The Evolution of L1 and L2 Retrotransposons in Therian Mammals and Monotremes

Alexander Stuart

a1749756

Supervised by:

Prof. David Adelson

The School of Biological Sciences

The University of Adelaide

2021

Declaration of Originality

This work contains no material previously accepted for the award of any other degree or diploma in any university or tertiary institution and, to the best of my knowledge, contains no material written and/or published by another person, except where due reference has been made in the text.

Alexander Stuart

22nd of October, 2021

Acknowledgements

I would like to thank my honours supervisor Dave Adelson for the fantastic opportunity to work on such an interesting project, and for introducing me to the wonderful world of transposable elements; your mentorship and guidance has been greatly appreciated.

I would also must make specific thanks to Joe McConnell, for his approachability and unfailing helpfulness during my time here. Eugene Lee, I have greatly enjoyed our chats, and I thank you for helping maintain my sanity throughout the year. Carey, your assistance with LaTeX was both appreciated and absolutely necessary, I couldn't have finished writing this thesis without your help. I would also like to thank all past and present members of the Adelson La, James, Dan, Erbo, Ha, Noz, Zhipeng, Yuka and Hanyuan for being so welcoming, and making the past year so enjoyable. I would also like to thank Frank Grützner and Linda Shearwin for their encouragement, interesting questions and contributions to this project.

Additionally, I would like to thank my good friend Xavier Montin for his support throughout this project, and our many, many coffees together. To my parents, your lifelong support of my interests and passions are what helped make this project possible, and I thank you. To all my other friends and family who encouraged, inspired and supported me on one way or another, I just have to say `thanks!`

```
                    /
_\|/\/_   /
/\/\\•>
  " "
```

Contents

List of Figures

List of Tables

Abbreviations

3'	Three Prime
5'	Five Prime
BED	Browser Exensible Data
BLAST	Basic Local Alignment Search Tool
bp	base pair
CARP	Comprehensive `ab initio` Repeat Pipeline
CR1	Chicken Repeat 1
EEN	Endonuclease
env	envelope
ERV	Endogenous Retrovirus
gag	group-specific antigen
GFF	General Feature Format
GTF	Gene Transfer Format
HMM	Hidden Markov Model
HPC	High Performance Computing
kb	kilobase
***k*-mer**	Length k subsequence
L1	Long Interspersed Nuclear Element 1
L2	Long Interspersed Nuclear Element 2
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
miRNA	micro RNA
mya	million years ago
NCBI	National Center for Biotechnology Information

nt	nucleotide
ORF	Open Reading Frame
pol	polymerase
RNA-seq	RNA-sequencing
RTE	Retrotransposable Element
RVT	Reverse Transcriptase
SINE	Short Interspersed Nuclear Element
siRNA	small interfering RNA
SLURM	Simple Linux Utility for Resource Management
SSR	Simple Sequence Repeat
TE	Transposable Element
TSD	Tandem Site Duplication
UCSC	University of California, Santa Cruz.
UTR	Untranslated Region

Abstract

Transposable elements (TEs) are known to be significant drivers of gene and genome evolution across the eukaryotic tree of life. L1 retrotransposons have undergone a relatively recent expansion in mammalian species, often being the dominant TE, with the exception of monotremes. Unlike therian mammals, monotreme genomes appear to have never been influenced by L1s, instead being dominated by a class of retrotransposon no longer active in therians, the L2s. As the only mammals with genomes untouched by L1s, this study initially characterised TEs within monotremes *ab initio*, and subsequently compared their insertion habits with L1s in therians.

The echidna genome has recently been made publicly available, but has had no in-depth TE analysis. This project sought to initially characterise new classes of potentially active repeat within the echidna, and use these sequences to reclassify dispersed repeat families identified through the comprehensive *ab initio* repeat pipeline, ultimately producing repeat annotations for the echidna and platypus. These annotations were used to see where L2s have been inserting into different exonic regions, and make comparisons to L1 insertions in therians.

Through this, I was able to detect 11 new subclasses of potentially active retrotransposon within the echidna, and compare transposon content and composition between the echidna and platypus. While a high number of L2 overlaps were observed when looking at insertions, the significant difference in predicted gene quality between monotremes and therians made a direct comparison not possible. Despite this, the generation of high quality monotreme specific repeat annotations makes a variety of other L1-L2 comparisons possible, and is planned for a future project.

Chapter 1: Literature Review and Project Details

1.1 Introduction

Eukaryotic genomes are dominated by repetitive elements. These regions are known to be largely derived from transposable elements (TE), selfish sequences capable of moving and multiplying throughout the genome (Doolittle and Sapienza, 1980). Their activity is known to influence genome size, genes and regulatory elements, making them powerful drivers of evolution (Piskurek and Jackson, 2012). Some TEs even have the potential to move between species, rather than from parent-to-offspring, in a process known as horizontal transfer (Finnegan, 1989).

The most abundant components of mammalian genomes are the LINE and SINE classes of retrotransposon; the L1 LINE and its associated SINEs were initially reported to cover over 30% of the human genome (Lander et al., 2001), and are thought to have entered the mammalian genome through a horizontal transfer event (Ivancevic et al., 2018). L1 elements are found in all therian mammals, yet are entirely absent from monotremes, a highly diverged clade of mammals including the platypus and echidna. Instead, monotreme genomes are dominated by L2s, a group of TEs now extinct in all other extant mammal lineages (Warren et al., 2008). Monotremes offer a unique glimpse into a mammalian genome untouched by L1s, representing an alternate path of evolutionary history. By investigating how L2s have contributed to monotreme genome evolution, we can explore the progression of mammalian evolution, and may better understand how L1s have impacted our own evolution.

1.1.1 The Evolution of Mammals

Mammals are a highly diverse class, and have radiated to fill a wide range of niches. While their dominance and variation makes them an important group for understanding

evolution as a whole, there is also a certain anthropocentric bias in wanting to study mammalian evolution. Despite their radiation after the Cretaceous-Paleogene extinction event 66 million years ago (mya), mammals were already quite diverse in the Mesozoic, with at least 25 lineages coexisting with non-avian dinosaurs (Luo, 2007; Grossnickle et al., 2019).

All extant mammals can be grouped into three clades, the eutherians (placental mammals), metatherians (marsupials) and monotremes. Eutherians exhibit the greatest contemporary diversity, with representatives found on every continent and in almost every environment. Extant metatherians originate from Gondwana, and are found only in Australia and the Americas. While monotreme fossils have been found in both Australia and South America, extant monotremes are limited to Australia and New Guinea (Luo, 2007). Two families of monotremes can be found today; Ornithorhynchidae, represented by a single species, *Ornithorhynchus anatinus* (platypus), and Tachyglossidae, comprised of four species of echidna. Of these three clades, eutherians and metatherians are the most closely related, having split approximately ∼167 mya (Upham et al., 2019). In contrast, the most recent common ancestor of monotremes and therians is estimated to have diverged ∼187 mya, in the early Jurassic (Zhou et al., 2021).

Monotremes

Endemic to south-eastern Australia and Tasmania, the platypus is a semi-aquatic organism with a duck-like, toothless beak. Echidnas are spiny, insectivorous creatures, superficially resembling hedgehogs. The short-beaked echidna (*Tachyglossus aculeatus*) is found throughout Australia, in a wide range of habitats. New Guinean long-beaked echidnas (*Zaglossus attenboroughi, bartoni, and bruijni*) are larger, with less prominent spines. The fossil record of monotremes is sparse, but phylogenetic analysis has estimated the split between these two families to have occurred approximately 55 million years ago (Zhou et al., 2021).

