   GTAAA......TTAGAAGGAA.....ACAAAA.....
Coelacanth40   ...........................
Coelacanth41   ...........................
Coelacanth42   ..GAA......TTCAATAAAG.....AAAA.......
Coelacanth43   ...........................
Coelacanth44   ..........TTTATAAAAA.....AA.........
Coelacanth45   ..AAA......TTTAATGACA.....AA.........
Coelacanth46   ..TAT......TTTATACAAA....AAAGAAAAGGA
Coelacanth47   ..ATT......TTTAATGAAA.....GAATATAATAA
Coelacanth48   ..A.........................
Coelacanth49   ...........................
Coelacanth50   ..AAA.......................
Coelacanth51   ..ATT......TTTAAAGAAA.....GAAGACAATAA
Coelacanth52   ...........................
Coelacanth53   ...........................
Coelacanth54   ..ATT......TTTAAAGAAA.....GAAGACAATAA
Coelacanth55   ...........................
Coelacanth56   ...........................
Coelacanth57   ...........................
Coelacanth58   ...........................
OutGroup       ..AAA......TTTAAAAAAA.....AAAAA......
PCBP2-Human    ..AAA......GTT..............
PCBP2-Chimp    ..AAA......GTT..............
PCBP2-Dog      ..AAA......GT...............
PCBP2-Mouse    ..AAA.........AAAAAAA.....AAA.......
PCBP2-Rat      ..AAA.......................
PCBP2-Opossum  ..AAA.......................
ISL1-Human     ..TAT......CTGATCAAAT.....AAAAA......
ISL1-Chimp     ..TAT......CTGATCAAAT.....AAAAA......
ISL1-Dog       ...........................
ISL1-Mouse     ..CAG......TCTATAAACA.....CAAA.......
ISL1-Rat       ...........................
ISL1-Opossum   ...........................
ISL1-Chicken   ...........................
ISL1-Frog      ...........................
DACH1-Human    ...........................
DACH1-Chimp    ...........................
DACH1-Dog      ...........................
DACH1-Mouse    ...........................
DACH1-Rat      ...........................
DACH1-Opossum  ...........................
```

Figure S14: **LF-SINE reconstruction (contd).**

39

Figure S15: **Human instances of the LF-SINE.** Alignment of the 50 most conserved human instances to the coelacanth LF-SINE. Notice that multiple instances match each portion of the SINE, and that several instances span nearly its entire length.

Figure S15: **Human instances of the LF-SINE (contd).**

Figure S15: **Human instances of the LF-SINE (contd).**

Figure S15: **Human instances of the LF-SINE (contd).**

Figure S15: **Human instances of the LF-SINE (contd).**

Table S1: **Available genomic DNA from the Indonesian coelacanth,** *Latimeria menadoensis*.

Region Of Homology	BACs	Non-overlapping Length (bp)	LF-SINE Instances
Protocadherin Cluster	AC150310 AC150283 AC150284 AC150309 AC150308	609,120	21
Hox Cluster	AC151571	187,392	21
Hox Cluster	AC147788	168,364	4
Genomic	AC140159	92,794	13
Total	8	1,057,670	59

BACs - bacterial artificial chromosome cloning vectors, listed using their GenBank accessions.

Table S2: **Whole genome drafts searched for LF-SINE instances.**

Species	UCSC Assembly	LF-SINE Detected	Species	UCSC Assembly	LF-SINE Detected
Homo sapiens	hg17	Yes	*Danio rerio*	danRer2	No
Pan troglodytes	panTro1	Yes	*Tetraodon nigroviridis*	tetNig1	No
Macaca mulatta	rheMac1	Yes	*Takifugu rubripes*	fr1	No
Mus musculus	mm6	Yes	*Ciona intestinalis*	ci1	No
Rattus norvegicus	rn3	Yes	*Strongylocentrotus purpuratus*	strPur1	No
Canis familiaris	canFam1	Yes	*Drosophila melanogaster*	dm2	No
Bos taurus	bosTau1	Yes	*Anopheles gambiae*	anoGam1	No
Monodelphis demestica	monDom1	Yes	*Caenorhabditis elegans*	ce2	No
Gallus gallus	galGal2	Yes	*Saccharomyces cerevisiae*	sacCer1	No
Xenopus tropicalis	xenTro1	Yes			

All assemblies are available at http://genome.ucsc.edu.

