Paleozoic Protein Fossils Illuminate the Evolution of Vertebrate Genomes and Transposable Elements

Martin C. Frith[1,2,3]

[1]Artificial Intelligence Research Center, AIST, Tokyo, Japan
[2]Graduate School of Frontier Sciences, University of Tokyo, Chiba, Japan
[3]Computational Bio Big-Data Open Innovation Laboratory, AIST, Tokyo, Japan

Abstract

Genomes hold a treasure trove of protein fossils: fragments of formerly protein-coding DNA, which mainly come from transposable elements (TEs) or host genes. These fossils reveal ancient evolution of TEs and genomes, and many fossils have been exapted to perform diverse functions important for the host's fitness. However, old and highly-degraded fossils are hard to identify, standard methods (e.g. BLAST) are not optimized for this task, and few Paleozoic protein fossils have been found.

Here, a recently optimized method is used to find protein fossils in vertebrate genomes. It finds Paleozoic fossils predating the amphibian/amniote divergence from most major TE categories, including virus-related Polinton and Gypsy elements. It finds 10 fossils in the human genome (8 from TEs and 2 from host genes) that predate the last common ancestor of all jawed vertebrates, probably from the Ordovician period. It also finds types of transposon and retrotransposon not found in human before. These fossils have extreme sequence conservation, indicating exaptation: some have evidence of gene-regulatory function, and they tend to lie nearest to developmental genes. Some ancient fossils suggest "genome tectonics", where two fragments of one TE have drifted apart by up to megabases, possibly explaining gene deserts and large introns.

This paints a picture of great TE diversity in our aquatic ancestors, with patchy TE inheritance by later vertebrates, producing new genes and regulatory elements on the way. Host-gene fossils too have contributed anciently-conserved DNA segments. This paves the way to further studies of ancient protein fossils.

Introduction

Genomes contain relics of formerly protein-coding DNA, which may be functionless and neutrally evolving, or in some cases have gained new, non-protein-coding functions. Most of them are derived from either transposable elements or host genes.

Transposable elements (TEs) are parasitic, or perhaps symbiotic, DNA elements that get copied or moved from one genome location to another. They have often proliferated greatly, so that e.g. the human genome has millions of TE-derived segments comprising at least ∼50% of the genome. Most of these segments are highly-mutated fragments, no longer active TEs.

TEs have had a massive impact on the evolution of their hosts (Warren et al. 2015; Etchegaray et al. 2021). They cause mutations by their prolif-

eration, and also by ectopic recombination between TE copies, causing deletions, inversions, and duplications. This can duplicate or inactivate genes (Barsh et al. 1983; Hayakawa et al. 2001), or change their tissue-specific expression (Ting et al. 1992). Some host genes have evolved from TEs, such as the vertebrate RAG genes that generate the diverse antibodies and T cell receptors of the immune system (Kapitonov and Koonin 2015), and syncytin genes that seem to enable cell fusion in placental development (Dupressoir et al. 2005). Some DNA elements that regulate gene expression have also evolved from TEs (Ting et al. 1992; Jordan et al. 2003).

A series of studies in 2006–2007 found thousands of TE-derived non-protein-coding elements with strong evolutionary conservation in mammals (Bejerano et al. 2006; Nishihara et al. 2006; Xie et al. 2006; Kamal et al. 2006; Lowe et al. 2007; Gentles et al. 2007). They often occur in gene deserts, and nearest to developmental genes (Lowe et al. 2007). These TE insertions often predate the placental/marsupial divergence (Mesozoic), but few clearly predate the mammal/bird divergence (Paleozoic), and an exceptional handful ("at least several") were shown to predate the amniote/amphibian divergence (Bejerano et al. 2006). It is thus remarkable that a later study claimed to find 133 TE insertions predating the divergence of humans and ray-finned fish, by comparing human TE fragments found by RepeatMasker to vertebrate genome alignments (Lowe and Haussler 2012).

The boundary between TEs and viruses is blurry, and an entire field, paleovirology, is mainly based on viral insertion fossils in eukaryote genomes (Barreat and Katzourakis 2021a). The oldest viral fossils found so far seem to be Mesozoic (Suh et al. 2014; Barreat and Katzourakis 2021a).

TEs are diverse and their classification is partly arbitrary (Storer et al. 2021; Kojima 2019), but eukaryotic TEs are conventionally split into *retrotransposons* which duplicate by reverse transcription of their RNA into DNA, and *DNA transposons* which do not. Major types of retrotransposon are: LINEs ("long interspersed nuclear elements"), LTR retrotransposons (which bear long terminal repeats), YR (tyrosine recombinase) retrotransposons, and Penelope-like elements. These are further subclassified, e.g. LINEs have clades and sub-clades such as Hero, Nimb, L1, I, and CR1. Major types of DNA transposon are: DDE transposons (named after 3 key amino acids in the transposase), Cryptons (YR transposons), Helitrons, and Polintons (also called Mavericks). These are also subdivided, e.g. DDE transposons have "superfamilies" such as Academ, hAT, Kolobok, and piggyBac. Finally, non-autonomous TEs such as SINEs ("short interspersed nuclear elements") typically encode no proteins, and propagate by hijacking enzymes from autonomous TEs.

Many types of TE are patchily present and absent in closely- and distantly-related eukaryotes (Yuan and Wessler 2011; Chalopin et al. 2015). This can sometimes be explained by ordinary vertical inheritance, with multiple losses of the TE family (Fawcett and Innan 2016). Contrarily, it has been suggested that long-term vertical persistence of TEs may be rare, so their long-term persistence depends on horizontal transfer (Gilbert and Feschotte 2018). Thus, in order to understand the evolution of TE families in eukaryotes, it is valuable to know what TE types were present in ancestral eukaryotes (Fawcett and Innan 2016).

Host-gene-derived protein fossils are often called "pseudogenes". They usually arise from duplication of (part of) a gene, such that one of the two copies is either not expressed or dispensable so evolves away from its protein-coding ancestry. Many such duplications are created by reverse-transcription of mRNA to DNA (e.g. by retro-

transposon enzymes), producing intron-depleted fossils termed "processed pseudogenes". There are also non-duplicated "unitary pseudogenes", e.g. the *GULO*/*GULOP* gene/pseudogene for making vitamin C, which is non-functional in primates and guinea pigs (Nishikimi et al. 1994).

Some pseudogenes seem to have significant functions, e.g. by being transcribed into an antisense RNA regulator of its cognate gene (Korneev et al. 1999), or regulating transcription (Huang et al. 2017), or generating small interfering RNAs (Tam et al. 2008). The Xist RNA involved in X chromosome inactivation has evolved partly from a formerly protein-coding gene, and partly from TEs (Elisaphenko et al. 2008). The boundary between protein fossils and functional protein-coding genes is fuzzy: they may produce peptides with marginal contribution to the organism's fitness, in the process of gene death or resurrection (Brosius and Gould 1992; Cheetham et al. 2020).

Genetic fossils are often found by comparing a genome to a database of TE or gene sequences (Storer et al. 2021; Harrison 2021). This can be done by either DNA-to-DNA or DNA-to-protein comparison. Protein-coding DNA tends to evolve by changes that preserve the encoded amino acids or replace them with similar ones: thus highly-diverged sequences can be detected more effectively at the protein level (States et al. 1991). On the other hand, protein fossils evolve without amino-acid conservation. Thus, new TE families are often found by protein-level matches to distantly-related families, whereas relics of known TE families are best detected by DNA-level matches to a model approximating the family's most-recent active ancestor. RepeatMasker files of such DNA-level matches are available for many genomes (Smit et al. 2015).

Protein-level matches have usually been sought with BLAST (Altschul et al. 1997), which is not optimized for fossils. Central to sequence matching methods are parameters defining the (dis)favorability of substitutions and gaps, which provide the definition of similarity. BLAST uses a 20×20 amino acid substitution matrix (BLOSUM or PAM), which is based on substitution rates in living proteins, so is likely suboptimal for fossils.