Monotremes have retained several traits ancestral to mammals. While oviparity is arguably the most famous 'archaic' trait unique to monotremes, the presence of a cloaca, and

certain unique skeletal features are traits which have been passed from early therapsids to monotremes, but have been lost in therian mammals (Musser, 2003).

Another way in which monotremes greatly differ from therians is in their method of sex determination. Maleness is determined by the Y chromosome in therians, specifically by the *SRY* gene (Sinclair et al., 1990). The process by which sex determination occurs in monotremes has evolved independantly, more closely resembling that of certain birds. Rather than an XX/XY system, monotremes have a multiple sex chromosome system, and lack an *SRY* gene (Grützner et al., 2004).

The unique phylogenetic position of monotremes makes them a excellent candidate for comparative genomics, to help better understand the position of mammals in the tree of life.

1.1.2 Transposable Elements

In all eukaryotic organisms sequenced to date, repetitive regions have been observed, and in many cases make up a large fraction of the genome (Deininger et al., 2003; Huang et al., 2012). Many of these regions have been found to originate from sequences capable of moving and replicating throughout the genome, known as transposable elements (TE), or transposons. Often described as 'selfish elements', TEs make use of cellular machinery to facilitate their movement. Based on their presence in almost all eukaryotic organisms, TEs appear ubiquitous to complex life as we know it (Bennetzen, 2000; Piskurek and Jackson, 2012).

Transposons can broadly be grouped into two classes by their transposition intermediate. Class I elements (retrotransposons) are the most abundant in eukaryotes, propagating through a 'copy and paste' mechanism (Smit, 1999; SanMiguel et al., 1996). Class II elements (DNA transposons) use no RNA intermediate, instead operating through a 'cut and paste' method (Hua-Van et al., 2005; Wicker et al., 2007). Retrotransposons can be classified further based on the presence of long terminal repeats (LTR).

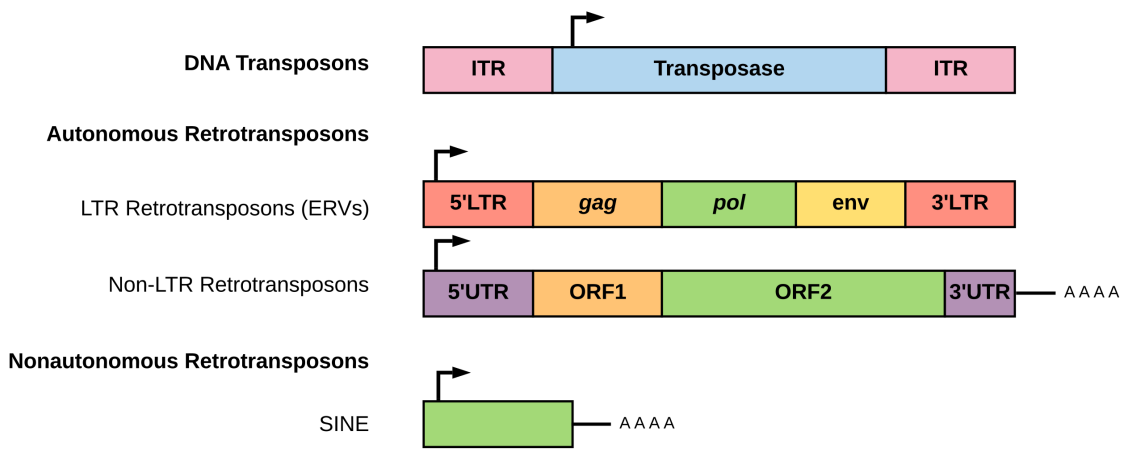


Figure 1.1: Schematic representation, showing the structures of four major groupings of transposable elements in eukaryotes. Figure made using Inkscape (0.92.5), based on (Saleh et al., 2019).

The LTR retrotransposons are a diverse group, characterised by their long terminal repeats, and the presence of at least two essential genes, *pol* (polymerase) and *gag* (group specific antigen) (Havecker et al., 2004). Additionally, an *env* gene may be present, forming coat proteins that may allow the retrotransposon to become exogenous; LTR retrotransposons with all three of these genes are referred to as endogenous retroviruses (ERVs) (Boeke and Stoye, 1997).

Non-LTR retrotransposons can be classified as autonomous or non-autonomous, based on their ability to encode the proteins required for their own retrotransposition. Sequences containing these internal proteins are referred to as long interspersed nuclear elements (LINEs), while non-autonomous sequences are called short interspersed nuclear elements (SINEs). LINEs contain an internal Pol II promoter to initiate transcription, and encode one or two proteins, ORF1 and ORF2 (Feng et al., 1996; Huang et al., 2012). ORF1 varies in function, often exhibiting RNA/DNA binding activity, while ORF2 will always encode a protein with reverse transcriptase and endonuclease activity. The endonuclease activity of the ORF2 protein produces nicks at specific short sequences, priming the retrotransposition of target RNA into the genome (Eickbush and Jamburuthugoda, 2008).

Unable to replicate on their own, SINEs use proteins expressed by other transposable elements to spread. Approximately 50-500 bp in length, these short interspersed elements do not originate directly from LINEs.; instead, the 5' end is derived from an endogenous sequence (such as a tRNA or ribosomal RNA) transcribed by a Pol III promoter, while the 3' end consists of a simple repeating sequence. Most SINEs also have an internal 'body' region, which often displays homology to a corresponding LINE (Okada et al., 1997; Kramerov and Vassetzky, 2011). Despite their small size, SINE elements often comprise a large percentage of many eukaryotic genomes, spreading rapidly in the presence of a corresponding LINE (Okada, 1991). In humans, approximately 11% of the genome is derived from a single SINE element, *Alu*, present in excess of 1 million copies. (Batzer and Deininger, 2002)