Table S3: **Distribution of LF-SINEs in available tetrapod genomes.**

Organism	5' UTR	3' UTR	Exonic Alt-Spliced	Total	Intronic	Intergenic	Total
Homo sapiens	1	0	12	13	68	163	245
Pan troglodytes	-	-	-	-	-	-	210
Macaca mulatta	-	-	-	-	-	-	229
Canis familiaris	-	-	-	-	-	-	235
Bos taurus	-	-	-	-	-	-	169
Mus musculus	0	1	7	8	25	57	91
Rattus norvegicus	-	-	-	-	-	-	87
Monodelphis domestica	-	-	-	-	-	-	394
Gallus gallus	0	1	2	3	244	451	699
Xenopus tropicalis	0	0	1	2	10	14	26

Assembly names are given in Table S2. Breakdown with respect to protein coding genes is shown in select species for which there is sufficiently independent gene annotation. Also see electronic supplement at http://www.cse.ucsc.edu/~jill/sine.html.

Table S4: **LF-SINE instances in partially sequenced tetrapods.**

```
                                                     Smallest Sum
                                             High    Probability
Sequences producing High-scoring Segment Pairs:   Score   P(N)     N

EM_OV:LC08034 U08034.1 Latimeria chalumnae clone 7 MHC cl...     792   5.8e-43   3
EM_OV:LC08043 U08043.1 Latimeria chalumnae clone 10 MHC c...     700   5.8e-35   2
EM_OV:AF131253 AF131253.1 Latimeria chalumnae rhodopsin (...     598   1.8e-19   1
EM_OV:AC147867 AC147867.3 Gopherus agassizii clone TAMGA-...     609   2.5e-19   1
EM_OV:AB125532 AB125532.1 Mauremys reevesii DNA, CryIIA S...     508   3.5e-15   1
EM_OM:AY208121 AY208121.1 Sus scrofa myostatin gene, comp...     396   1.9e-11   2
EM_OV:AB125431 AB125431.1 Macroclemys temminckii DNA, Cry...     367   9.4e-09   1
EM_OM:AC117941 AC117941.6 Felis catus clone RP86-565015, ...     292   7.6e-07   2
EM_OV:AF131260 AF131260.1 Latimeria chalumnae RH2 opsin (...     314   1.1e-06   1
EM_OV:CMDIMIB L42234.1 Chelonia mydas (clone Cm84) DNA re...     300   2.9e-06   1
EM_OV:AB125486 AB125486.1 Sternotherus odoratus DNA, CryI...     282   8.5e-05   1
EM_OV:AC147865 AC147865.3 Gopherus agassizii clone TAMGA-...     245   0.0071    2
EM_OM:AC123971 AC123971.3 Lemur catta clone LB2-255J23, c...     262   0.031     1
EM_OM:AC124905 AC124905.3 Equus caballus clone CH241-1840...     261   0.060     2
```

```
best hit  %ident  len.  mis   gap   e-value  bit    #hits  db    Species
(gnl|ti)              match  open         score  <1e-4

595468933 80.83   386   40    27    5e-73    283    >250   20G   Ornithorhynchus anatinus
819506662 79.62   373   42    25    1e-62    246    30     5G    Cavia porcellus
610716009 79.62   373   41    25    7e-62    244    66     7G    Oryctolagus cuniculus
311856343 79.41   374   40    27    1e-60    239    20     2G    Sus scrofa
716628186 79.14   374   41    27    2e-58    233    60     8G    Echinops telfairi
717304885 79.82   342   36    28    8e-57    227    44     5G    Pongo pygmaeus
598681119 78.88   374   37    31    2e-55    223    91     8G    Dasypus novemcinctus
477555815 78.98   352   39    26    3e-53    216    49     8G    Loxodonta africana
667813765 80.26   228   28    14    4e-37    162    57     6G    Felis catus
821401047 80.69   145   20    7     4e-23    110    4      400M  Pongo pygmaeus abelii
637106021 79.29   140   15    12    2e-17    89.1   35     33M   Alligator mississippiensis
343261025 78.38   111   15    8     4e-11    66.1   6      6M    Gopherus agassizii
426822539 87.72   57    5     2     2e-09    68.0   21     1G    Otolemur garnettii
723018327 92.31   39    3     0     3e-07    58.4   2      219M  Atelerix albiventris
497806131 84.13   63    6     4     3e-07    60.3   4      872M  Callithrix jacchus
645968230 82.35   68    7     5     4e-07    56.4   8      77M   Meleagris gallopavo
130066429 81.94   72    8     5     6e-07    58.4   8      464M  Papio cynocephalus
740209113 87.50   48    6     0     6e-07    58.4   4      536M  Rhinolophus ferrumequinum
774400718 80.82   73    5     9     1e-05    50.7   1      50M   Taeniopygia guttata
```

WU-Blastn hits to the EMBL nucleotide database (top) and Discontiguous Megablast Trace repository hits (bottom) from additional, partially sequenced tetrapods. While the Trace archive covers several tetrapods more extensively than the nucleotide database, it is also highly redundant. Columns 1-7 of the Trace hits table relate to the best hit in that species. The next two columns give the number of hits with E-value $< 10^{-4}$, and the total length of the redundant trace reads, respectively.