Therefore, we recently developed a new DNA-to-protein matching method (Yao and Frith 202x), implemented in LAST (`https://gitlab.com/mcfrith/last`). Its main advantage is that it sets the substitution, gap, and frameshift parameters by maximum-likelihood fit to given sequence data. It uses a richer 64×21 substitution matrix, allowing e.g. preferred matching of asparagine (encoded by `aac` or `aat`) to `agc` than to `tca`, which both encode serine. It judges homology based on not just one alignment, but on many alternative ways of aligning the putative homologs. This proved more sensitive than BLAST for finding human TE protein fossils, and for the first time it found YR retrotransposon fossils in the human genome (Yao and Frith 202x).

Here, this method is used to find new protein fossils in human and slowly-evolving Lagerstätte genomes: alligator, turtle, coelacanth (a lobe-finned fish closely related to land vertebrates), and chimaera (a non-bony cartilaginous fish); and also frog due to its intermediate phylogenetic position (table 1). The number of new fossils is relatively small, but they are especially ancient and include types of TE not found in human before. They thus illuminate the evolutionary history of TE content, and reveal strongly conserved ancient exaptations, including of host-gene fossils.

Results and Discussion

Protein-Fossil-Finding Pipeline

For each organism, homologous segments were found between the genome and a set of protein sequences comprising TE proteins from RepeatMasker plus proteins encoded by host genes of that organ-

Table 1. Genome versions and TE protein fossils

	organism	genome assembly (from NCBI or UCSC)	RepeatMasker version (source)	TE fossils	of which novel*	(%)
human	*Homo sapiens*	UCSC `hg38.analysisSet`	4.0.7 (UCSC)	546821	1641	(0.3)
alligator	*Alligator mississippiensis*	`ASM28112v4`	4.0.6 (NCBI)	410092	46065	(11)
turtle	*Chrysemys picta bellii*	`Chrysemys_picta_bellii-3.0.3`	4.0.6 (NCBI)	430459	63301	(15)
frog	*Xenopus tropicalis*	`UCB_Xtro_10.0`	4.0.8 (NCBI)	135507	14837	(11)
coelacanth	*Latimeria chalumnae*	UCSC `latCha1`	4.0.5 (rmsk†)	286944	279710	(97)
chimaera	*Callorhinchus milii*	UCSC `calMil1`	4.0.3 (UCSC)	105995	31098	(29)

*Not found by this version of RepeatMasker

†repeatmasker.org

ism. When multiple homologies overlapped in the genome, only the strongest was kept, to avoid homologies between different types of TE or between TEs and host genes. Homologies overlapping annotated protein-coding segments of the genome were removed. Finally, host-gene homologies were discarded if they overlapped TEs annotated by RepeatMasker: this removes true-but-unwanted homologies due to host gene protein-coding segments that evolved from e.g. SINEs. The resulting fossils, including a genome browser hub, are available at `https://github.com/mcfrith/protein-fossils`.

The homology search used a significance threshold of one expected random match to the set of proteins per 10^9 bp, so the human genome would have ∼3 matches if the sequences were perfectly random. The real false-positive rate will be higher, and was estimated by comparing the reversed (but not complemented) human genome to the proteins, producing 19 spurious matches (Yao and Frith 202x).

New TE Fossils

The number of TE protein fossils found per genome ranges from ∼100 000 to ∼500 000, most of which correspond to known TE fragments in public RepeatMasker files (table 1). The human genome has especially few new TE fossils, indicating how thoroughly human TEs have been analyzed. The coelacanth fossils are almost all new relative to the RepeatMasker annotations, simply because those annotations have very few TE types, illustrating that TE analysis is lacking for some genomes at any snapshot in time (Sotero-Caio et al. 2017).

Table 2. Classifying unknown repeats

unknown repeat	TE type
REP-2_CPB	CR1 (LINE)
REP-3_CPB	L2 (LINE)
REP-6_CPB	CR1 (LINE)
REP-21_CPB	CR1 in PIF/Harbinger
REP-22_CPB	hAT-Tag1
REP-28_CPB	CR1 (LINE)
REP-31_CPB	Gypsy (LTR)
SAT-928_Crp	Penelope
UCON84	PIF/Harbinger

Classifying Unknown Repeats

RepeatMasker genome annotations include repeats of unknown type, which might not be TEs. In alligator and turtle (but not the other genomes), some of these unknown repeats could be classified based on large and consistent overlaps with TE protein fossils (table 2). One of these repeats, UCON84, also occurs in the human genome: it is derived from a DDE transposon in the PIF/Harbinger superfamily (fig. 1). The UCON84 consensus sequence, obtained from Dfam (Storer et al. 2021), has shorter and weaker (but significant) homology to PIF/Harbinger proteins (not shown). The consensus is expected to approximate an ancestral sequence and thus have clearer homology, but it is hard to make an accurate consensus of ancient fragments.

An interesting case is REP-21_CPB in turtle: its central region consistently overlaps CR1 LINE protein matches, while the left and right flanks con-

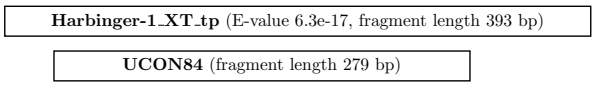

Fig. 1. Overlap between a TE protein fossil (upper box) and a repeat of unknown type (lower box) in the alligator genome (at coordinate 15 466 729 in `NW_017707593.1`).

sistently overlap PIF/Harbinger matches. Thus, it seems that a CR1 fragment was inserted in a PIF/Harbinger transposon, then this hybrid element proliferated, producing thousands of copies. We may speculate that it proliferated as a typical non-autonomous DNA transposon, with damaged protein-coding sequence but intact termini recognized by the transposase.

Inter-Genome Homology

The age of genetic fossils can be inferred by comparing different genomes. For example, fig. 2 shows a human TE fossil aligned to an L1 LINE protein, alongside mammal genome alignments from the UCSC genome browser (Kent et al. 2002; Harris 2007). This L1 insertion is present in various monkey genomes but absent from bushbaby and other placental mammals, showing that the insertion occurred in a common ancestor of monkeys. It is thus curious that the L1 insert is aligned to two marsupial genomes, opossum and tasmanian devil. Marsupials also have L1s, and these marsupial regions are indeed annotated as L1s by RepeatMasker. Thus, these human and marsupial *inserts* are true homologs, because all L1s share common ancestry, but the *insertions* are not homologous: not descended from a common ancestral insertion. The inserts might even be orthologs, if their common ancestor is no older than the placental/marsupial divergence.

Why, then, do these marsupial alignments extend into flanking sequence beyond the insert? It is hard to determine the precise endpoint of homology between distantly-related sequences: alignments

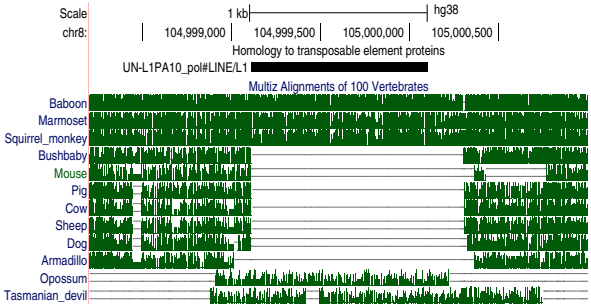

Fig. 2. A TE protein fossil in human chromosome 8, with confusing inter-genome homology. Black bar near top: alignment of an L1 LINE protein. Green tracks: alignments between the human and other genomes. Screen shot from `http://genome.ucsc.edu`.

overshoot or undershoot. These human-marsupial alignments were made with the HoxD55 substitution matrix and gap parameters that are prone to large overshoots (Frith et al. 2008).

For this study, new pair-wise genome alignments were made, by finding homologous regions (Frith and Noé 2014) and cutting them down to most-similar one-to-one alignments (Frith and Kawaguchi 2015). This tends to find higher-similarity alignments than those from UCSC and elsewhere, indicating that a higher fraction of the alignments are orthologous (Frith and Kawaguchi 2015). This probably does not avoid non-homologous TE insertions, so a new step was added: isolated alignments were discarded, by only keeping groups of alignments that are nearby in both genomes. Some examples are in fig. 3: each panel shows one TE fossil in the human genome (central vertical stripe) that overlaps an inter-genome alignment (diagonal lines/dots). The alignments are not isolated: they are flanked by other alignments, indicating homology of not just the TE insert but also the flanking regions. Because these are distantly-related genomes, most of the DNA lacks similarity and is unaligned. The alignable fragments are probably conserved by natural selection.