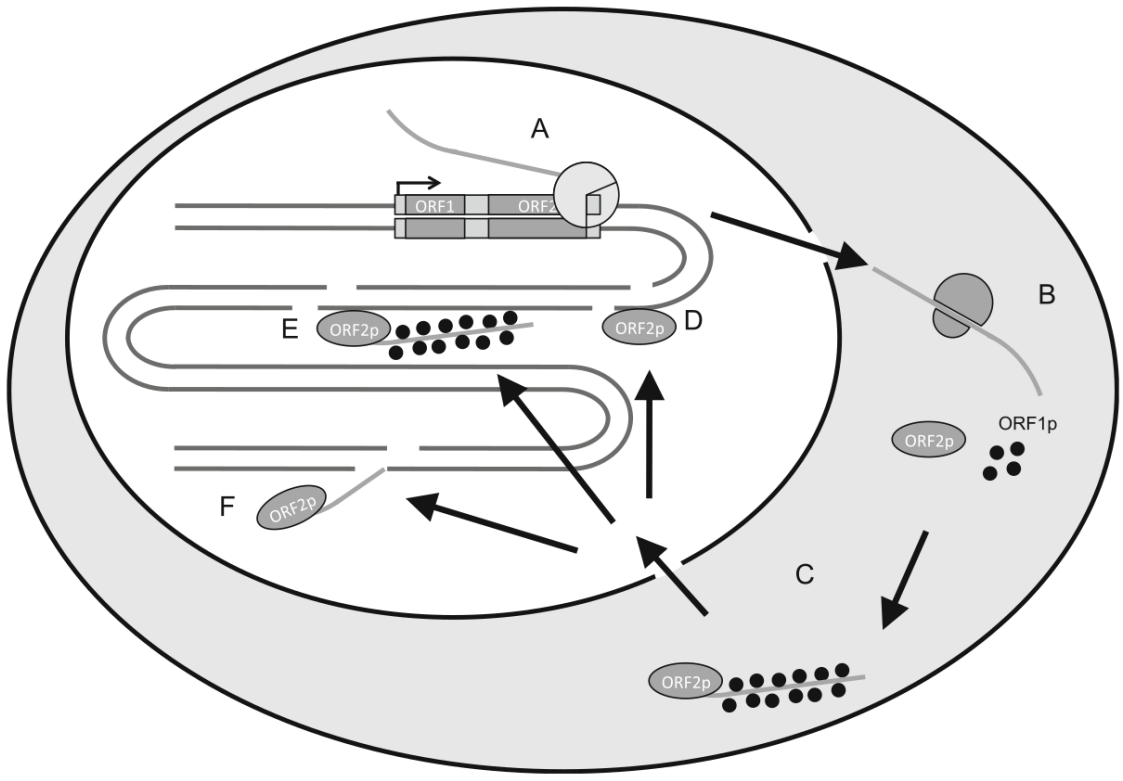


Figure 1.2: General life cycle of a non-LTR retrotransposon: **A.** ORFs are transcribed by RNA Pol II and exported to cytoplasm. **B.** ORFs are translated. ORF2 protein (ORF2p) has both endonuclease and reverse transcriptase activities. **C.** ORF2 forms a complex with TE RNA known as a ribonuclear protein. **D.** ORF2p creates double stranded breaks at target sequences. **E.** TE RNA is used as a template for the reverse transcription activity of ORF2p. **F.** TE RNA inserts into DNA. Figure taken from (Adelson et al., 2015)

Retrotransposon Regulation

The unchecked spread of transposable elements has the potential to wreak havoc on a genome; mutagenic transposition events have been observed in a wide variety of taxa, with several human cancers associated with LINE-1 deregulation (Ogino et al., 2008; Payer and Burns, 2019). There are two primary ways by which eukaryotes are able to supress transposons; through transcriptional repression and post-transcriptional degradation (Yoder et al., 1997). DNA methylation of the internal promoter, histone tail modifications and alterations to chromatin condensation have all been shown to contribute to the transcriptional repression of transposons (Eric M. Ostertag and Kazazian, 2001)

If transcription occurs, RNA interference (RNAi) may be used to silence the TE mRNA molecules. Proteins from the Dicer family can generate two types of interfering RNA through the cleavage of dsRNA: the 21-30 nucleotide small interfering RNAs (siRNA) and the ∼21 nucleotide micro RNAs (miRNA). These molecules are incorporated into an RNA-induced silencing complex (RISC), which is able to bind to complementary mRNA, leading to its cleavage and inactivation (Gebert and Rosenkranz, 2015). The diversity of RNAi mechanisms in eukaryotes is vast, often having coevolved with specific transposon types. The only constant in the forms which RNAi takes is the end function: controlling the spread of transposable elements (Gebert and Rosenkranz, 2015).

Despite the many mechanisms that keep TEs in check, they are still able to avoid complete suppression.

Transposable Elements and Evolution

The exponential rise in sequencing data generated over the past two decades has shown that TEs play a major role in the evolution of both genes and genomes (Bennetzen, 2000; Piskurek and Jackson, 2012; Suh et al., 2015; Platt et al., 2018; Cosby et al., 2021).

TEs are a major source of non-coding regulatory RNAs. While the mechanism of these RNA molecules can vary greatly, all come to influence gene expression in some manner (Chuong et al., 2017; Zeng et al., 2018). Regulatory sequences within the untranslated regions (UTRs) of several genes have been linked to a TE origin. *Alu* derived sequences have been detected in the UTRs of human genes; when inserted at the 3' end, they can influence mRNA decay and localisation. Within the 5' end, *Alu* has been shown to modulate translation efficiency (Deininger, 2011; Shen et al., 2011).

Many regulatory RNAs, such as miRNA and siRNA, seem to be derived from TEs (Feschotte, 2008; Petri et al., 2019). Computational studies have found that TEs within transcripts are able to act as templates for RNA binding proteins, including the machinery of miRNA pathways (Filshtein et al., 2012). Several functionally active miRNAs derived

from LINEs (L1, L2 and L3) and SINEs (*Alu* and MIR) are expressed in humans, and can be highly conserved (Piriyapongsa et al., 2007; Spengler et al., 2014). Despite the fact that L2 elements have been inactive in the human genome for at least 100 million years, L2-derived miRNAs appear to be expressed in multiple human cell types, with some such as miR-95 and miR-151 controlling lysosomal function and cell migration (Petri et al., 2019). Most miRNA are extremely ancient, so their origins are uncertain; however it is clear that TEs in the past have influenced RNA regulatory networks in eukaryotes, and have the potential to become parts of future regulatory networks.

While the disruption of coding regions is often detrimental, in some cases the uptake of TEs to exonic regions may give rise to novel genes, driving adaption and speciation (Joly-Lopez and Bureau, 2018). This phenomenon is known as exaption, and can be observed in eukaryotes and prokaryotes. Some well known examples of TE exaption are the *RAG* genes involved in vertebrate immunity, the CRISPR-associated system in prokaryotes, *FHY3* transcription factors in plants (Jangam et al., 2017) and Comparative genomics has linked many well researched genes with a TE origin. Additionally, TE insertions have been shown to rewire transcriptional networks through alterations to promoter/enhancer regions (Kunarso2010, Dermitzakis2002)

Horizontal Transfer

The way genetical material is usually transmitted in eukaryotes is through vertical transfer, that is by parent to offspring. While the transmission of genetic material through means other than reproduction, known as horizontal transfer, occurs frequently in prokaryotes it is far rarer in eukaryotes. Recently, through comparative sequencing, there has been a growing body of evidence for non-retroviral horizontal transfer in eukaryotes, at a significantly higher rate than would be expected (Ivancevic et al., 2018).

The direct cell-cell horizontal transfer observed in prokaryotes does not apply here, and, aside from some retroviruses, most TEs do not encode a capsule. Thus, a vector of some sort is required for horizontal transfer to occur in higher organisms (Walsh et al., 2013).

Several candidates have been proposed as vectors, including viruses, arthropods (like ticks and other parasites), snails, etc (Ivancevic et al., 2018). Many examples have been observed of horizontal transfer occuring between highly divergent species.

The insertion of foreign TEs into a genome has the clear potential to influence genome function and evolution, in much the same way as an endogenous TE has (Ivancevic et al., 2013). An example of this can be seen in the retrotransposable family Bovine-B (BovB), a TE which has spread across several phyla in the last 50 mya. Originating in squamates, it is now found in a range of genomes, including therians, monotremes, arthropods and echinoderms (Walsh et al., 2013). BovB has undergone a relatively recent major expansion in ruminants and afrotheria, becoming the dominant repetitive element in these organisms.

L1 elements have a patchy distribution throughout the tree of life, particularly in plants and fungi (Ivancevic et al., 2016). Sequence similarity analysis of L1s seems to show evidence of possible cross-phylum HT events in marine eukaryotes. However, exhaustive analysis of therians seems to indicate that there has been little to no active L1 HT in mammals. Interestingly, while L1s have been found in every therian mammal sequenced to date, they are entirely absent from monotremes. As a genetic element known for leaving a large impact on the genome, the absence of any L1 traces is striking. While it is possible that L1s have been removed from the genome of monotremes in the distant past, even organisms lacking active L1 activity (such as the megabats and birds) display evidence of past L1 activity. Because of this, Ivancevic et al. (2018) has proposed that L1s were not present in the ancestor of modern mammals, entering the genome at some point after the monotreme-therian split 160-191 mya.