Table S5: **Chromosomal distribution of LF-SINEs in the human genome.**

Chromosome	hits	exp	hits/exp	Chromosome	hits	exp	hits/exp
chr1	14	19.15	0.73	chr13	14	8.21	1.71
chr2	26	20.41	1.27	chr14	11	7.59	1.45
chr3	18	16.72	1.08	chr15	9	6.99	1.29
chr4	19	16.08	1.18	chr16	6	6.78	0.89
chr5	23	15.27	1.51	chr17	3	6.68	0.45
chr6	13	14.38	0.90	chr18	13	6.41	2.03
chr7	10	13.30	0.75	chr19	2	4.79	0.42
chr8	12	12.25	0.98	chr20	3	5.11	0.59
chr9	12	10.12	1.19	chr21	3	2.94	1.02
chr10	11	11.31	0.97	chr22	0	2.99	0.00
chr11	6	11.27	0.53	chrX	7	12.92	0.54
chr12	10	11.19	0.89	chrY	0	2.14	0.00

The expected number of hits per chromosome is computed assuming uniform distribution over all sequenced bases of human genome draft hg17. The observed values diverge from this null model (chi-square and simulation $p \sim 0.01$), albeit no single chromosome, except the very short ones, deviates more than two fold from the expected value.

Table S6: **Tetrapodal genes containing an LF-SINE exonized instance.**

Species	Gene	Function	Exon
M	PCBP2	mRNA stability and shuttling	9
M	SMARCA4	SWI/SNF related matrix associated	27
M,C	EEF1B2	translation elongation factor	3
M	TCERG1	transcription factor	22
M	PTDSR	promote cell differentiation	5
M,C	RORA	DNA binding nuclear receptor	3
M,C	GRID1	neurotransmitter receptor	1
M,C	ATF2	transcription factor	14
M	FLJ22833	nucleic acid binding	4
M	ARHGAP6	GTPase activator	13
M,C	KIAA1409	unknown	34
M	NT5C2	5'-nucleotidase	9
M,C,F	LRP1B	low density lipoprotein receptor related	90
M	DHX30	unwind double-stranded nucleic acids	4
M,C	gg-DMTF1	transcription factor	12
C	gg-PPP2R2C	subunit of protein phosphatase	2
C	gg-SHFM1	subunit of 26S proteasome	3
F	xt-MBNL1	RNA binding / development	4
F	JGI-49280	unknown	5

Species: M - Mammals (Human, Mouse, Possum), C - Chicken, F - Frog. Also see Table S7.

Table S7: **LF-SINE instances overlapping tetrapodal exons.**

Gene Name	Sp.	Exon No.	Alt-Spl.	CDS/UTR	Ins. F.S.	Ins. Stop	Trig. NMD	3' S.S. ex5'	3' S.S. Seq.	5' S.S. ex3'	5' S.S. Seq.