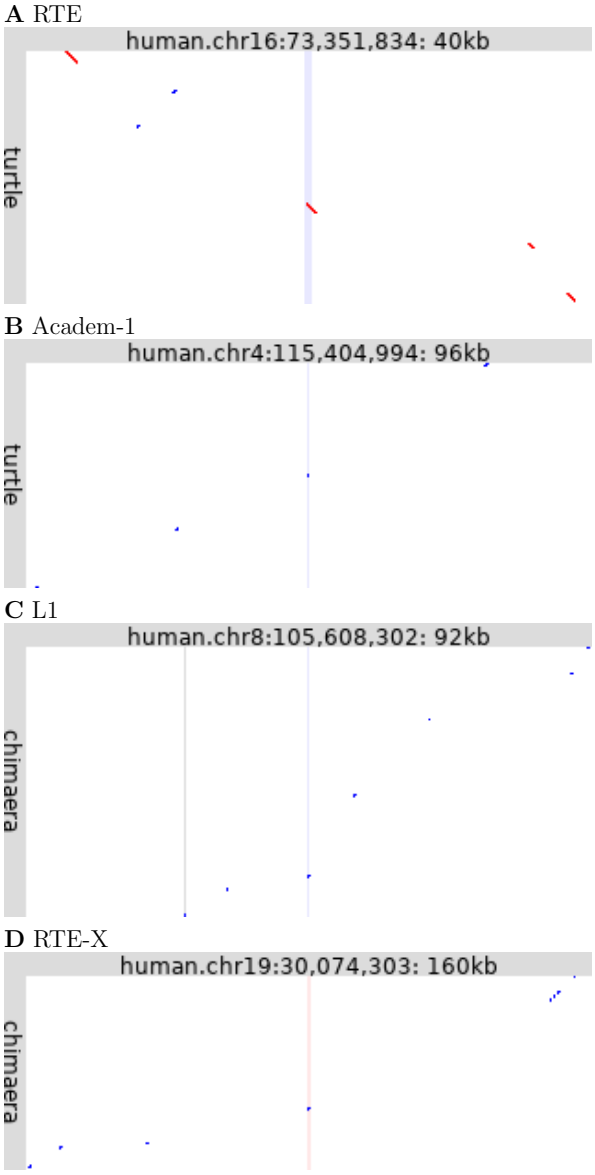

Fig. 3. Ancient conserved TE insertions. Each panel shows alignments between part of the human genome (horizontal) and turtle (**A**, **B**) or chimaera (**C**, **D**). Red dots indicate same-strand alignments, blue dots opposite-strand alignments. The central vertical lines show the location in human of the TE fossil (pink: forward strand, blue: reverse strand). The vertical gray line in panel **C** shows a protein-coding exon of *ZFPM2*.

A possible objection is that these examples might be independent insertions of an abundant TE into homologous regions of two genomes. This cannot be ruled out, but the key point is that these alignments are not only homologies but most-similar one-to-one homologies: it would be a strong coincidence for these single-best matches to independently be in homologous regions.

TE Types Newly Found in Human

The human TE protein fossils include several types of TE that have not been found in human before (table 3). These are all LINEs or DDE transposons, and are in addition to the first human YR retrotransposons (DIRS and Ngaro) and first-but-one Polintons we recently reported (Yao and Frith 202x). Some were found directly in human, others were found in another genome and mapped to human via the inter-genome alignments ("found in" column). The E-value indicates significance/confidence of the DNA-protein homology: it is the expected number of times to find such a similarity between the genome and the set of proteins, if they were random sequences. Some of the E-values are quite high, indicating lower confidence. On the other hand, most of these putative DNA-protein homologies overlap human/non-mammal genome alignments, which would be a strong coincidence if they were random similarities (fig. 3).

Some of these TE fossils are shown in more detail in fig. 4. It can be seen that they lie in human genome regions conserved in non-mammals, and are not annotated by RepeatMasker. These regions have strong evolutionary conservation in mammals according to phastCons (Siepel et al. 2005), independent of their conservation in non-mammals. Some of these fossils overlap candidate regulatory elements or known transcription factor binding sites (Moore et al. 2020; Lesurf et al. 2016): the Hero fossil in fig. 4B overlaps a CEBPB binding site, and the RTE fossil in fig. 4C overlaps binding sites for GATA2, STAT1, JUND, FOS, and JUN. Fig. 4A shows two Nimb fragments that coincide with conserved DNA

Table 3. TE fossils of types newly found in human

type	chrom-osome	start	nearest gene	intergene length	intron (kb)	found in	E-value	age
Retrotransposons								
I	3	139262204	*MRPS22*	299		alligator	0.026	amniote
I	9	32218473	*ACO1*	3171		turtle	0.97	amniote
Nimb	2	4986057	*SOX11*	1844		alligator	0.027	amniote
Nimb	3	70169802	*MDFIC2*	226		turtle	0.014	amniote
Nimb	10	76759130	*KCNMA1*	309		alligator	1.6	amniote
Nimb	13	53319346	*OLFM4*	4089		human	3.1e-05	tetrapod
Nimb	X	7635163	*VCX*	488		alligator	0.00012	amniote
Nimb	X	87287424	*KLHL4*	685		human	5.6e-14	amniote
Nimb	X	87289312	*KLHL4*	685		human	2.5	tetrapod
L2-Daphne	15	76184366	*TMEM266*		16	alligator	5.8e-05	amniote
L2-Kiri	3	157982382	*SHOX2*	592		turtle	0.012	amniote
L2-Kiri	16	53518165	*AKTIP*	95		turtle	0.053	amniote
L2-Kiri	18	25027321	*ZNF521*	582		alligator	0.0012	amniote
R2-Hero	2	118705507	*EN1*	731		alligator	0.48	amniote
R2-Hero	4	13163078	*RAB28*	1939		alligator	1.3e-06	amniote
R2-Hero	6	72553710	*KCNQ5*	219		turtle	6.5e-05	tetrapod
R2-Hero	7	36766591	*AOAH*	128		alligator	6.9e-06	amniote
R2-Hero	8	71079407	*EYA1*	461		turtle	0.064	amniote
R2-Hero	8	76862008	*ZFHX4*		7	alligator	4.4e-12	amniote
R2-Hero	11	91081544	*CHORDC1*	2002		human	0.0099	amniote
R2-Hero	14	53387314	*DDHD1*	796		alligator	3.2e-09	amniote
R2-Hero	15	67559478	*MAP2K5*		13	human	3.4e-06	amniote
R2-Hero	X	31319778	*DMD*		57	human	2.3e-08	amniote
RTE	1	88450891	*PKN2*	1335		human	0.74	.
RTE	1	216744731	*ESRRG*		82	human	0.2	amniote
RTE	2	198403329	*PLCL1*	1120		human	0.0067	amniote
RTE	2	204082119	*ICOS*	584		human	2.2e-13	amniote
RTE	3	67685630	*SUCLG2*	337		alligator	5e-05	amniote
RTE	3	169026953	*MECOM*	988		alligator	0.0053	amniote
RTE	3	172228179	*FNDC3B*		21	human	0.026	.
RTE	10	33939004	*PARD3*	775		turtle	3.3e-12	amniote
RTE	10	82724195	*NRG3*		317	alligator	2e-07	amniote
RTE	13	58819808	*DIAPH3*	1936		alligator	0.029	amniote
RTE	16	73371834	*ZFHX3*		138	human	9.3e-16	amniote
RTE	X	97057978	*DIAPH2*		108	human	0.07	.
Rex1/Babar	4	34782619	*ARAP2*	4919		turtle	5.2e-13	amniote
Rex1/Babar	4	130232079	*C4orf33*	4033		alligator	0.019	amniote
DNA Transposons								
Academ-1	4	115452994	*NDST4*	1970		turtle	0.0067	amniote
EnSpm	2	129063904	*HS6ST1*	1661		turtle	0.0008	amniote
EnSpm	2	180729359	*UBE2E3*	973		alligator	0.04	amniote
EnSpm	3	180734783	*CCDC39*	233		human	2	amniote
Ginger1	3	147687919	*ZIC1*	1281		alligator	1.2e-20	amniote
hAT19	1	3746554	*CCDC27*	16		coelacanth	1.5	sarcopterygian
hAT19	2	104030381	*POU3F3*	2036		human	7.6e-08	.
hAT19	2	143799848	*ARHGAP15*	170		human	3.2e-21	amniote
hAT19	4	34434122	*ARAP2*	4919		human	5.1e-11	.
hAT19	7	57524151	*ZNF716*	6572		human	3.8e-14	.
hAT19	16	75903914	*CPHXL*	551		alligator	3.5e-11	amniote
hAT19	20	26187394	*ZNF337*	5561		human	3.1e-08	.
hAT5	18	38666092	*CELF4*	4389		alligator	7.1e-05	tetrapod

