bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Re-evaluating inheritance in genome evolution: widespread transfer of LINEs between species

Authors: Atma Ivancevic, Daniel Kortschak, Terry Bertozzi, David Adelson*

*Correspondence to: david.adelson@adelaide.edu.au

Horizontal transfer[1,2] (HT) is the transmission of genetic material by means other than parent-to-offspring: a phenomenon primarily associated with prokaryotes. However, eukaryotic genomes contain transposable elements[3] (TE), colloquially known as 'jumping genes' for their ability to replicate to new genomic locations. Long interspersed element (LINE) retrotransposons are TEs which move using a "copy and paste" mechanism, resulting in gene disruptions, chromosome rearrangements and numerous diseases such as cancer[4-7]. LINEs are autonomous; they can move into a new genome and immediately commence replicating. This makes them good candidates for HT. Growing evidence[8-11] shows that HT is more widespread than previously believed, although questions still remain about the frequency of HT events and their long-term impact. Here we show that LINE-1 (L1) and Bovine-B (BovB)[12,13], the two most abundant retrotransposon families in mammals, were initially introduced as foreign DNA via ancient HT events. Using a 503-genome dataset, we identify multiple ancient L1 HT events in plants and show that L1s infiltrated the mammalian lineage after the monotreme-therian split, in contrast with the current literature[14]. We also extend the BovB paradigm by identifying: more than twice the number of estimated transfer events compared to previous studies[8,11]; new potential blood-sucking parasite vectors and occurrences in

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

new lineages (e.g. bats, frog). Given that these retrotransposons make up nearly half of the genome sequence in today's mammals[3], our results provide the first evidence that HT can have drastic and long-term effects on the new host genomes. This revolutionizes our perception of genome evolution to consider external factors, such as the natural introduction of foreign DNA. With the advancement of genome sequencing technologies and bioinformatics tools, we anticipate our study to be the first of many large-scale phylogenomic analyses exploring the role of HT in genome evolution.

Three criteria are typically used to detect HT candidates: 1) a patchy distribution of the TE across the tree of life; 2) unusually high TE sequence similarity between divergent taxa; and 3) phylogenetic inconsistencies between TE tree topology and species relationships[15]. To comprehensively test these criteria, we performed large-scale phylogenomic analyses of over 500 eukaryotic genomes (plants and animals) using iterative similarity searches of BovB and L1 sequences.

Our findings show that there are two phases in HT: effective insertion of the TE, followed by expansion throughout the genome. Figure 1 shows that both BovB and L1 elements have been horizontally transferred because of their patchy distribution across eukaryotes. Both are absent from most arthropod genomes yet appear in relatively primitive species such as sea urchins and sea squirts. Furthermore, both TEs are present in a diverse array of species including mammals, reptiles, fish and amphibians. The main difference between BovB and L1 lies in the number of colonised species. BovBs are only present in 60 of the 503 species analysed, so it is easy to trace their horizontal transfer between the distinct clades (e.g. squamates, ruminants). In contrast, L1s encompass a total of 407 species,

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

within plants and animals, and they are ubiquitous across the well-studied therian mammals. However they are surprisingly absent from platypus and echidna (monotremes). There are only two possible explanations for this; either L1s were expunged shortly after the monotreme-therian split but before they had a chance to accumulate, or monotremes never had L1s. The first scenario is unlikely in the context of L1 distributions in other eukaryotes. Consider the 60 currently available bird genomes: full-length L1s have all but been eradicated from the avian lineage, but every bird species bears evidence of ancient/ancestral L1 activity through the presence of fragments[16]. In contrast, there are no L1 fragments in monotremes. We therefore conclude that L1s were inserted into a common ancestor of therian mammals, between 160 and 191 Million Years Ago (MYA), and have since been vertically inherited (see below).

The abundance of TEs differs greatly between species. As shown in Fig. 1, mammalian genomes are incredibly susceptible to BovB and L1 expansion. More than 15% of the cow genome is formed by these TEs (12% BovB, 3% L1). This is without considering the contribution of TE fragments[17] or derived Short INterspersed Elements (SINEs), boosting retrotransposon coverage to almost 50% in mammalian genomes[3]. Even within mammals there are noticeable differences in copy number; for example, bats and equids have a very low number of full-length BovBs (<50 per genome) compared to the thousands found in ruminants and Afrotherian mammals. The low copy number here is TE-specific rather than species-specific; there are many L1s in bats and equids. Hence, the rate of TE propagation is determined both by the genome environment (e.g. mammal versus non-mammal) and the type of retrotransposon (e.g. BovB versus L1).