PCBP2	H	9	+	C	-	-	-	381	cacagGACAG	289	TAGAGgtgag
	M		+	C	-	-	-	381	cacagGACAG	289	TAGAGgtgag
SMARCA4	H	27	+	C	-	-	-	381	cacagAGCAG	289	TAAAGgtgag
	M		+	C	-	-	-	381	cacagAGCAG	289	TAAAGgtgag
EEF1B2	H	3	+	C	-	+	$+^a$	376	aacagGTAGA	101	TCATAgtgag
TCERG1	H	22	+	C	+	+	$+^a$	469	tatagTTAAT	377	AACAGgtaca
	M		+	C	+	+	$+^a$	469	tatagTTAAT	377	AACAGgtaca
PTDSR	H	5	+	C	+	+	+	376	aacagATACA	289	TCGGGgtaag
	M		+	C	+	+	+	376	aacagATACA	289	TCGGGgtaag
RORA	H	3	+	C	+	+	+	366	cgcagGGCAG	289	AGGTGgtaag
	M		+	C	+	+	+	366	cgcagGGCAG	289	AGGTGgtaag
GRID1	H	1	+	5U						289	GTAGGgtaag
ATF2	H	14	+	C	+	+	$+^a$	450	taaagTGAAT	377	ACCAGgtaca
	M		+	$3U^b$								
FLJ22833	H	4	+	C	+	+	+	381	cacagTCTAG	280	GTAAGgtaat
ARHGAP6	H	13	+	C	-	+	+	381	cccagAACAA	289	TAAGGgtgag
	M		+	C	-	+	+	381	cacagAACCA	289	TAAGGgtgag
KIAA1409	H	34	-	C	-	-	-	381	cacagAACAG	281	GAGAGgttag
	M		-	C	-	-	-	381	cacagAACAG	281	GAGAGgttag
	C		-	C	-	-	-	381	cacagAGCAG	281	CAGAGgttag
NT5C2	H	9	+	C	+	+	+	450	taaagTGAAT	107	GGATGgtaat
	M		+	C	+	+	+	450	taaagTGAAT	240	GGGAGgtttg
LRP1B	H	90	+	C	-	-	-	381	tacagGCCAG	289	CTGGGgtgag
DHX30	H	4	+	C	-	-	-	381	cccagATCGG	289	TCGAGgtaag
gg-DMTF1	C	12	+	C	+	+	+	450	tttagTGAAT	377	AACAGgtaca
gg-PPP2R2C	C	2	+	C	+	+	+	381	cacagGAGAG	289	TGGAGgtgag
gg-SHFM1	C	3	+	$3U^b$							
xt-MBNL1	F	4	+	C	-	-	-	381	cacagGCCAG	289	TATGGgtgag
JGI-49280	F	5	+	C	+	+	$+^a$	450	tctagTGATT	377	AACAGgtaaa

^{a} The very first fully encoded codon is a stop codon.
^{b} Instance is embedded inside a larger 3' UTR region.
NMD - nonsense mediated decay. Predicted where the inserted stop codon appears more than 50 nucleotides upstream of the final splice junction (Lewis *et al.*, 2003).
H-Human, M-Mouse, C-Chicken, F-Frog.
All splice region sequences are lowercase when in an intron and uppercase when in a coding region. The numbering of the splice sites corresponds to the first and last coding base of the exon according to the alignments in Fig. S9. Exon number is computed according to the longest annotated transcript in the primary species. As annotation for these genomes improves, more instances of the LF-SINE may be shown to overlap coding exons.

Table S8: **Exapted Human LF-SINE splice site distribution.**

SS	Location	Conversion	Changes	Genes
3'	366	TG→AG	1	RORA
3'	376	AG→AG	0	EEF1B2, PTDSR
3'	381	AG→AG	0	PCBP2, SMARCA4, FLJ22833, ARHGAP6, KIAA1409, LRP1B, DHX30, PPP2R2C, MBNL1
3'	450	TG→AG	1	ATF2, NT5C2, DMTF1, JGI-49280
3'	469	TT→AG	2^a	TCERG1
5'	101	GAC→AGT	3^a	EEF1B2
5'	107	GGT→GGT	0	hg-NT5C2
5'	240	TGC→GGT	2	mm-NT5C2
5'	280	GGT→GGT	0	FLJ22833
5'	281	GGG→GGT	1	KIAA1409
5'	289	GGT→GGT	0	RORA, PTDSR, PCBP2, SMARCA4, ARHGAP6, LRP1B, DHX30, PPP2R2C, MBNL1
5'	377	GGT→GGT	0	ATF2, DMTF1, JGI-49280, TCERG1

[a] The perfect mismatch in these cases may suggest that the LF-SINE sequence had diverged in the local region beyond our ability to align it correctly.