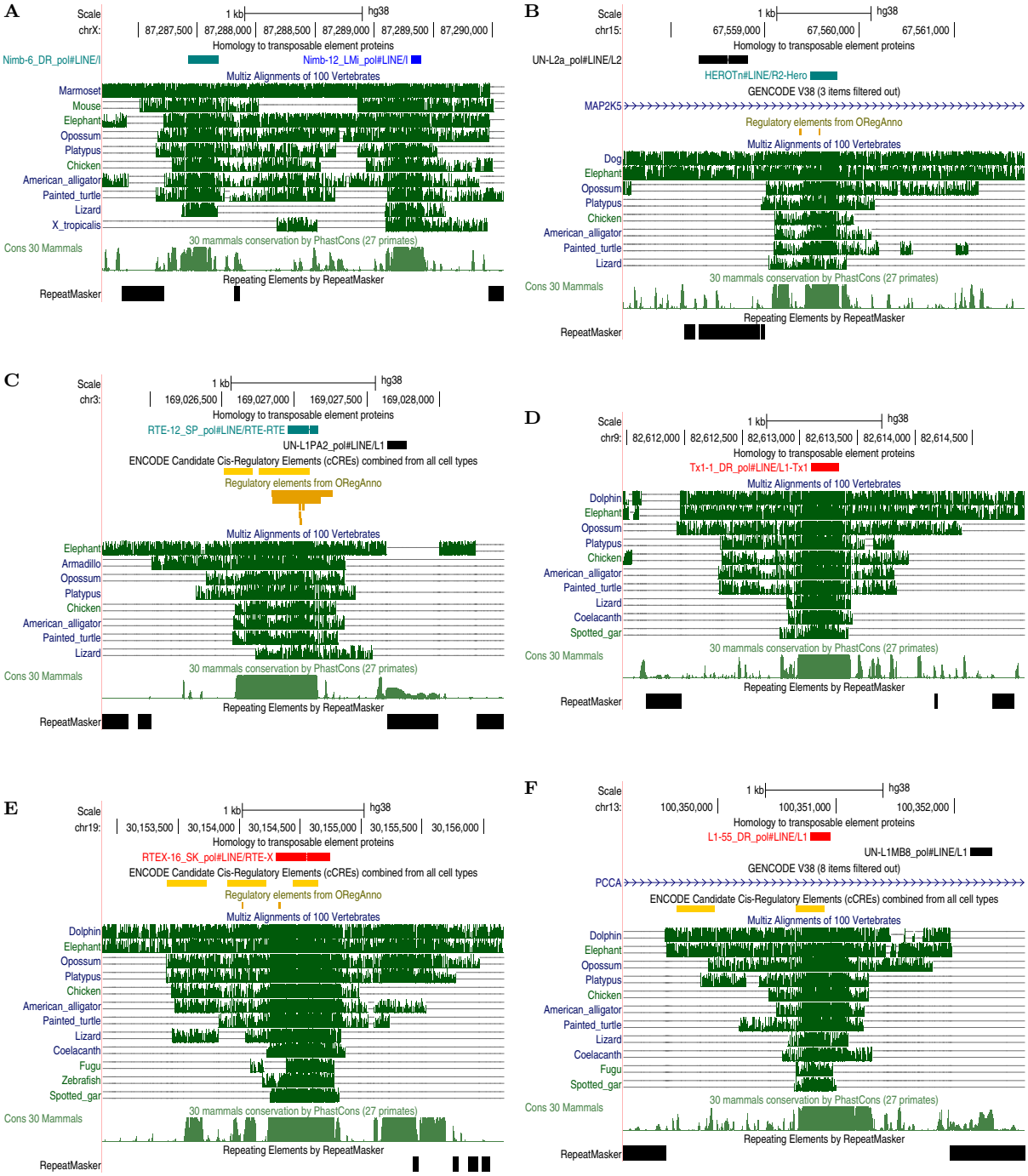

Fig. 4. Ancient conserved TE insertions in the human genome. Each panel shows, from top to bottom: TE protein fossils, alignments of the human genome to other vertebrate genomes, evolutionary conservation in mammals (phastCons), and repeats found by RepeatMasker. Some panels also show annotations of regulatory elements and Gencode genes (introns). Screen shots from `http://genome.ucsc.edu`.

segments: presumably they come from one Nimb insertion, which predates the amniote/amphibian divergence. (Only one of these Nimb fragments is aligned to frog: the other may be deleted or not detected in frog.)

These fossils clarify the historical presence of TE types in vertebrates. They make presence of several TE types less patchy among vertebrates, thus explicable by vertical inheritance rather than horizontal transfer. Nimb-type LINEs have been found in insects, mollusks, teleost (bony) fish (Kapitonov et al. 2009; Chalopin et al. 2015), and turtle (Smit et al. 2015): here Nimb relics are found from ancient tetrapods, and also in coelacanth. This makes the presence of Nimb in vertebrates less patchy, and suggests vertical inheritance from the common ancestor of bony vertebrates. The Hero clade was found in sea urchin, lancelet, and fish (Kojima and Fujiwara 2004; Kapitonov et al. 2009): its presence in ancient tetrapods fits with vertical inheritance from deuterostome ancestors. Hero LINEs are unusual in having a restriction-like endonuclease (Kojima and Fujiwara 2004), unlike all other human LINE relics except Mam_R4. The I clade was previously found in fish and some invertebrates (Kapitonov et al. 2009): here hundreds are found in turtle and a few in alligator. Daphne was previously found in sea urchin and arthropods (Schön and Arkhipova 2006), plus lancelet and zebrafish (Smit et al. 2015): here 67 fragments are found in coelacanth, 6 in chimaera, 6 in turtle, and 4 in alligator, rounding out its historical presence in vertebrates. The Rex1/Babar clade has been found patchily in non-sarcopterygian fish excluding chimaera, plus frog and lizard (Chalopin et al. 2015; Smit et al. 2015): here it is found in ancestral amniotes and also coelacanth and chimaera, rendering its distribution non-patchy.

RepeatMasker distinguishes two types of RTE-like LINE, BovB and RTE: it finds only BovB in human, and it finds RTE in turtle and zebrafish. Previous reports of RTE in human seem to be BovB elements that were not classified separately (Kojima 2018). This study finds RTEs from amniote ancestors, and thousands in coelacanth, again suggesting vertical inheritance from ancestors of bony vertebrates.

These fossils also provide support for TE origin of genes. Ginger1 transposons were previously found in some invertebrates including lancelet (Bao et al. 2010), but not in sarcopterygians (Yuan and Wessler 2011; Chalopin et al. 2015). Their relics are found here in alligator, turtle, coelacanth, and many in frog. This makes it more plausible that the human *GIN1* gene was indeed exapted from Ginger1 in ancestral amniotes (Bao et al. 2010). At least one pre-amniote Ginger1 relic was also exapted for non-protein-coding function (table 3). Similarly, hAT19 fossils from amniote ancestors support the hAT19 origin of the amniote-specific gene *CGGBP1* (Yellan et al. 2021), which binds CGG repeats and regulates gene expression (Singh and Westermark 2015). hAT19 fragments have been exapted for non-protein-coding functions too.

In the four tetrapod genomes just one hAT5 fragment is found, which is conserved in all of them: the single exapted relic of an ancient hAT5 infection. hAT5 was previously found in some invertebrates (Putnam et al. 2007) and fish (Smit et al. 2015), and is unusual in having 5 bp TSDs (target site duplications), whereas all previously-known hATs have 8 bp TSDs (Putnam et al. 2007).