1.1.3 Non-LTR Retrotransposons in Mammals

The diversity and composition of TEs in vertebrates varies enormously; the more species are sequenced, the more TEs are found (Malik et al., 1999). As additional genome assemblies are annotated for TEs, it appears that more deeply branching lineages, such as fish and amphibians, exhibit a greater diversity of TEs than more recent radiations, such as

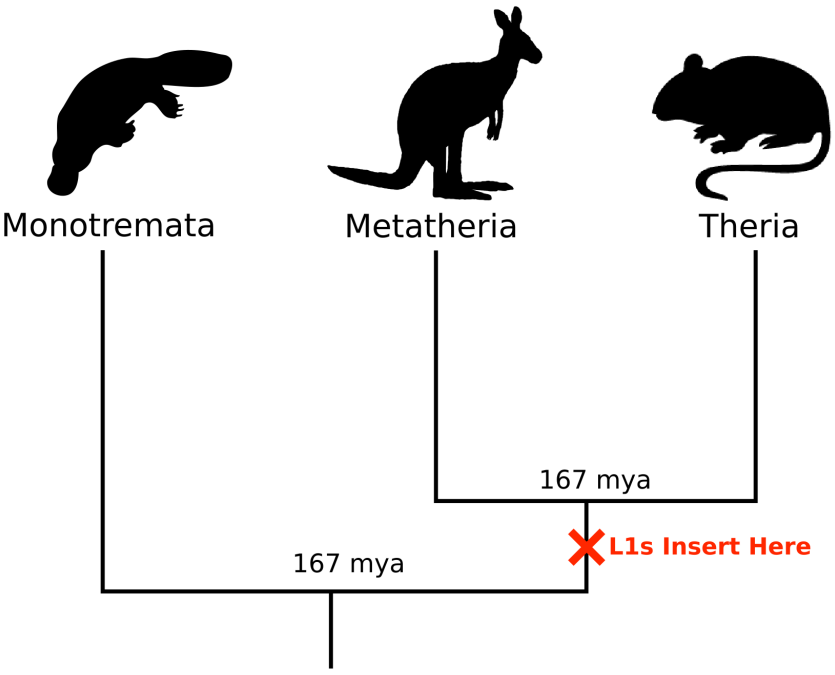


Figure 1.3: Phylogenetic tree of extant mammalian clades, Monotremata, Metatheria and Eutheria, with estimated divergence times at 187 mya and 167 mya, based on (Zhou et al., 2021) and (Upham et al., 2019) respectively. Based on L1 and L2 distribution in modern representatives of these clades, it is thought that L1s were introduced into the ancestral therian genome after the monotremata-therian split (Ivancevic et al., 2018)

birds and mammals.

LINE-1 Elements

L1 elements are one of the most well known TEs. Found in all major eukarytotic phyla, L1s are ancient and diverse in structure. Capable of rapid genome expansion, they are powerful agents of genomic change. Although their exact composition varies, active L1s are generally defined as 6-8kb elements containing a 5' UTR, and two ORFs. ORF1p is an RNA-binding protein, believed to aid in the entry of L1 RNA into the nucleus, while ORF2 encodes a protein with apurinic/apyramidic (AP) endonuclease and reverse transcriptase activity (Mathias 1991, Babushok 2007).

LINE-2 Elements

Although LINE-2 elements share a similar name with LINE-1s, they are not closely related, instead belonging to the Jockey superfamily of retrotransposons. Although they majorly differ in sequence, LINE-2s are structurally very similar to LINE-1 elements, with a 5' UTR and one-two ORFs, although its overall size is only ∼4.5 kb. ORF1p exhibits RNA binding activity, while ORF2 encodes a protein with AP endonuclease and reverse transcriptase activity (Lovšin et al., 2001). Although highly degraded L2 associated elements have been found in therian mammals, they are functionally extinct, and have been for at least 100 mya (Petri et al., 2019).

1.1.4 Conclusion

Transposable elements are key drivers of genome evolution. L1s are a highly active and abundant class of TEs in therians, and are known to have greatly influenced their evolution. As L1 elements are absent in monotremes, they represent the only contemporary example of a mammalian genome untouched by L1 activity. Thus, analysis of L2 activity in extant monotreme genomes can be used to better understand mammalian evolution as a whole.

1.2 Aims

The overarching goal of this project will be to compare how L1s and L2s have contributed to the genome evolution of therians and monotremes respectively. This will be completed by adressing three aims:

As the repeat content of the echidna has never been analysed, several pipelines will be used to identify and characterise potentially active LTR and non-LTR retrotransposons. Potentially active elements will be identified in a way that allows the generation of a Hidden Markov Model profile, so that these sequences can be submitted to Dfam.

The second aim will be to produce manual genome annotations for repeats in *Ornithorhynchus anatinus* and *Tachyglossus aculeatus*. High quality assemblies for these two species are available publicly and privately. Species specific dispersed repeat families will be identified and classsified using the Comprehensive *ab initio* Repeat Pipeline (CARP) (Zeng et al., 2018), and subsequently mapped to the genome.

Once the positions, phylogeny and nature of L2s within the monotreme genome have been characterised, a range of comparisons can be made to therians and L1s. By identifying intervals where L2s overlap with exonic and untranslated regions, genes with L2 insertions will be identified in the monotreme genomes; the frequency and nature of these insertions can be compared to L1-associated insertions in therian genomes.

1.3 Data

1.3.1 Genomes and Gene Data

The platypus (*Ornithorhynchus anatinus*, GCF_004115215), echidna (*Tachyglossus aculeatus*, GCF_004115215.4), human (*Homo sapiens*, GCF_000001405.39), horse (*Equus caballus*, GCF_002863925.1) and dog (*Canis familiaris*, GCF_000002285.5) genomes and gene annotations were sourced from the NCBI genome browser. For the purposes of *de novo* repeat identification, a higher quality assembly of the echidna genome was sourced privately, with permission from Linda Shearwin.

1.3.2 Repeats

The March 2021 collection of representative vertebrate repeat sequences was downloaded from RepBase Bao et al. (2015) Annotated transposons identifed *de novo* in the tuatara were used for L2 phylogeny analysis, with permission from Lu Zeng. `RepeatMasker` whole genome annotations associated with the human, horse and dog were downloaded from NCBI.

1.3.3 HMM Profiles

The Pfam current release 34.0 was used for HMM models, available at http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz

Chapter 2: Data and Methods

2.1 Initial *ab initio* repeat identification, using CARP

The Comprehensive *ab initio* Repeat Pipeline outlined in Zeng *et al.* was performed to identify and classify dispersed repeat families within the platypus and echidna. The overall workflow of this pipeline is outlined in Figure 2.1, and can be split into two distinct sub-pipelines; the identification of dispersed repeat families, and their classification.

2.1.1 Dispersed family identification

The identification of dispersed families was performed entirely using the bìogo packages `krishna` and `igor` [https://github.com/biogo/biogo] (Daniel Kortschak et al., 2017). The platypus and echidna genomes were first partitioned into 80 MB files using `seqsplit`. `krishna` was executed with default parameters of 400bp and 94% identity, based on prior runs by the laboratory. `igor` generated cluster repeat families from this pairwise alignment data, with a consensus sequence at 90% identity selected.

2.1.2 Classifying consensus sequences

The consensus sequences identified using bìogo were initially annotated using the program `RepeatMasker`, with species specific consensus sequences as the subject, and RepBase representative vertebrate repeat sequences as a query. Consensus sequences were also classified as protein, transposable element derived or retroviral based on homology to corresponding databases using `blastx` and `blastn`, as described in Zeng et al. (2018). After removing SSRs with `phobos`, the java `GenerateAnnotatedLibrary` was used to annotate the consensus sequences based on these inputs.

Observing the output of this pipeline, it was clear that active repeats needed to be specifically identified in the echidna *de novo*.