The table shows an inverse correlation between the number of mutations necessary to activate a splice site, and the number of characterized instances that use it. The numbering of the splice sites corresponds to the first and last coding bases of the exons according to the alignments in Fig. S9.

Table S9: **RNA folding potential of LF-SINE instances.**

Rank	Orient.	qFrom	qTo	tLength	%GC	MFE	*p*-value	Comment
latMen	+	1	480	476	47%	-172.30	0/10000	Coelacanth LF-SINE
#1	+	103	463	359	57%	-163.10	0/10000	PCBP2 exon
#183	+	55	460	384	47%	-140.40	0/10000	C20orf17 intron
#2	-	1	457	413	47%	-130.40	0/10000	
#4	-	212	481	262	49%	-102.40	1/10000	AF084530 intron
#13	-	78	418	299	42%	-96.00	1/10000	
#131	+	5	462	423	36%	-123.90	2/10000	CR749820 intron
#8	-	75	463	353	39%	-98.60	3/10000	
#117	-	65	287	208	46%	-76.46	3/10000	NAV2 intron
#34	-	68	464	342	42%	-107.89	5/10000	TOP2B intron
#99	-	70	348	264	45%	-85.60	5/10000	
#152	-	137	391	238	53%	-96.20	8/10000	
#225	+	58	480	391	36%	-100.20	10/10000	
#238	+	1	288	252	45%	-81.40	10/10000	
#47	+	127	478	348	39%	-90.61	12/10000	
#74	-	24	306	271	41%	-77.40	12/10000	
#107	-	1	458	450	49%	-146.40	13/10000	
#102	+	197	479	280	39%	-79.40	19/10000	
#133	+	20	185	164	37%	-45.41	21/10000	
#241	+	5	253	251	40%	-70.90	25/10000	
#138	+	148	383	232	53%	-84.20	26/10000	
#217	-	18	289	230	43%	-65.40	27/10000	
#30	-	1	362	304	47%	-95.50	30/10000	
#179	+	165	469	281	43%	-73.80	36/10000	DPYD intron
#201	+	162	477	288	36%	-73.42	38/10000	
#37	-	3	427	372	45%	-106.00	42/10000	VTI1A intron
#119	+	41	447	396	42%	-110.08	46/10000	
#173	-	28	358	308	48%	-94.80	51/10000	
#96	+	10	310	269	44%	-83.20	51/10000	
#236	-	118	459	324	34%	-84.73	53/10000	
#35	-	65	295	229	41%	-61.72	53/10000	
#243	+	23	464	408	40%	-111.20	55/10000	GRID2 intron
#180	+	130	466	277	41%	-73.49	59/10000	MAP4K3 intron
#151	+	6	321	282	46%	-91.30	60/10000	
#106	-	45	446	399	48%	-140.10	61/10000	
#103	+	1	160	158	41%	-48.50	65/10000	
#9	+	110	471	364	44%	-109.77	67/10000	TCERG1 exon
#129	-	33	428	374	49%	-131.20	68/10000	DNCI1 intron
#209	+	17	225	198	48%	-68.90	69/10000	
#20	-	52	467	444	47%	-146.40	70/10000	IRAK3 intron
#178	-	71	259	184	48%	-60.20	75/10000	
#147	+	22	448	398	40%	-108.06	76/10000	
#156	+	1	469	426	45%	-133.46	94/10000	
#92	-	62	428	337	42%	-96.00	98/10000	FMR2 intron
#140	+	125	428	258	47%	-85.10	98/10000	

Rank measures how similar the LF-SINE instance is relative to all others, against the coelacanth LF-SINE. Orientation is given for the strand with the lower predicted minimal folding energy (MFE), and is in reference to the LF-SINE sequence. qFrom, qTo give the coordinates of the match with respect to the LF-SINE. tLength, %GC and MFE relate to the human instance. Gene name is provided for all genic instances.