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To develop a method for identifying horizontal transfer events, we used BovB, a TE known to undergo HT. We clustered and aligned BovB sequences (both full-length nucleotide sequences and amino acid reverse-transcriptase domains) to generate a representative consensus for each species, and infer a phylogeny (Fig 2a shows the nucleotide-based tree). The phylogeny supports previous results[8] — with the topology noticeably different from the tree of life (Fig. 1) — although we were able to refine our estimates for the times of insertion. For example, the cluster of equids includes the white rhino, *Ceratotherium simum*, suggesting that BovBs were introduced into the most recent common ancestor before these species diverged. The low copy number in equids and rhino, observed in Fig. 1, is not because of a recent insertion event. The most likely explanation is that the donor BovB inserted into an ancestral genome, was briefly active, lost its ability to retrotranspose and was subsequently inherited by its descendants.

The placement of arthropods is intriguing, revealing potential HT vectors and the origin of BovB retrotransposons. For example, BovBs from butterflies, moths and ants appear as a basal monophyletic group, sister to sea squirt *Ciona savignyi* BovB. The presence of BovB in all these species suggests that BovB TEs may have arisen as a subclass of ancient RTEs, countering the belief that they originated in squamates[11]. The next grouping consists of two scorpion species (*Mesobuthus martensii* and *Centruroides exilicauda*) nestled among the snakes, fish, sea urchin and leech — a possible vector. But the most interesting arthropod species is *Cimex lectularius*, the common bed bug, known to feed on animal blood. The full-length BovB sequence from *Cimex* shares over 80% identity to viper and cobra BovBs; their reverse transcriptase domains share over 90% identity at the amino acid level. Together, the bed bug and leech support the idea[8] that blood-sucking parasites can transfer retrotransposons between the animals they feed on.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We extended the BovB paradigm to include 10 bat species and one frog (*Xenopus tropicalis*). The bats were not included in the phylogenetic analysis because their BovB sequences were too divergent to construct an accurate consensus. Instead, we clustered all individual BovB sequences to identify two distinct subfamilies (Fig. 2b); one containing all the horse and rhino BovBs as well as eight bat sequences, and the other containing the remaining bat BovBs as well as the single BovB from *Xenopus*. We also included three annotated sequences from a public database[18] to resolve an apparent discrepancy between the naming of BovB/RTE elements. Our results have several implications: first, bat BovBs can be separated into two completely distinct clades, suggesting bat BovBs arose from independent insertion events; second, the BovBa-1-EF bat clade may have arisen from an amphibian species, or vice versa; and third, the naming conventions used in RepBase[18] need updating to better distinguish BovB and RTE sequences. This third point is discussed in the Supplementary Information (see Supp. Fig. 1).

In order to exhaustively search for all cases of BovB HT, we replicated the all-against-all BLAST[19] approach used in El Baidouri *et al.*[20] to detect individual HT candidate sequences. Briefly, this compares all sequences within a database to generate BovB clusters or families. We identified 215 HT candidate families which contained BovBs belonging to at least two different eukaryotic species. Many of these were closely related species; so to find the HT families most likely to be true events we restricted the analysis to families that linked species in different eukaryotic Orders (e.g. Afrotheria and Monotremata). We performed *a machina* validation for each candidate HT family: pairwise alignments of the flanking regions to rule out possible contamination or orthologous regions, and phylogenetic reconstructions to confirm discordant relationships.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A total of 22 HT families passed all of the tests, indicating at least 22 cross-Order HT events. Many HT families included one or two reptile BovBs, and numerous mammalian BovBs (see Supp. Table 6). This is important for determining the direction of transfer. BovBs are thought to have entered ruminants after squamates[11]. The single reptile element in a family is therefore likely to be the original transferred sequence, supporting the theory that retrotransposons undergo HT to escape host suppression or elimination[20]. Altogether, our results demonstrate that the horizontal transfer of BovB elements is even more widespread than previously reported, providing one of the most compelling examples of eukaryotic horizontal transfer to date.

We carried out the same exhaustive search in L1s, which presented a challenge because of greater divergence and a strong vertical background. Producing a consensus for each species was impractical as most species contained a divergent mixture of old, degraded L1s and young, intact L1s. Instead, we used the all-against-all clustering strategy on the collated dataset of L1 nucleotide sequences over 3kb in length (>1 million sequences total). 2815 clusters contained L1s from at least two different species; these were our HT candidates. As with BovBs, to improve recognition of HT we looked for families displaying cross-Order transfer. Most non-mammalian L1s (insects, reptiles, amphibians) had already been excluded because they definitively grouped into species-specific clusters, even at low (50%) clustering identity. The remaining families were from plants and mammals. After the validation tests, we found that all the mammalian candidate families were very small (e.g. one L1 element per species), and located in repeat-dense, orthologous regions in the genome most likely explained by vertical inheritance (see Supp. Fig. 3). Thus, we found no evidence for recent L1 transfer since their insertion into the therian mammal lineage and subsequent shaping of modern therian genomes.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Nevertheless, four plant families presented a strong case for L1 horizontal transfer (Fig 3a). High sequence identity was restricted to the elements themselves, there were more than two L1 elements in each family, the sequences encoded open reading frames or had intact reverse-transcriptase domains, and the phylogenetic reconstructions showed evolutionary discordance. The number of elements in each family mimicked the patterns seen with BovBs: very few elements from the 'donor species', and a noticeable expansion of L1s in the 'host species'. This indicates that transferred L1s can retain activity and expand within their new host. Moreover, it contradicts the belief that L1s are exclusively vertically inherited, and supports our conclusion that a similar event introduced L1s to mammals. At this stage, we do not know the vector of transfer since none of the analysed arthropods showed similarity to plant L1 sequences.