2.2 Characterising Potentially Active Repeats in *Tachyglossus aculeatus*

Reviewing the initial output of CARP did not show any evidence of currently active DNA transposons within the echidna, with no matches of significant % identity to known sequences. Two pipelines were developed to identify potentially active non-LTR and LTR retrotransposons, the latter of which is outlined in Figure 2.2.

2.2.1 Non-LTR retrotransposons

Initially, consensus sequences mapping to known repeats at over 70% homology were manually inspected, to see common families; this composition can be seen in Table 2.1.

Table 2.1: CARP (Zeng et al., 2018) classification of dispersed repeat family consensus sequences identified in the echidna, displaying over 70% with homology to query.

Repeat Class	Count
L2	8724
RTE	10
Non-LTR	216
Other	61
Total	**9011**

From this output, the 10 RTE sequences were individually inspected, but no sequences were of sufficient length to be potentially active. For subsequent analysis of non-LTR retrotransposons, only L2s were considered.

Consensus sequences mapping to L2s greater than 3000 nucleotides in length were isolated. The program `ORFfinder` was used to identify sequences with open reading frames greater than 1500 nt. These ORFs were used in a `hmmsearch`, using the RVT_1 profile as the hmmfile. Sequences with RVT_1 and EEN domain matches were selected, generating a list of potentially active sequences. These sequences were aligned, trimmed at the 5' and 3'

ends, and realigned. `CD-HIT-EST` was used to cluster these sequences at 94% identity, into 5 new subclasses of L2.

2.2.2 LTR retrotransposons

Due to the presence of flanking LTRs in LTR retrotransposons, CARP tends to misannotate potentially active LTR retrotransposons as chimeric. To correct for this, a *de novo* pipeline independent of CARP was developed for LTR retrotransposon identification.

`LTRharvest` was run using parameters outlined in Appendix C on the whole echidna genome, followed by `LTRretriever`, configured to look for canonical LTRs. The output of `LTRretriever` was queried using the `RepeatMasker` webserver, with sequences mapping to non-LTR retrotransposons discarded. A hmmsearch of these sequences was performed using the entire Pfam HMM profile database, with identified protein domains displayed using the `-domtblout` argument. Output protein domains were manually searched, and designated as *gag, pol* or *env* associated if appropriate. Sequences containing domains associated with *gag, pol* and RVT were clustered at 90% identity using `CD-HIT-EST`.

A sequence from each cluster was randomly chosen, and all sequences were aligned using `MAFFT`. Observing this alignment showed 5 distinct groups, and a member from each was chosen as a consensus sequence. These sequences were queried against the whole genome using `blastn`. Blast hits within 80% of the query length were extracted using `bedtools`, and extended by 2000 bp on either end. These extended sequences were realigned, and the presence of identical flanking LTRs was confirmed. These alignments were manually curated, and separated into subgroups. Subgroups were realigned, and an appropriate member was chosen as the consensus. ORFs from these consensus sequences were extracted using `ORFfinder`, and subsequently run through the NCBI Conserved Domains online tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), to confirm the previously identified domains are contained within an ORF. The LTRs flanking these retrotransposons were classified seperately. Potential LTRs from the extended alignment were extracted from

both the 5' and 3' ends, and realigned to each other. These alignments were examined, and groups with over 4 members were extracted and realigned. A consensus sequence was chosen, and named according to its origin.

2.2.3 Creating HMM Profiles for Dfam

Dfam (https://www.dfam.org) (Storer et al., 2021) is an open source database of repetitive DNA elements, organized around multiple sequence alignments of TE families. Rather than use single representative repeat sequences, Dfam uses seed alignments, which allow both a traditional consensus sequence and a profile hidden Markov model (HMM) to be built. L2 sequences identified in both the platypus and echidna were used to generate HMM profiles, along with the LTR retrotransposons identified *de novo* in the echidna.

Platypus L2 sequences identified as full length and containing ORF2 within RepBase, along with L2 families identified *de novo* as potentially active in the echidna, were queried against their respective genomes using `blastn`. Matches over 1000bp were extracted using `bedtools`, and aligned with `MAFFT`. A profile HMM was built from this sequence alignment using `hmmbuild`. A plurality based consensus sequence was built from this profile using `hmmemit`, and compared to the original query, to check for the presence of segmental duplications.

The LTR retrotransposons identified *de novo* in the echidna were used to generate HMM profiles. The flanking LTRs and internal regions of these retrotransposons were considered separately, according to Dfam convention. For both regions, a chosen consensus sequence was queried against the genome with `blastn`. Matches over 500bp were extracted with `bedtools`, and aligned with `MAFFT`.

2.2.4 Phylogeny of L2s and CR1s

Repeats classified as L2s or CR1s were extracted from the RepBase vertebrate library, originating from the anole, crocodile and coelocanth, along with L2 sequences from

the tuatara. Sequences identified *de novo* as L2s in the echidna and platypus over 2000 bp were included. ORFs over 200bp were extracted and used as queries for `hmmsearch`, using RVT_1 as the profile. Domains were extracted from the genome in nucleotide format using bedtools, and aligned using MAFFT. IQ-TREE was used to create phylogenies. This pipeline was performed by modifying the custom bash script `L2_CR1_extractor.sh`, shown in Appendix B.2.

2.3 Secondary *ab initio* repeat identification and annotation

Upon further review of the `RepeatMasker` search algorithm, it was decided to try using the bìogo repeat identification/annotation tool `ins` [https://github.com/kortschak/ins], developed as alternative to the tool `CENSOR`. `ins` was executed using an updated library, containing the potentially active L2s and ERVs detected in the echidna *de novo*, along with the RepBase vertebrate repeats libary.

Only sequences with over 75% coverage by a repeat were included in the collapsed libary. As the majority of L2 dispersed repeat families were 5' truncated, only sequences over 1800bp long were considered. While the over 75% rule was used to initially reduce the library size, each sequence was manually examined, to remove sequences overlapping with unrelated repeats. These *de novo* CARP derived libraries were merged with the RepBase vertebrate repeat library, for used as queries in a `RepeatMasker` search against the entire genome of each monotreme respectively, to generate a repeat annotation for the whole genome.

Quantifying Repeat Composition and Divergence

The `RepeatMasker .align` files were parsed using the perl script `parseRM.pl` [https://github.com/4ureliek/Parsing-RepeatMasker-Outputs], to give the repeat composition and divergence per chromosome. The files output by this script were subsequently run through the custom R scripts `divergence.R` and `L2_subfamily_divergence.R` (described

in Appendix B.3), to visualise composition and divergence of all repeats and L2s respectively.

2.4 Comparison of L2s to L1s

Gene data for the five mammals investigated was downloaded from NCBI in gtf format, and parsed using the perl script `agat_convert_sp_gxf2gxf.pl` [https://github.com/NBISweden/AGAT], converting the file to gff, and simultaneously giving 5' and 3' UTR region coordinates. Genes identified as protein coding were further extracted from this gff file using the "gene_biotype" flag. For the human, horse and dog genomes, `RepeatMasker` data was downloaded from NCBI, and used alongside the *de novo* `RepeatMasker` whole genome annotations previously generated for the echidna and platypus. The genomic search engine `GIGGLE` (Layer et al., 2018) was used to index these `RepeatMasker` outputs, according to the methods outlined in [https://github.com/ryanlayer/giggle].

The bash script `OverlapExtractor.sh`, shown in Appendix B.6, was used to determine overlaps between L1s (in therians) or L2s (in monotremes), and exonic, 3' UTR or 5' UTR regions, corresponding to protein coding genes. The % coverage of these elements was subsequently graphed and analysed.