Anciently Conserved TE Fossils

The human genome contains diverse TE protein fossils that are older than the amniote/amphibian divergence (table 4). It is striking that they include nearly all major types of TE: LINEs, Penelope elements, LTR retrotransposons (Gypsy), YR retrotransposons (DIRS), DDE transposons, a Crypton, and Polintons. Eight of them (7 LINEs and a Crypton) are shared by human and chimaera, making

Table 4. Pre-tetrapod TE fossils found in human

type	chrom-osome	start	nearest gene	intergene length (kb)	intron	found in	E-value	age
Retrotransposons								
CR1	4	13161246	*RAB28*	1939		alligator	1.9e-06	tetrapod
CR1	4	111235163	*PITX2*	1503		chimaera	0.0078	gnathostome
CR1	5	94809895	*MCTP1*		69	alligator	0.0049	tetrapod
CR1	6	98370405	*POU3F2*	1551		turtle	0.022	tetrapod
CR1	16	78813330	*WWOX*		779	turtle	0.022	tetrapod
CR1	18	72559688	*CBLN2*	198		alligator	9.9e-08	tetrapod
Nimb	13	53319346	*OLFM4*	4089		human	3.1e-05	tetrapod
Nimb	X	87289312	*KLHL4*	685		human	2.5	tetrapod
L1	3	70416318	*MDFIC2*	642		coelacanth	0.61	gnathostome
L1	8	105654302	*ZFPM2*		154	human	1	gnathostome
L1	9	2050774	*SMARCA2*		7	human	0.093	tetrapod
L1	13	100350784	*PCCA*		28	human	0.035	gnathostome
L1	18	55784184	*TCF4*	961		alligator	6e-07	tetrapod
L1-Tx1	9	82613098	*RASEF*	984		human	9.6e-05	gnathostome
L1-Tx1	10	129519611	*MGMT*		69	coelacanth	3e-16	gnathostome
L2	5	166711049	*TENM2*	3554		turtle	0.4	tetrapod
L2	6	45881206	*CLIC5*	347		alligator	0.032	tetrapod
L2*	7	108869078	*DNAJB9*	2088		human	2.9e-25	tetrapod
L2-Crack	1	216137121	*USH2A*		77	alligator	3.5e-08	tetrapod
L2-Crack	5	109502097	*PJA2*	280		human	0.00013	tetrapod
R2-Hero	6	72553710	*KCNQ5*	219		turtle	6.5e-05	tetrapod
RTE-X	19	30154303	*ZNF536*	209		human	0.00078	gnathostome
Penelope	4	187183800	*FAT1*	1272		alligator	0.05	tetrapod
Penelope	13	106429488	*EFNB2*	999		human	0.00017	tetrapod
Gypsy	1	38669733	*RRAGC*	791		human	5.9e-11	tetrapod
Gypsy	2	57598618	*VRK2*	1521		alligator	8.3e-12	tetrapod
Gypsy	19	30111148	*URI1*	209		human	0.017	tetrapod
Gypsy	X	98972571	*PCDH19*	2687		human	4.6e-23	tetrapod
DIRS	3	55911374	*ERC2*		62	alligator	1.1e-08	tetrapod
DIRS	9	13728242	*NFIB*	802		human	0.0037	tetrapod
DIRS	9	20189423	*MLLT3*	553		turtle	0.86	tetrapod
DIRS	10	16734086	*RSU1*		57	alligator	0.84	tetrapod
DIRS	16	78152629	*WWOX*		49	turtle	3.2e-17	tetrapod
DNA Transposons								
PIF/Harbinger	2	145279820	*ZEB2*	3324		human	4.4e-05	tetrapod
PIF/Harbinger	2	176676027	*MTX2*	875		human	0.0012	tetrapod
hAT-Blackjack	7	14272601	*DGKB*		160	alligator	6.5e-10	tetrapod
hAT-Tip100	4	4696650	*STX18*	317		alligator	0.0021	tetrapod
hAT-Tip100	4	129428560	*C4orf33*	4033		human	1.8e-11	tetrapod
hAT19	1	3746554	*CCDC27*	16		coelacanth	1.5	sarcopterygian
hAT5	18	38666092	*CELF4*	4389		alligator	7.1e-05	tetrapod
Crypton-A	12	14468609	*ATF7IP*		9	alligator	2.2e-70	gnathostome
Polinton	3	114555204	*ZBTB20*		83	human	2.8	tetrapod
Polinton	20	52577239	*ZFP64*	781		turtle	4.4e-15	tetrapod
Polinton	20	55516705	*CBLN4*	1346		turtle	7.4e-11	tetrapod

*Previously found by RepeatMasker

them older than the last common ancestor of all jawed vertebrates. Three of these most ancient fossils are shown in fig. 4 D–F: their ancient exaptation is supported by their conserved presence in mammal, reptile, and bony-fish genomes, their strong conservation in mammals (phastCons), and sometimes by evidence of regulatory function.

A further 882 TE protein fossils that predate the mammal/reptile divergence were found in the human genome (table 5). Most of these (745, 84%)

Table 5. Other pre-amniote TE fossils in human

type	number	of which known*
Retrotransposons		
CR1	204	92
Dong-R4	1	0
Vingi	1	0
L1	137	22
L1-Tx1	12	0
L2	167	14
L2-Crack	25	1
BovB	28	1
RTE-X	4	0
Penelope	54	1
Gypsy	83	3
DIRS	37	0
Ngaro	28	0
DNA Transposons		
Kolobok-T2	1	0
PIF/Harbinger	18	1
PiggyBac	7	0
TcMar-Mariner	1	0
TcMar-Pogo	2	0
TcMar-Tc1	4	0
TcMar-Tigger	5	0
hAT-Ac	7	1
hAT-Blackjack	12	1
hAT-Charlie	6	0
hAT-Tip100	21	0
Crypton-A	3	0
Polinton	14	0

*Previously found by RepeatMasker

are novel (not annotated by RepeatMasker), as are all but one of the pre-tetrapod fossils (table 4). These ancient TE fossils are often in megabase-scale gene deserts or large ($\sim 10^5$ bp) introns (table 3, 4). The nearest genes are significantly enriched in developmental functions such as nervous system development, cell morphogenesis, and axonogenesis (Mi et al. 2021).

These TE relics from ancient vertebrates help us to understand the ancestral mobilome, which has been difficult, especially since TEs might have been horizontally transferred (Chalopin et al. 2015). For example, it has been suggested that mammal L1s were introduced by horizontal transfer into a common ancestor of therian (live-bearing) mammals (Ivancevic et al. 2018). We now have direct evidence that L1-like TEs were present in a common ancestor of jawed vertebrates, and hundreds of L1 fragments predate the amniote divergence (table 5). Actually, repeatmasker.org lists 353 L1 fragments in the platypus genome (`ornAna1`), so perhaps L1s were vertically inherited by mammals, but became inactive early in the monotreme lineage. There are also L1-Tx1 fossils from gnathostome ancestors (table 4): this supports the suggestion that L1 clades including Tx1 diverged in a common ancestor of mammals and fish (Ichiyanagi et al. 2007), which was not certain since Tx-like L1s are prone to horizontal transfer between marine hosts (Ivancevic et al. 2018).

For other TE types too – DIRS, Polinton, and PIF/Harbinger – their previously-noted patchiness among tetrapods (Chalopin et al. 2015) is explained by ancient loss of activity, since they were present in tetrapod ancestors. The emerging picture is that ancient vertebrates had many diverse types of TE, like present day teleost fish but unlike mammals or birds (Chalopin et al. 2015).

The pre-amniote BovB fossils (table 5) are particularly informative, because BovB has frequently been horizontally transferred (Ivancevic et al. 2018). Interestingly, the phylogeny of BovB elements differs greatly *but not entirely* from the phylogeny of their host organisms: amniote BovBs are all in a central branch of the tree and fish BovBs on outer branches (Ivancevic et al. 2018, fig. 2a). Knowing that BovBs were present in amniote ancestors, it seems likely that BovB initially entered amniotes by vertical inheritance, perhaps specifically into squamate reptiles, before being horizontally transferred among amniotes and arthropod vectors.