During our mining of candidate L1 HT families, we serendipitously discovered a chimeric L1-BovB element present in cattle genomes (*Bos taurus* and *Bos indicus*), shown in Fig. 3b. This particular element most likely arose from a recently active L1 element (98% identical to the canonical *Bos* L1-BT[18]) inserting into an active BovB (97% identical to *Bos* BovB[18]). In fact, L1s and BovBs have accumulated to such extents in these two genomes that they have created the ideal environment for chimeric repetitive elements. With two reverse-transcriptase domains and high similarity to currently active L1/BovB elements, this chimeric element has the potential to still be functional - presenting the possibility for L1 elements to be horizontally transferred throughout mammals by being transduced by BovBs.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In conclusion, both BovB and L1 retrotransposons can undergo HT, albeit at different rates. We extracted millions of retrotransposon sequences from a 503-genome dataset, demonstrating the similarly patchy distributions of these two LINE classes across the eukaryotic tree of life. We further extended the analysis of BovBs to include blood-sucking arthropods capable of parasitising mammals and squamates, as well as two distinct clades of bat BovBs and the first report of BovB in an amphibian. Contrary to the belief of exclusive vertical inheritance, our results with L1s suggest multiple ancient HT events in plants and, strikingly, HT into the early therian mammal lineage. This transfer allowed subsequent expansion of L1s and associated SINEs, transforming genome structure, regulation and gene expression in mammals[7] and potentially catalysing the therian radiation.

References

1 Piskurek, O. & Jackson, D. J. Transposable Elements: From DNA Parasites to Architects of Metazoan Evolution. *Genes-Basel* **3**, 409-422, doi:10.3390/genes3030409 (2012).

2 Ivancevic, A. M., Walsh, A. M., Kortschak, R. D. & Adelson, D. L. Jumping the fine LINE between species: Horizontal transfer of transposable elements in animals catalyses genome evolution. *Bioessays* **35**, 1071-1082, doi:10.1002/bies.201300072 (2013).

3 Prak, E. T. L. & Kazazian, H. H. Mobile elements and the human genome. *Nat Rev Genet* **1**, 134-144 (2000).

4 Chen, J. M., Stenson, P. D., Cooper, D. N. & Ferec, C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human genetics* **117**, 411-427, doi:10.1007/s00439-005-1321-0 (2005).

5 Kaer, K. & Speek, M. Retroelements in human disease. *Gene* **518**, 231-241, doi:10.1016/j.gene.2013.01.008 (2013).

6 Kazazian, H. H., Jr. Mobile elements and disease. *Current opinion in genetics & development* **8**, 343-350 (1998).

7 Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**, 71-86, doi:10.1038/nrg.2016.139 (2017).

8 Walsh, A. M., Kortschak, R. D., Gardner, M. G., Bertozzi, T. & Adelson, D. L. Widespread horizontal transfer of retrotransposons. *Proceedings of the National*

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Academy of Sciences of the United States of America **110**, 1012-1016, doi:10.1073/pnas.1205856110 (2013).

9 Sormacheva, I. *et al.* Vertical Evolution and Horizontal Transfer of CR1 Non-LTR Retrotransposons and Tc1/mariner DNA Transposons in Lepidoptera Species. *Mol Biol Evol* **29**, 3685-3702, doi:10.1093/molbev/mss181 (2012).

10 Suh, A. *et al.* Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. *Nat Commun* **7**, doi:10.1038/Ncomms11396 (2016).

11 Kordis, D. & Gubensek, F. Horizontal transfer of non-LTR retrotransposons in vertebrates. *Genetica* **107**, 121-128 (1999).

12 Kazazian, H. H., Jr. Genetics. L1 retrotransposons shape the mammalian genome. *Science* **289**, 1152-1153 (2000).

13 Kordis, D. & Gubensek, F. Molecular evolution of Bov-B LINEs in vertebrates. *Gene* **238**, 171-178 (1999).

14 Waters, P. D., Dobigny, G., Waddell, P. J. & Robinson, T. J. Evolutionary history of LINE-1 in the major clades of placental mammals. *PloS one* **2**, e158, doi:10.1371/journal.pone.0000158 (2007).