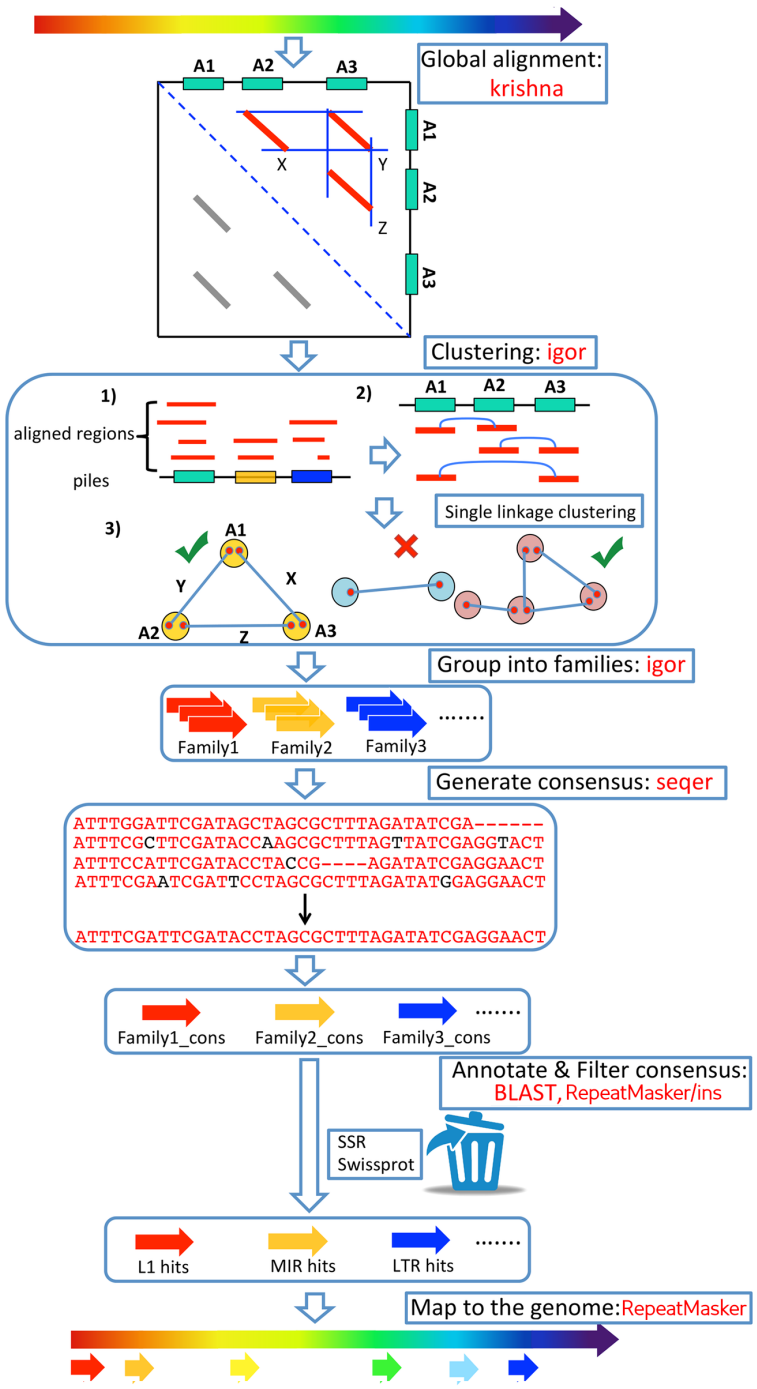


Figure 2.1: Detailed steps for the Comprehensive *ab initio* Repeat Pipeline (CARP). `krishna`: identifies repetitive regions though all to all pairwise alignment. `igor`: performs single linkage clustering to produce families of repetitive sequences. `seqer`: globally aligns families to generate a consensus sequence. `BLAST` & `RepeatMasker`/`ins`: Filters consensus sequences for non-TE protein coding genes, and annotates using RepBase, along with a library of retrovirus and reverse transcriptase sequences. `RepeatMasker` (Second run): Uses annotationed consensus sequences in reannotation of the genome, to identify lower-identity sequences excluded in earlier steps. Figure taken from (Zeng et al., 2018)

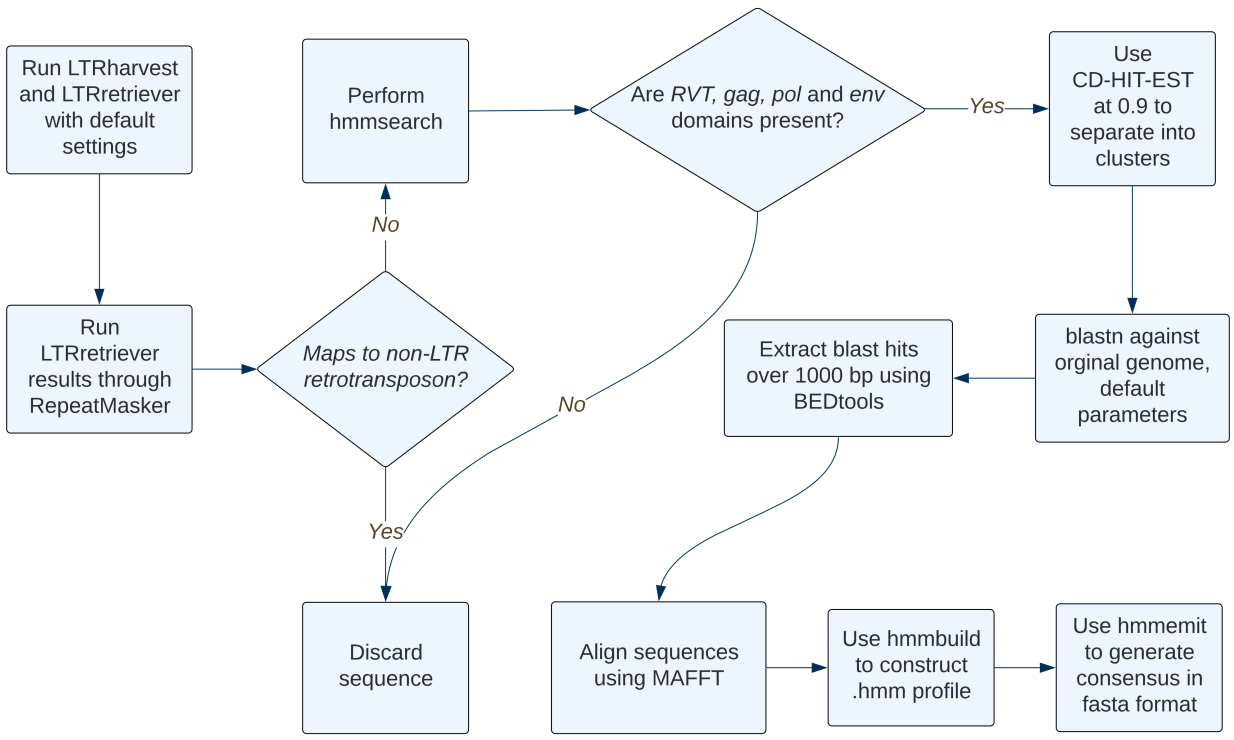


Figure 2.2: *de novo* LTR retrotransposon discovery pipeline from `LTRretriever` output, coupled with generation of HMM model for Dfam. Additional information about this workflow can be found in Appendix C.

Chapter 3: Initial *ab initio* repeat annotation

3.1 Results

Dispersed Family Identification

After briefly attempting to run an end to end `krishna` alignment of the platypus genome in a single run, it was decided that the genome needed to be split into chromosomes, which was accomplished using `bundle`. An `awk` script modified from [https://github.com/jo-mc/carp-alts] was used to write SLURM scripts for use on the Adelaide University HPC, so each pairwise alignment could be run in parallel.

While the majority of jobs submitted to the HPC succeeded, several jobs crashed due to out of memory errors. To combat this, these jobs were iteratively run with increasing memory allocations until all jobs were completed, going up to a maximum of 1 terabyte.