Regarding LTR retrotransposons, it is intriguing that ancient Gypsy-like fossils are found (table 4, 5), but ancient ERV (endogenous retrovirus) fossils are not. The origin of vertebrate retroviruses has been debated (Hayward 2017): ERVs may have evolved from Gypsy-like elements in a common ancestor of

amniotes and amphibians (Hellsten et al. 2010).

The age of the oldest Polinton insertions is greatly increased from 95 million years (Barreat and Katzourakis 2021b) to ∼350 million years (the amniote/amphibian divergence). It is possible, however, that these protein fossils are much younger than their insertions, if the intact TE benefits host fitness so remains intact (i.e. protein coding) by natural selection of the host. Intact TEs are usually thought not to benefit host fitness, but intact Polintons might protect the host from viruses, in particular iridoviruses that infect cold-blooded vertebrates (Barreat and Katzourakis 2021b). Nevertheless, the human Polinton relics are no longer intact, yet some have strong phastCons conservation in mammals indicating exaptation.

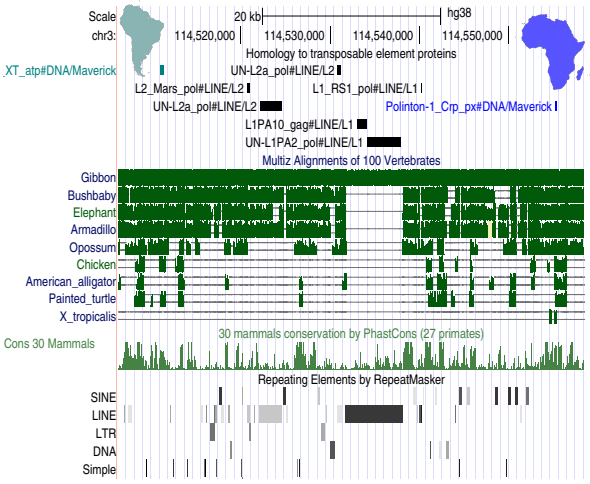

Fig. 5. Ancient Polinton/Maverick fragments in an intron of *ZBTB20* on human chromosome 3. The two fragments are labeled by South America and Africa, with younger TE fossils in between.

Conserved RepeatMasker Fossils

For sake of comparison, the age of previously known TE fossils (from RepeatMasker) was inferred in the same way. RepeatMasker includes many more TE fossils, especially non-protein-coding SINEs. It is tuned to have a false positive fraction of 0.2% (Hubley et al. 2016), which corresponds to $\sim 10^4$ false hits in the human genome. There are 133 RepeatMasker hits in human that are conserved in frog, of which 84 (63%) are especially ancient types of repeat: UCON, Eulor, LFSINE, and AmnSINE1 (Gentles et al. 2007; Bejerano et al. 2006; Nishihara et al. 2006). Most of these are unknown types of repeat, and may not be TEs. In contrast, there are 73 RepeatMasker hits in human that are conserved in coelacanth, which are not obviously enriched in ancient repeat types. They include primate-specific L1P and SVA elements, which are surely false-positive RepeatMasker annotations. A few may be real, but it is hard to know which ones or have confidence in them. Unfortunately, RepeatMasker files do not state the significance (E-value) of each hit. In summary, the oldest confident minimum age for previously-known TE insertions (apart from TE-derived genes) is the amniote/amphibian divergence (Bejerano et al. 2006).

This casts doubt on the previously reported TE insertions predating the human/teleost divergence (Lowe and Haussler 2012). Aside from false RepeatMasker hits, that study mentioned no countermeasures for non-homologous insertions (fig. 2).

The tetrapod TEs found here (table 4) are almost completely disjoint from previously known ones: the latter are mostly unknown repeat types or SINEs. The newly-found LINEs might be the autonomous counterparts of the ancient SINEs, in particular, AmnSINE1 was thought to be mobilized by an undiscovered L2-like LINE (Nishihara et al. 2006).

Genome Tectonics

Sometimes, two TE fossils of the same type lie strikingly near each other in the human genome. An example is in fig. 5: two Polinton relics are separated by 44 kb, which is remarkably close considering there are only 40 Polinton fragments in the genome. They

might come from two independent insertions into a Polinton hotspot, but a simpler explanation is that they come from one Polinton, and drifted apart due to younger TE insertions between them. It is well known that old TEs get fragmented by younger insertions, but it is interesting to consider how far apart they can drift. If there is a locally higher rate of insertion than deletion, this might over time produce large introns and gene deserts. Ancient fossils can be markers of such long term rifting. Among the pre-amniote TE fossils, there are a few hundred such pairs separated by 30–3000 kb.

Host-Gene-Derived Protein Fossils

This study found 27 240 host-gene-derived protein fossils in the human genome, of which 4303 (16%) are new: not in Gencode V37 or RefSeq pseudogenes, or RetroGenes V9 (Pruitt et al. 2014; Harrow et al. 2012; Baertsch et al. 2008). They do not overlap known protein-coding regions, but some may be unknown protein-coding exons rather than fossils. Frameshifts or premature stop codons are present in 71.3% of the new segments and 72.4% of the non-new ones, suggesting a similar (presumably low) fraction of unknown coding exons.

Ancient fossils were sought in the same way as for TEs, but there is an extra difficulty. While we may find a fossil in the human genome that overlaps an alignment to (say) chimaera, it might have encoded a functional protein for most of this evolutionary history, becoming a fossil only recently in the human lineage (Sheetlin et al. 2014). The aligned region of chimaera was also required to be noncoding, but it may have independently become a fossil, or simply be an unannotated protein-coding exon. The non-human genomes presumably have less thorough gene annotation. This was assessed by manually examining UCSC phyloP graphs showing basewise evolutionary conservation in 100 vertebrates (Pollard et al. 2010). In some cases, there was a pattern

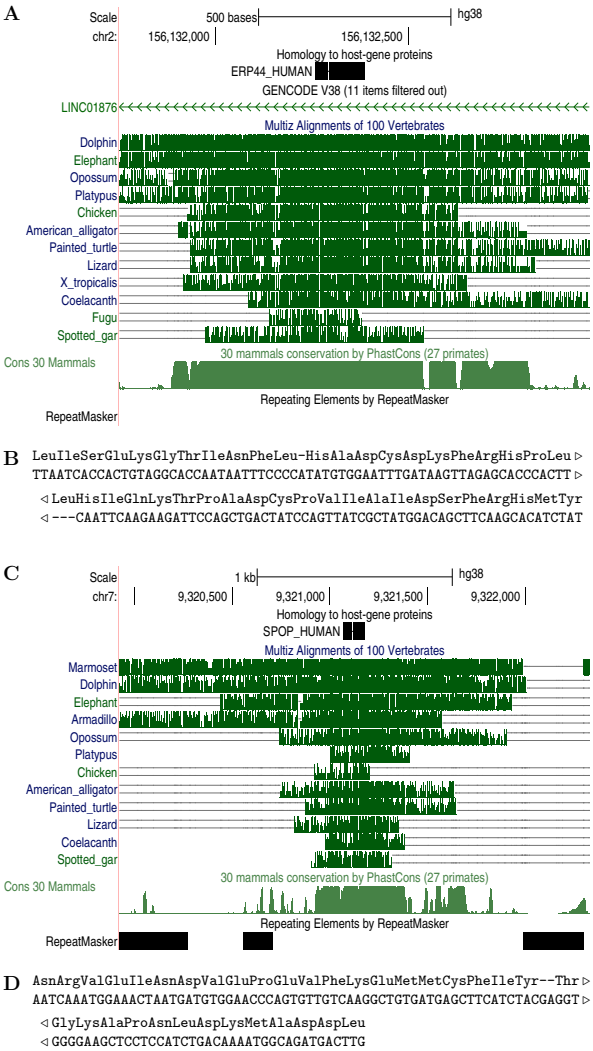

Fig. 6. Ancient conserved pseudogenes in the human genome. **A** Match between endoplasmic reticulum resident protein 44 and chromosome 2, showing conservation in vertebrates. **B** Base-level alignment of the above. **C** Match between speckle-type POZ protein and chromosome 7. **D** Base-level alignment thereof.

of every 3rd base being less conserved, indicating that natural selection conserved the encoded amino acids, for at least part of the history.