15 Schaack, S., Gilbert, C. & Feschotte, C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in ecology & evolution* **25**, 537-546, doi:10.1016/j.tree.2010.06.001 (2010).

16 Ivancevic, A. M., Kortschak, R. D., Bertozzi, T. & Adelson, D. L. LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. *Genome biology and evolution* **8**, 3301-3322, doi:10.1093/gbe/evw243 (2016).

17 Adelson, D. L., Raison, J. M. & Edgar, R. C. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12855-12860, doi:10.1073/pnas.0901282106 (2009).

18 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-467, doi:10.1159/000084979 (2005).

19 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

20 El Baidouri, M. *et al.* Widespread and frequent horizontal transfers of transposable elements in plants. *Genome research* **24**, 831-838, doi:10.1101/gr.164400.113 (2014).

21 Maddison, D. R. & Schulz, K. S. *The Tree of Life Web Project*, <http://tolweb.org/> (2007).

22 Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242-W245, doi:10.1093/nar/gkw290 (2016).

23 Kumar, S. & Hedges, S. B. TimeTree2: species divergence times on the iPhone. *Bioinformatics* **27**, 2023-2024, doi:10.1093/bioinformatics/btr315 (2011).

24 Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. (2007).

25 Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics* **7**, 474, doi:10.1186/1471-2105-7-474 (2006).

26 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461, doi:10.1093/bioinformatics/btq461 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

27 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37, doi:10.1093/nar/gkr367 (2011).
28 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
29 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).
30 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).
31 Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649, doi:10.1093/bioinformatics/bts199 (2012).
32 Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC bioinformatics* **12**, 116, doi:10.1186/1471-2105-12-116 (2011).

Supplementary Information

Additional supplementary material is provided in the attached PDF.

Acknowledgements

The authors wish to acknowledge Olivier Panaud and Steve Turner for helpful discussions, Reuben Buckley and Lu Zeng for ideas and moral support, Brittany Howell for proof reading and providing a much needed sanity check and Matt Westlake for HPC support above and beyond the call of duty.

Author Information

Affiliations

Department of Genetics and Evolution, Biological Sciences, The University of Adelaide, South Australia, Australia

Atma M. Ivancevic, R. Daniel Kortschak, Terry Bertozzi, David L. Adelson

Evolutionary Biology Unit, South Australian Museum, South Australia, Australia

Terry Bertozzi

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Contributions

A.M.I. performed the analysis, interpreted the results and wrote the manuscript. R.D.K, T.B. and D.L.A. supervised the development of work and assisted in analysing the results and writing the manuscript. T.B. provided access to DNA samples and performed laboratory validation experiments.