To avoid these out of memory errors, the echidna genome was split into smaller sections than the platypus, into segments with a maximum size of 80MB. Even with smaller alignments, a large number of out of memory errors still occurred, and the process of iterative memory allocation increases had to be repeated.

Once the `krishna` alignment for both genomes was complete, `igor` and `seqer` were used to report dispersed family groupings and generate consensus sequences respectively. The final number of consensus sequences can be seen in Table 3.1, along with the genome size.

Table 3.1: Key statisistics from the initial run of CARP

Species	Genome Size (Gb)	No. Consensus Sequences
Platypus	1.9	16202
Echidna	2.2	27192

Classifying Consensus Sequences

`RepeatMasker` was run using recommended settings, masking low complexity repeats. The CARP specific java `GenerateAnnotatedLibrary` was used to classify the previously generated consensus sequences.

3.2 Discussion

Dispersed Family Identification

While it is essential to split the genome up before attempting to run `krishna`, deciding what resources to allocate to each job based on fasta length alone is difficult, as the memory requirement of each pairwise alignment is dependent on repeat abundance, length and composition.

The increased number of consensus sequences in the echidna relative to the platypus indicates a higher percent repeat content, especially when genome size is considered; if true, this would account for the increased memory requirements of the echidna.

Classifying Consensus Sequences

Although `RepeatMasker` was chosen as the program for annotating these libraries, the CARP pipeline was designed for use with the tool `CENSOR`. The `CENSOR` tool has lost support since 2016, and is no longer considered an optimal choice for genome annotation. An alternative to `CENSOR`, `ins` was initially considered, but frequent errors in inititalisation made `RepeatMasker` the tool used for this first run of CARP.

Chapter 4: Classifying Potentially Active Retrotransposons in *Tachyglossus aculeatus*

4.1 Results

4.1.1 Non-LTR Retrotransposons

Through completion of the non-LTR retrotransposon pipeline outlined in the methods, only 10 consensus sequences contained all the characteristics indicative of a potentially active L2, from the total of 24,847 consensus sequences identified as L2s (the distribution of which can be seen in Figure 4.1). The constituents of these consensus sequences were first extracted, then extended by 1,000 bp in either direction, to ensure potential 5' and 3' ends were included. These sequences were subsequently aligned to several platypus derived L2s, to observe features common to the 5' and 3' ends, which can be seen in Figure 4.2.

This 3' end in Figure 4.2 can be distinguished quite easily, as the echidna L2s appear to share a similar terminating k-mer. Initially there was some difficulty in determining the 5' end of the echidna, as alignments seemed to widely vary both between and within families, and did not align to any platypus derived repeats. Upon closer inspection, it was observed that these potentially active L2 sequences had a k-mer of 103 bp that repeated between 1 and 6 times at the 5' end, an example of which is shown in Figure 4.3.

Once the 5' and 3' ends had been resolved, sequences were manually inspected to ensure there were no large indels present. The 18 remaining sequences were clustered via `CD-HIT-EST` at 94 % identity, giving 5 representative sequences labelled `Tacu_L2a` to `Tacu_L2e`, representatives of which can be found in Appendix E.1.

4.1.2 LTR Retrotransposons

Observing the output of CARP, there wasn't a single sequence which mapped to a LTR retrotransposon with over 75% coverage, greater than 3000 bp in length. To see whether there were any LTR retrotransposons being misannotated, the software `LTRharvest` and `LTRretriever` was used, as an alternative *ab initio* transposon detection approach to CARP.

The tool `LTRharvest`, run using settings outlined in Appendix C, identified 50,656 potential LTRs over 1500bp. To remove redundancy and reduce the rate of false positives, these sequences were used as the inputs for `LTRretriever`, which detected 232 non-redundant canonic LTR retrotransposon candidates.

Although a TE specific .hmm profile provided through `LTRretriever` was supposed to remove L2s during this search based on the presence of an L2 specific RVT domain, a `RepeatMasker` run using the vertebrate representative RepBase libary on this output showed many false positives. After manually removing sequences mapping to non-LTR retrotransposons, 80 sequences remained.

To ensure potentially active sequences with low homology to previously characterised repeats were still considered, a hidden Markov model search was used to detect protein domains indicative of an active LTR retrotransposon. Using the pipeline outlined in the methods, consensus sequences containing at least *gag* and *pol* associated domains were separated, aligned and clustered at 85% identity using `CD-HIT-EST`, separating into 6 distinct clusters.

Sequences from each cluster were run through the online ORFfinder tool manually, to confirm that these protein domains overlapped fully with an ORF. From this, a single sequence containing appropriate ORFs and domains was selected as the representative sequence for that class. These sequences were run through the `RepeatMasker` online tool, against the entire RepBase repeat library, shown in Figures 4.4 and 4.5, to determine whether the sequences belonged to a class of LTR retrotransposon previously identified in platypus. Representative sequences from each of the 6 families can be found in Appendix E.2.

4.1.3 HMM Model Generation for Dfam

The pipeline for generating HMM models specifically for submission to Dfam was performed in a non-canonic way, directly using `blastn` (outlined in `dfamgenerator.sh`, Appendix B.1), instead of through Perl scripts provided by the tool `RepeatModeler`. Despite this, the alternate method of library generation provided a large number of samples for inclusion.

4.1.4 L2 Phylogenies

To better observe the evolutionary relationships between L2s of different species, a maximum likelihood phylogeny of aligned RVT domains from L2s in several vertebrates was inferred using `IQ-TREE`, shown in Figure 4.6.

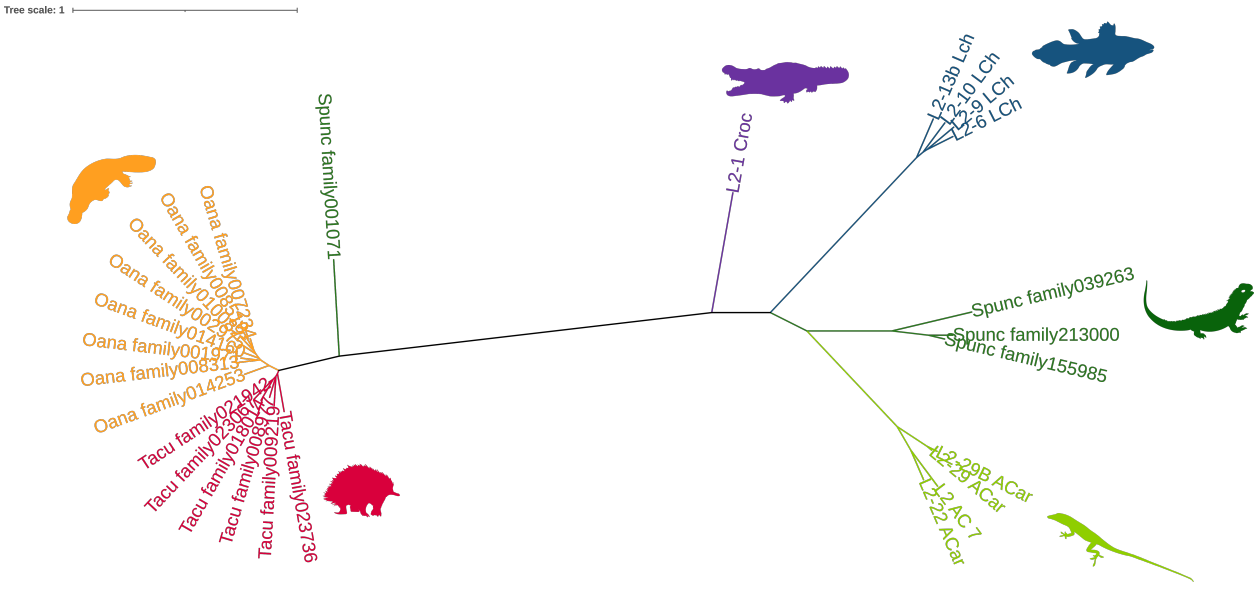


Figure 4.6: Maximum likelihood tree, showing the phylogeny of the reverse transcriptase domain contained within ORF2 of the LINE-2 retrotransposons, found in vertebrates. Species included have multi-copy full length LINE-2 elements, which are presumed active. Clockwise, species included are the short-beaked echidna (*T. aculeatus*), platypus (*O. anatinus*), tuatara (*S. punctatus*), saltwater crocodile (*C. porosus*), coelocanth (*L. chalumnae*) and green anole (*A. carolinensis*). Silhouettes sourced from creazilla.com and phylopic.org.