In the end, two strong candidates were found for host-gene-derived fossils predating the last common ancestor of jawed vertebrates (fig. 6). These human regions are aligned to alligator, turtle, coelacanth,

and chimaera, and are not annotated as protein-coding in any of these genomes. The DNA-protein alignments have frameshifts (fig. 6 B, D), and the basewise conservation does not suggest 3-periodicity (not shown). Their ancient conservation, and strong phastCons conservation in mammals, testifies to their exaptation for some critical but unknown function.

Conclusions and Prospects

This study greatly increases the number and variety of Paleozoic protein fossils. Fossils of most major TE categories (except Helitrons) are found that predate the amphibian/amniote divergence. The oldest fossils, from both TEs and host genes, predate the last common ancestor of jawed vertebrates. The detection of some TE types in ancestral genomes makes their distribution in vertebrates less patchy, suggesting that ancient vertebrates had a high diversity of TEs that were vertically inherited in some lineages but lost activity in others. There are hints that marine or aquatic vertebrates are prone to horizontal TE transfer (Ivancevic et al. 2018; Zhang et al. 2020; Barreat and Katzourakis 2021b), which might explain the high ancestral diversity. These ancient fossils have strong sequence conservation, indicating exaptation, and some have evidence of regulatory function. Not only TEs but also host-gene fossils were anciently exapted with strong sequence conservation. Ancient fossils can be markers of long-term genome tectonics.

It is hoped that these fossil-finding methods can easily be adapted for future studies. They are especially beneficial for finding TEs in less-studied genomes, reducing reliance on de novo repeat-finding and confusion between low copy-number TEs, multi-gene families, and TE-derived genes (Arkhipova 2017; Makałowski et al. 2019). The fitting of substitution and gap rates could perhaps be improved: here it was done naively by comparing a genome to known TE proteins. The choice of sequence data for parameter-fitting seems important for finding ancient or unknown types of fossil. Fossil-finding could also be aided by ancestralizing the genome sequence, e.g. reverting recent substitutions and TE insertions.

One promising application is paleovirology: few Mesozoic and no Paleozoic viral fossils have been found so far (Barreat and Katzourakis 2021a). If Gypsy-like elements (Metaviridae) or Polintons are counted as viruses, Paleozoic fossils predating ∼350 million years are found here (table 4).

A great challenge is to infer ancient genetic sequences from their fossil fragments, much as ancient organisms are inferred from mineral fossils. This inference might be assisted by LAST's ability to estimate the probability that each column of a sequence alignment is correct.

Materials and Methods

The pipeline scripts are available at: `https://gitlab.com/mcfrith/protein-fossils`.

Genome Data

Genome sequences and their RepeatMasker annotations were downloaded from UCSC, NCBI, or repeatmasker.org (table 1). The human RepeatMasker annotations are from UCSC's `rmskOutCurrent`.

TE protein sequences were taken from the file `RepeatPeps.lib` in RepeatMasker version 4.1.2-p1. For each non-human genome, proteins encoded by host genes were taken from NCBI's `.faa` file for that genome. For human, with the aim of getting reliable proteins, non-TE proteins with existence level 1–3 were taken from `uniprot_sprot_human.dat` in UniProt release 2021_02 (The UniProt Consortium 2020).

Protein-coding regions of the human genome were taken from the union of `wgEncodeGencodeCompV37` and `ncbiRefSeq` from UCSC (Pruitt et al. 2014;

Harrow et al. 2012). For each non-human genome, protein-coding regions were obtained from NCBI's `.gff` file for that genome.

Finding Protein Fossils

The DNA/protein substitution and gap rates were found separately for each genome, by comparing it to the TE proteins, using LAST version 1250:

```
lastdb -q -c myDB RepeatPeps.lib

last-train -P8 --codon -X1 --pid=50
                          myDB genome.fa > te.train
```

The `--pid=50` option makes it only use homologies with $\leq 50\%$ identity, with the aim of focusing on old fossils. Next, the genome was matched to TE and host-gene proteins:

```
fasta-nr hostProteins RepeatPeps.lib |
  lastdb -q -c pDB

lastal -D1e9 -K0 -m500 -p te.train
                            pDB genome.fa > aln.maf
```

Option `-D1e9` sets the significance threshold to one false hit per 10^9 bp, `-K0` omits hits that overlap stronger hits in the genome, and `-m500` makes it more slow and sensitive. (With lower values of `m`, occasionally a host-gene-derived fossil was missed and instead wrongly aligned to a TE protein.)

It turns out the RepeatMasker proteins include exapted genes: they were excluded, by omitting hits to proteins whose names contain `_HSgene`, `_Hsa_`, `UN-GIN`, or `_Xtr_eg_tp`.

Finally, alignments >10% covered by protein-coding annotation were removed, as were host-protein alignments >10% covered by RepeatMasker TE annotations other than `Low_complexity` and `Simple_repeat`.

Genome Alignments

Pair-wise genome alignments were made like this:

```
lastdb -P8 -uMAM8 gDB genome1.fa

last-train -P8 --revsym -D1e9 --sample-number=5000
                          gDB genome2.fa > g.train

lastal -P8 -D1e9 -m100 -p g.train gDB genome2.fa |
  last-split -fMAF+ > many-to-one.maf

last-split -r many-to-one.maf |
  last-postmask > one-to-one.maf
```

The `-uMAM8` and `-m100` options make it extremely slow and sensitive (Frith and Noé 2014). These one-to-one alignments are available at `https://github.com/mcfrith/last-genome-alignments`.

Next, isolated alignments were removed by defining two alignments to be "linked" if, in both genomes, they are separated by at most 10^6 bp and by at most 5 other alignments. Alignments were retained if linked, directly or indirectly, to at least two others.

Ancient Protein Fossils

A protein fossil was inferred to be ancient if it overlaps an inter-genome alignment. However, spurious overlaps are caused by the DNA-protein or inter-genome alignments overshooting beyond the end of homology: this often happens when the fossil is near a protein-coding exon. Therefore, the set of alignments between two genomes was reduced to those that do not overlap protein-coding annotations in either genome, and then each fossil was considered conserved if at least 30% of it is covered by alignments between those two genomes. This 30% threshold was determined empirically (25% was occasionally insufficient). There is likely a better way using LAST's ability to estimate the probability of each column in an alignment.

Novelty

For tables 4 and 5, a TE fossil was deemed "known" if it has non-zero same-strand overlap with a RepeatMasker genome annotation of the same "class" (DNA, LINE, LTR, etc). For table 1, a TE fossil was deemed novel if at most 10% of it is covered

by same-strand RepeatMasker TE annotations with known "class/family". A host-gene protein fossil was deemed novel if at most 10% of it overlaps same-strand known pseudogenes.

Nearest genes

The nearest genes were found from among those with `NM_` accession numbers in `ncbiRefSeqCurated` from UCSC.

Acknowledgments

I thank the Frith and Asai labs' members, past and present, for useful discussions and feedback, and Ayaka Ishiguro for a preliminary survey of protein fossils in eukaryote genomes.

References

Altschul, S. F., Madden, T. L., Schäffer, 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 research*, 25(17):3389–3402.

Arkhipova, I. R. (2017). Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mobile DNA*, 8(1):1–14.

Baertsch, R., Diekhans, M., Kent, W. J., Haussler, D., and Brosius, J. (2008). Retrocopy contributions to the evolution of the human genome. *BMC genomics*, 9(1):1–19.

Bao, W., Kapitonov, V. V., and Jurka, J. (2010). Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. *Mobile DNA*, 1(1):1–10.

Barreat, J. G. N. and Katzourakis, A. (2021a). Paleovirology of the DNA viruses of eukaryotes. *Trends in Microbiology*.

Barreat, J. G. N. and Katzourakis, A. (2021b). Phylogenomics of the *Maverick* virus-like mobile genetic elements of vertebrates. *Molecular biology and evolution*, 38(5):1731–1743.

Barsh, G. S., Seeburg, P. H., and Gelinas, R. E. (1983). The human growth hormone gene family: structure and evolution of the chromosomal locus. *Nucleic acids research*, 11(12):3939–3958.

Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., Rubin, E. M., Kent, W. J., and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441(7089):87–90.

Brosius, J. and Gould, S. J. (1992). On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proceedings of the National Academy of Sciences*, 89(22):10706–10710.

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, 7(2):567–580.