Corresponding author

Correspondence to: David L. Adelson**,** david.adelson@adelaide.edu.au

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

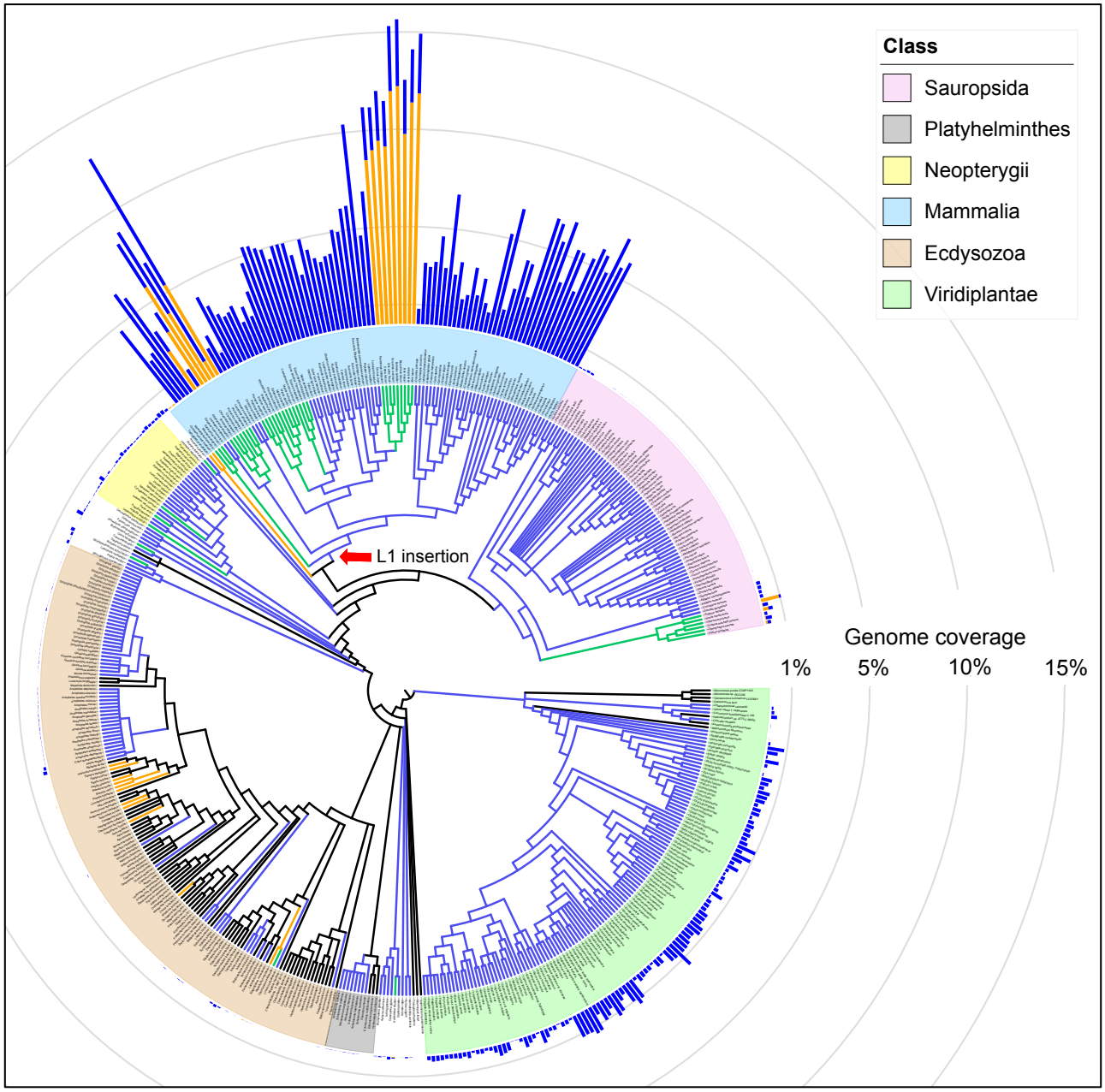


Figure 1: **Presence and coverage of L1 and BovB elements across eukaryotes.** The Tree of Life[21] was used to infer a tree of the 503 species used in this study; iTOL[22] was used to generate the bar graph and final graphic. The red arrow marks the L1 horizontal transfer event into therian mammals between 163-191MYA. Branches are coloured to indicate which species have both BovB and L1 (green), only BovB (orange), only L1 (blue), or neither (black). Bar graph colours correspond to BovB (orange) and L1 (blue). An interactive vesion of this figure is available at: http://itol.embl.de/shared/atma .

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

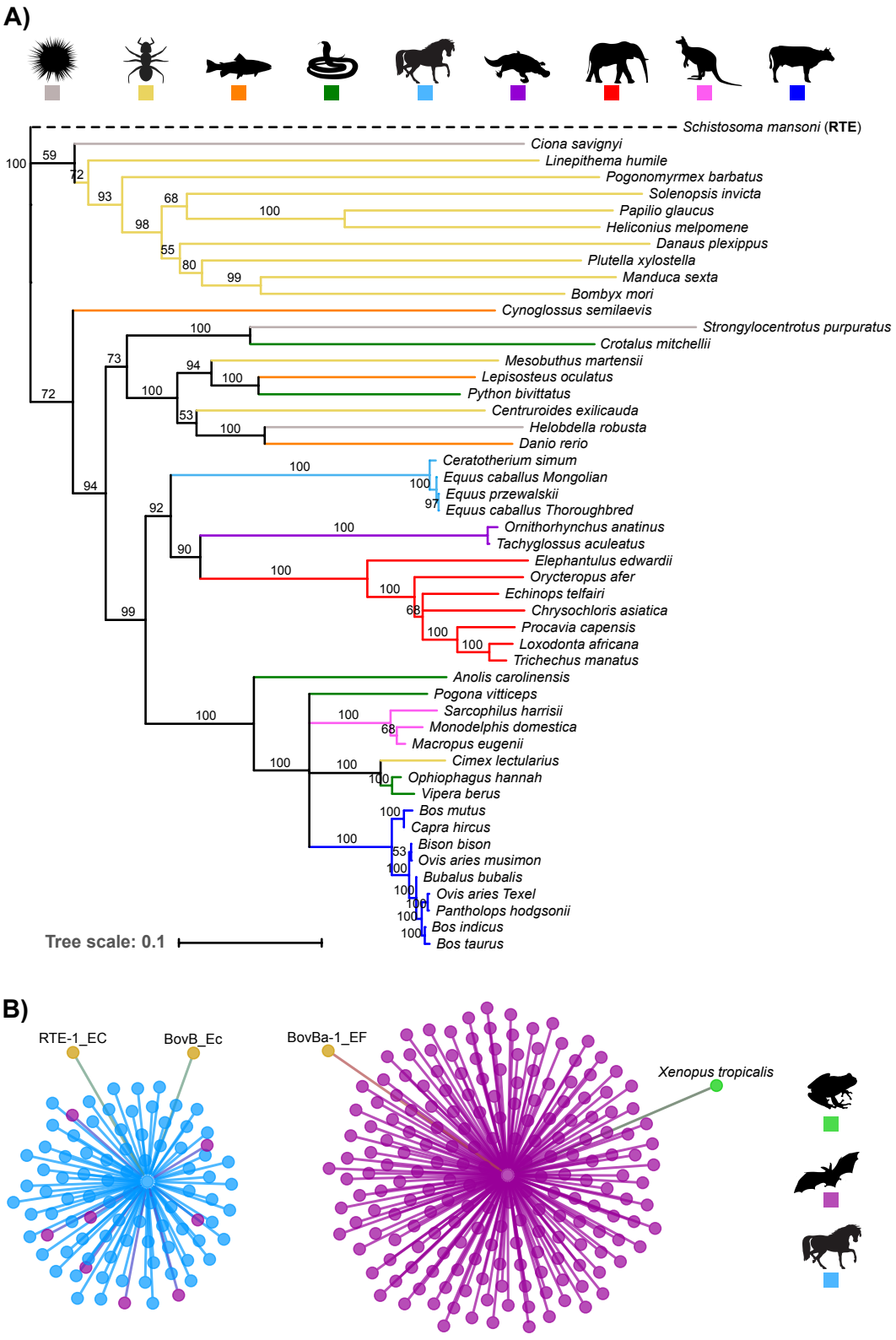
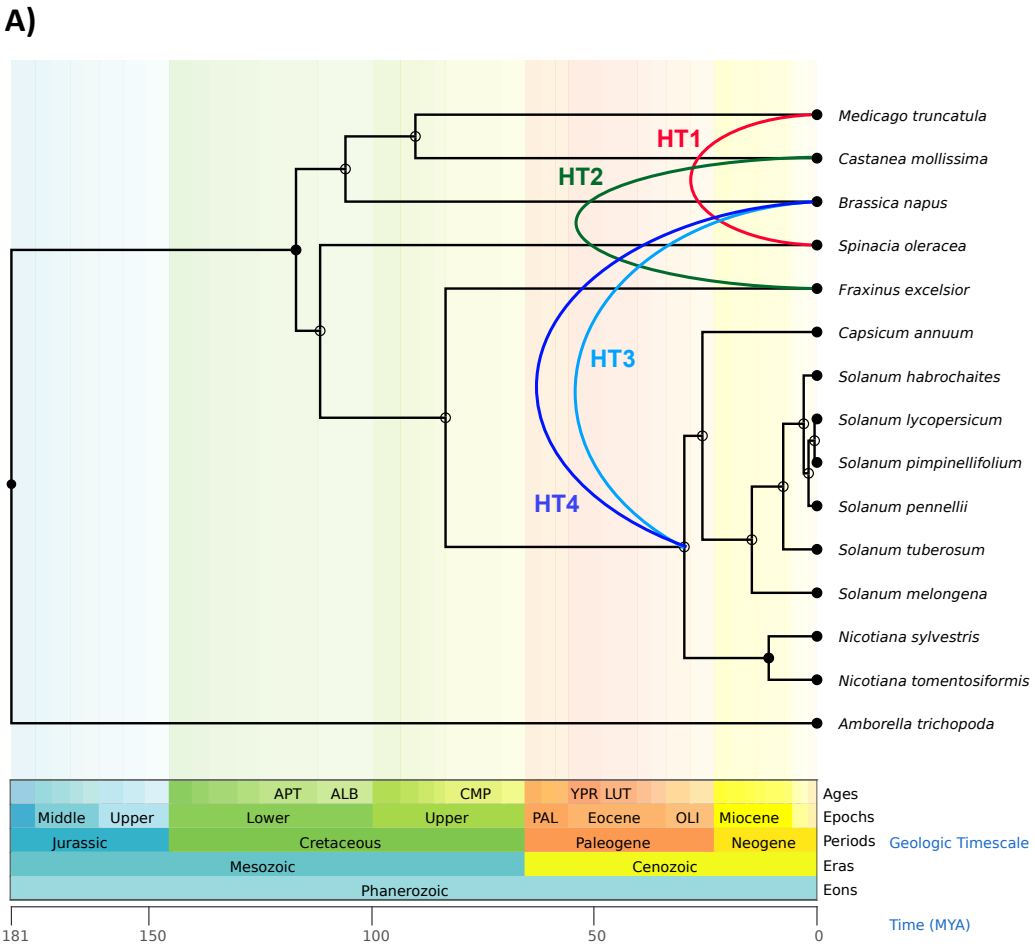