Cheetham, S. W., Faulkner, G. J., and Dinger, M. E. (2020). Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nature Reviews Genetics*, 21(3):191–201.

Dupressoir, A., Marceau, G., Vernochet, C., Bénit, L., Kanellopoulos, C., Sapin, V., and Heidmann, T. (2005). Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. *Proceedings of the National Academy of Sciences*, 102(3):725–730.

Elisaphenko, E. A., Kolesnikov, N. N., Shevchenko, A. I., Rogozin, I. B., Nesterova, T. B., Brockdorff, N., and Zakian, S. M. (2008). A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. *PloS one*, 3(6):e2521.

Etchegaray, E., Naville, M., Volff, J.-N., and Haftek-Terreau, Z. (2021). Transposable element-derived sequences in vertebrate development. *Mobile DNA*, 12(1):1–24.

Fawcett, J. A. and Innan, H. (2016). High similarity between distantly related species of a plant SINE family is consistent with a scenario of vertical transmission without horizontal transfers. *Molecular biology and evolution*, 33(10):2593–2604.

Frith, M. C. and Kawaguchi, R. (2015). Split-alignment of genomes finds orthologies more accurately. *Genome biology*, 16(1):1–17.

Frith, M. C. and Noé, L. (2014). Improved search heuristics find 20 000 new alignments between human and mouse genomes. *Nucleic acids research*, 42(7):e59–e59.

Frith, M. C., Park, Y., Sheetlin, S. L., and Spouge, J. L. (2008). The whole alignment and nothing but the alignment: the problem of spurious alignment flanks. *Nucleic acids research*, 36(18):5863–5871.

Gentles, A. J., Wakefield, M. J., Kohany, O., Gu, W., Batzer, M. A., Pollock, D. D., and Jurka, J. (2007). Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome research*, 17(7):992–1004.

Gilbert, C. and Feschotte, C. (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current opinion in genetics & development*, 49:15–24.

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. PhD thesis, The Pennsylvania State University.

Harrison, P. M. (2021). Computational methods for pseudogene annotation based on sequence homology. In *Pseudogenes*, pages 35–48. Springer.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome research*, 22(9):1760–1774.

Hayakawa, T., Satta, Y., Gagneux, P., Varki, A., and Takahata, N. (2001). Alu-mediated inactivation of the human CMP-N-acetylneuraminic acid hydroxylase gene. *Proceedings of the National Academy of Sciences*, 98(20):11399–11404.

Hayward, A. (2017). Origin of the retroviruses: when, where, and how? *Current opinion in virology*, 25:23–27.

Hellsten, U., Harland, R. M., Gilchrist, M. J., Hendrix, D., Jurka, J., Kapitonov, V., Ovcharenko, I., Putnam, N. H., Shu, S., Taher, L., et al. (2010). The genome of the western clawed frog *Xenopus tropicalis*. *Science*, 328(5978):633–636.

Huang, P., Keller, C. A., Giardine, B., Grevet, J. D., Davies, J. O., Hughes, J. R., Kurita, R., Nakamura, Y., Hardison, R. C., and Blobel, G. A. (2017). Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. *Genes & development*, 31(16):1704–1713.

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F., and Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic acids research*, 44(D1):D81–D89.

Ichiyanagi, K., Nishihara, H., Duvernell, D. D., and Okada, N. (2007). Acquisition of endonuclease specificity during evolution of L1 retrotransposon. *Molecular biology and evolution*, 24(9):2009–2015.

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., and Adelson, D. L. (2018). Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome biology*, 19(1):1–13.

Jordan, I. K., Rogozin, I. B., Glazko, G. V., and Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics*, 19(2):68–72.

Kamal, M., Xie, X., and Lander, E. S. (2006). A large family of ancient repeat elements in the human genome is under strong selection. *Proceedings of the National Academy of Sciences*, 103(8):2740–2745.

Kapitonov, V. V. and Koonin, E. V. (2015). Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biology direct*, 10(1):1–8.

Kapitonov, V. V., Tempel, S., and Jurka, J. (2009). Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene*, 448(2):207–213.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, 12(6):996–1006.

Kojima, K. K. (2018). Human transposable elements in Repbase: genomic footprints from fish to humans. *Mobile DNA*, 9(2).

Kojima, K. K. (2019). Structural and sequence diversity of eukaryotic transposable elements. *Genes & Genetic Systems*, 94(6):233–252.

Kojima, K. K. and Fujiwara, H. (2004). Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. *Molecular biology and evolution*, 21(2):207–217.

Korneev, S. A., Park, J.-H., and O'Shea, M. (1999). Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *Journal of Neuroscience*, 19(18):7711–7720.

Lesurf, R., Cotto, K. C., Wang, G., Griffith, M., Kasaian, K., Jones, S. J., Montgomery, S. B., Griffith, O. L., and Consortium, O. R. A. (2016). ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic acids research*, 44(D1):D126–D132.

Lowe, C. B., Bejerano, G., and Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the national academy of sciences*, 104(19):8005–8010.

Lowe, C. B. and Haussler, D. (2012). 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS One*, 7(8):e43128.

Makałowski, W., Gotea, V., Pande, A., and Makałowska, I. (2019). Transposable elements: Classification, identification, and their use as a tool for comparative genomics. In *Evolutionary Genomics*, pages 177–207. Springer.

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., and Thomas, P. D. (2021). PANTHER

version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic acids research*, 49(D1):D394–D403.

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710.

Nishihara, H., Smit, A. F., and Okada, N. (2006). Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome research*, 16(7):864–874.

Nishikimi, M., Fukuyama, R., Minoshima, S., Shimizu, N., and Yagi, K. (1994). Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *Journal of Biological Chemistry*, 269(18):13685–13688.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., et al. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic acids research*, 42(D1):D756–D763.

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V. V., et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, 317(5834):86–94.

Schön, I. and Arkhipova, I. R. (2006). Two families of non-LTR retrotransposons, Syrinx and Daphne, from the Darwinulid ostracod, *Darwinula stevensoni*. *Gene*, 371(2):296–307.

Sheetlin, S. L., Park, Y., Frith, M. C., and Spouge, J. L. (2014). Frameshift alignment: statistics and post-genomic applications. *Bioinformatics*, 30(24):3575–3582.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050.

Singh, U. and Westermark, B. (2015). CGGBP1–an indispensable protein with ubiquitous cytoprotective functions. *Upsala journal of medical sciences*, 120(4):219–232.

Smit, A., Hubley, R., and Green, P. (2013–2015). RepeatMasker open-4.0. `http://www.repeatmasker.org`.

Sotero-Caio, C. G., Platt, R. N., Suh, A., and Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome biology and evolution*, 9(1):161–177.

States, D. J., Gish, W., and Altschul, S. F. (1991). Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods*, 3(1):66–70.

Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., and Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, 12(1):1–14.

Suh, A., Weber, C. C., Kehlmaier, C., Braun, E. L., Green, R. E., Fritz, U., Ray, D. A., and Ellegren, H. (2014). Early Mesozoic coexistence of amniotes and Hepadnaviridae. *PLoS genetics*, 10(12):e1004559.

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., et al. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–538.

The UniProt Consortium (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489.

Ting, C.-N., Rosenberg, M., Snow, C., Samuelson, L., and Meisler, M. (1992). Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes & development*, 6(8):1457–1465.

Warren, I. A., Naville, M., Chalopin, D., Levin, P., Berger, C. S., Galiana, D., and Volff, J.-N. (2015). Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Research*, 23(3):505–531.

Xie, X., Kamal, M., and Lander, E. S. (2006). A family of conserved noncoding elements derived from an ancient transposable element. *Proceedings of the National Academy of Sciences*, 103(31):11659–11664.

Yao, Y. and Frith, M. C. (202x). Improved DNA-versus-protein homology search for protein fossils.

Yellan, I., Yang, A. W., and Hughes, T. R. (2021). Diverse eukaryotic CGG-binding proteins produced by independent domestications of hAT transposons. *Molecular biology and evolution*, 38(5):2070–2075.

Yuan, Y.-W. and Wessler, S. R. (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proceedings of the National Academy of Sciences*, 108(19):7884–7889.

Zhang, H.-H., Peccoud, J., Zhang, X.-G., Gilbert, C., et al. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. *Nature communications*, 11(1):1–10.