Figure 2: **HT of BovB retrotransposons**. (2a) Neighbour-joining tree (1000 bootstrap replicates) inferred using full-length nucleotide BovB consensus sequences, representing the dominant BovB family in each species. Nodes with confidence values over 50% are labelled and branches are coloured taxonomically. RTE sequence from *Schistosoma mansoni* was used as the outgroup. (2b) Network diagram representing the two distinct BovB clades in bats. Nodes are coloured taxonomically apart from the RepBase[18] sequences (light brown). RTE-1_EC and BovB_Ec are shown to belong to a single family, while BovBa-1_EF-like bat sequences form a separate family containing a single full-length BovB from the frog *Xenopus*.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



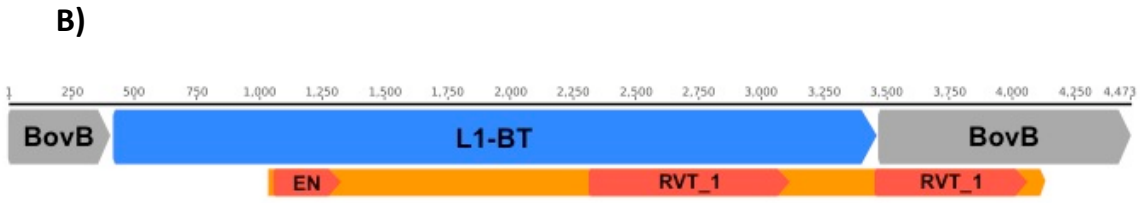


Figure 3: **HT of L1 in plants and newfound chimeric L1-BovB element**. (3a) TimeTree[23] illustrating the putative L1 horizontal transfer events between plant species. Shows only the species involved in HTs, and *Amborella trichopoda* as the outgroup. Background is coloured to match the ages in the geological timescale. (3b) Chimeric L1-BovB retrotransposon found in cattle genomes (*Bos taurus* and *Bos indicus*). L1-BT and BovB correspond to RepBase names[18], representing repeats which are known to have been recently active. RVT_1 = reverse-transcriptase, EN = endonuclease domain. The orange bar is the length of the entire open reading frame.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Extraction of L1 and BovB retrotransposons from genome data

To extract the retrotransposons of interest, we used the methods and genomes previously described in Ivancevic *et al.* (2016)[16]. Briefly, this involved downloading 499 publicly available genomes (and acquiring 4 more from collaborations), then using two independent searching strategies (LASTZ[24] and TBLASTN[19]) to identify and characterise L1 and BovB elements. A third program, CENSOR[25], was used with the RepBase library of known repeats[18] to verify hits with a reciprocal best-hit check. The raw L1 results have been previously published in Ivancevic *et al.*[16] (Supplementary Material); the BovB results are included in the Supplementary Material.

Extraction and clustering of conserved amino acid residues

Starting with BovBs, USEARCH[26] was used to find open readings frames (ORFs), with function -fastx_findorfs and parameters -aaout (for amino acid output) and -orfstyle 7 (to allow non-standard start codons). HMMer[27] was used to identify reverse transcriptase (RT) domains within the ORFs. RT domains were extracted using the envelope coordinates from the HMMer domain hits table (-domtblout), with minimum length 200 amino acid residues. The BovB RT domains from all species were collated into one file and clustered with UCLUST[26]. This was done as an initial screening to detect potential horizontal transfer candidates. The process was then repeated with L1 elements.

Clustering of nucleotide sequences to build one consensus per species

The canonical BovB retrotransposon is 3.2 kb in length[8,18], although this varies slightly between species. In this study, we classified BovB nucleotide sequences ≥2.4kb and ≤4kb as full-length. We wanted to construct a BovB representative for each species.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Accordingly, for each species, UCLUST[26] was used to cluster full-length BovB sequences at varying identities between 65-95%. A consensus sequence of each cluster was generated using the UCLUST -consout option.

The ideal cluster identity was chosen based on the number and divergence of sequences in a cluster. E.g. for species with few BovBs, a lower identity was allowed; whereas for species with thousands of BovBs, a higher identity was needed to produce an alignable cluster. The final clustering identity and cluster size for each species are given in Supp. Table 1. Note that the bat species are not included in this table - they were clustered separated, due to the high level of divergence between BovBs.

This method was tested on L1 retrotransposons, but the results were not ideal; most species simply had too many L1 sequences. Other methods tested on both BovBs and L1s included using centroids instead of consensus sequences (this gave better alignments but was less representative of the cluster), and using the same clustering identity for all species (e.g. 80% - this did not work well for species with less than 100 elements in the genome).

Inferring a phylogeny from consensus sequences

Consensus sequences were aligned with MUSCLE[28]. The multiple alignment was processed with Gblocks[29] to extract conserved blocks, with default parameters except min block size: 5, allowed gaps: all. FastTree[30] was used to infer a maximum likelihood phylogeny using a general time reversible (GTR) model and gamma approximation on substitution rates. Geneious Tree Builder[31] was used to infer a second tree using the neighbour-joining method with 1000 bootstrap replicates.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Distinguishing RTEs from BovBs

All sequences which identified as BovB or RTE were kept and labelled accordingly to their closest RepBase classification[18]. However, there appeared to be numerous discrepancies with the naming: e.g. some RTE sequences shared >90% identity to BovBs, and vice versa. BovB retrotransposons are a subclass of RTE, and they were only discovered relatively recently. It is likely that several so-called RTE sequences are actually BovBs.

To determine which species had BovB sequences, and which only had RTEs, we used the species consensus approach to build a BovB/RTE phylogeny (see Supp. Figure 1). This effectively separated BovB-containing species from RTE-containing species. The RTE sequences were not included in further analyses.

Clustering of nucleotide BovB sequences from bats and *Xenopus*

A reliable BovB consensus could not be generated for any of the ten bat species because the sequences were too divergent and degraded. Some bat BovBs seemed similar to equid BovBs; others did not. Likewise, the single full-length BovB from frog *Xenopus tropicalis* was very different to canonical BovBs, sharing highest identity with the bats.

In an effort to characterise these BovBs into families, we grouped all full-length BovB sequences from the bats, frog, equids and white rhino into a single file. We also added two RepBase equid sequences (RTE-1_EC and BovB_Ec) and 1 RepBase bat sequence (BovBa-1_EF)[18]. After clustering, we expected to find one family of equid BovBs, the equid RTE sequence as an outlier, and numerous families containing bat and frog BovBs. The actual findings are described in the text (Fig. 2b). We used UCLUST[26] to cluster the sequences (function -cluster_fast with parameters -id, -uc, -clusters). The highest identity

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

at which there were only 2 clusters/families was 40%. At higher identities, the equid BovBs stayed together but the bat and frog BovBs were lost as singletons.

To confirm the clustering, we also used MUSCLE to align all the sequences and FastTree to infer a phylogeny (see Supp. Figure 2).

HT candidate identification - BovBs and L1s

We compiled all confirmed BovB and L1 sequences into separate multi-fasta databases (316,017 and 1,048,478 sequences respectively). The length cut-off for BovBs was ≤2.4kb and ≥4kb; for L1s, ≥3kb. BovBs were analysed first to identify characteristics of horizontal transfer events.

To detect HT candidates, we used the all-against-all clustering strategy described in El Baidouri *et al.*[20]. Briefly, this method used a nucleotide BLAST[19] to compare every individual sequence in a database against every other sequence; hence the term all-against-all. BLAST parameters were as follows: -r 2, -e 1e-10, -F F, -m 8 (for tabular output). The SiLiX program[32] was then used to filter the BLAST output and produce clusters or families that met the designated identity threshold.

For BovB sequences, we tested identities of 40-90%. High identity thresholds were useful for finding very recent HT events (e.g. over 90% identity between the bed bug and snakes). However, the majority of clusters contained several copies of the same BovB family from a single species - indicative of vertical inheritance. Using a lower identity threshold was more informative for capturing ancient HT events. At 50% identity, the clustering preserved the recent, high-identity HT events while also finding the ancient, lower-%identity HT events. We concluded that this was the best %identity to use for our particular dataset, considering it includes widely divergent branches of Eukaryota.

bioRxiv preprint doi: https://doi.org/10.1101/106914; this version posted February 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Clusters were deemed HT candidates if they contained BovB elements belonging to at least two different species. To reduce the number of possible HT clusters, we went one step further and kept only the clusters which demonstrated cross-Order transfer (e.g. BovBs from Monotremata and Afrotheria in the same cluster). All potential HT candidates were validated by checking that they were not located on short, isolated scaffolds or contigs in the genome. The flanking regions of each HT candidate pair were extracted and checked (via pairwise alignment) to ensure that the high sequence identity was restricted to the BovB region. This was done to check for contamination or orthologous regions. Phylogenies of HT candidate clusters were inferred using maximum likelihood and neighbour-joining methods (1000 bootstraps).

The same procedure was performed to screen for nucleotide L1 HT candidates. As an extra step for L1s, we also used all ORF1 and ORF2 amino acid sequences from a previous analysis[16] to conduct similar all-against-all BLAST searches. However, the amino acid clusterings did not produce any possible HT candidates.

Data availability statement

Data generated or analysed during this study are included in the main text and Supplementary Material. Raw sequences are provided upon request.