4.2 Discussion

Determining whether a transposon has the potential for active transposition is difficult; while the best way would be to directly infer germline or somatic transpositions, this is not possible without specialised analysis of long-read sequencing data (Pendleton et al., 2015). In this analysis I have taken the preliminary step of identifying sequences with potential retrotransposition activity, which may be strengthened using RNA-seq data, or the methods described above.

4.2.1 Non-LTR Retrotransposons

Using the pipeline outlined in the methods, only 139 of the 24,847 consensus sequences mapped to L2s with RepeatMasker contained ORFs over 1500bp, the approximate size of ORF2 in platypus L2s. The majority of L2s were too short to contain the components needed for L2 activity, which would be somewhere between 4000 and 5000 bases.

Three Prime End

Figure 4.2 shows an alignment between potentially active echidna L2 sequences identified *de novo*, and Playtpus L2s submitted to RepBase. From observation, it appears that there are four k-mer variants amongst these monotreme L2s, TGA, TGAT, TGAA and TAC. While echidna L2s all possess either the TGA or TGAT variant, these two k-mers can be seen within members of the same consensus family, `family014611`. Upon closer examination of the RepBase descriptions of platypus L2s with this TGAT k-mer (L2_Plat1m to L2_Plat1t), it can be seen that these sequences are all severely 5' truncated, at 2-3kb each, and have been constructed from inactive elements.

The fact that this k-mer sequence is shared between inactive platypus L2s and potentially active echidna L2s is unexpected; this could be an indication that echidna L2s represent an archaic form of this LINE family. It should also be noted that even with extension of these echidna sequences at the 3' end, the expected TSD downstream of these

sequences could not be found.

Five Prime End

The five prime end of these retrotransposons was particularly hard to resolve, as a repeating 103bp sequence caused a large degree of variation, even between families. This sequence can be clearly seen in Figure 4.3, in which the sequence repeats four times. This variation in sequence count between potentially active L2s is visible in the supporting figures of Appendix S2, which contains extended self- alignments of all potentially active echidna L2s.

This repeating sequence does not appear in the platypus genome at all using a `blastn` search with the `megablast` algorithm, and only low length, low confidence hits are reported using the `blastn` algorithm. Additionally, extending and self-aligning platypus L2s did not show a different repeating sequence in any scenario, and no 5' repeating sequences are mentioned by Warren et al. (2008). Repeating this blast gave a high number of hits in the echidna, on every chromosome, shown in Supplementary Figure S1. Interestingly, it appears that over half of these hits have mapped to unplaced contigs in the echidna, which only make up 3% of the total assembly.

The ubiquitous presence of this repeat within all L2s identified as potentially active strongly indicates that this sequence is being replicated along with the rest of the element, and thus a single copy was included within the representative sequences. It would be very interesting to look at transcriptomics data to see if this repeat is included in the mRNA transcript, and to see if and how it is included in somatic retrotransposition.

4.2.2 LTR Retrotransposons

Through the completion of this LTR pipeline, it appears that there is greater diversity of potentially active LTR retrotransposons than L2s. All LTR retrotransposons identified also seem to be ERVs, as they possess an *env* gene. Figure 4.4 shows that three of these ERVs belong to the same family as platypus ERVs; conversely, Tacu_ERV4, Tacu_ERV5

and Tacu_ERV6 all have high coverage, low homology matches to various other ERVs and LTR retroelements, and represent classes of ERV not previously identified in the platypus.

One ERV of particular interest is Tacu_ERV4. Upon closer inspection, this repeat appears to have a 1kb domain, belonging to the TLV Coat superfamily. The presence of this domain would classify this ERV as a member of the Orthoretrovirinae, a subfamily of retroviruses heavily implicated in disease, capable of individual to individual transmission. A `blastn` search shows that this virus displays $\sim 70\%$ homology to a variety of Reticuloendotheliosis viruses, an immunosuppressive pathogen with a variety of avian hosts. A wider scale study into the expression of this virus, along with its presence would be of interest.

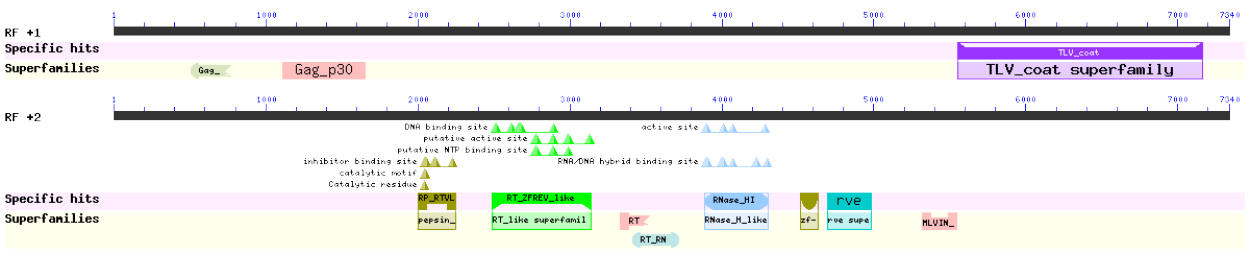


Figure 4.7: Protein domains of Tacu_ERV4, identified using the NCBI conserved domain search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), using the CDD v3.19 database.

4.2.3 HMM Model Generation for Dfam

Although HMM profiles are one of the best ways to represent a repeat family, monotreme derived repeat HMM profiles have never been generated. The inclusion of these sequences within the public Dfam library will allow them to be used in future research, and provide a more accessible public record of these potentially active elements. L2 sequences from both platypus and echidna, along with echidna ERVs have been prepared for submission to Dfam in MSA Stockholm format (according to the guidelines set out in Storer et al. (2021)); these sequences are scheduled to be uploaded at the same time as an upcoming paper.

4.2.4 L2 Phylogenies

Figure 4.6 clearly shows that L2s from echidna and platypus are distinctly grouped from L2s in the other vertebrates analysed, except for in the case of the tuatara. In (Gemmell et al., 2020), it was shown there are two distinct families of tuatara L2, suggesting a possible horizontal transfer event with monotremes; this finding has been confirmed here. Additionally, it can be seen that within monotremes, L2s are clustered by species. Assuming these repeats have remained active, this indicates that active L2s in these two species have been evolving separately. The absence of a paraphyletic group would also suggest that there is a single group of L2s active in each monotreme, unlike in the tuatara.

Additional analysis comparing L2s to CR1s can be seen in Supplementary Figure S6. This was performed due to the fact that platypus L2s are labelled in RepBase as belonging to the CR1 subclass, not L2. Although this was assumed to be a submission error, a maximum phylogeny tree was constructed in the same manner as Figure 4.6, including RVT sequences from the ORF of CR1s as well. This confirmed that L2s are distinct from CR1s, and that the RepBase submission was most likely misannotated.

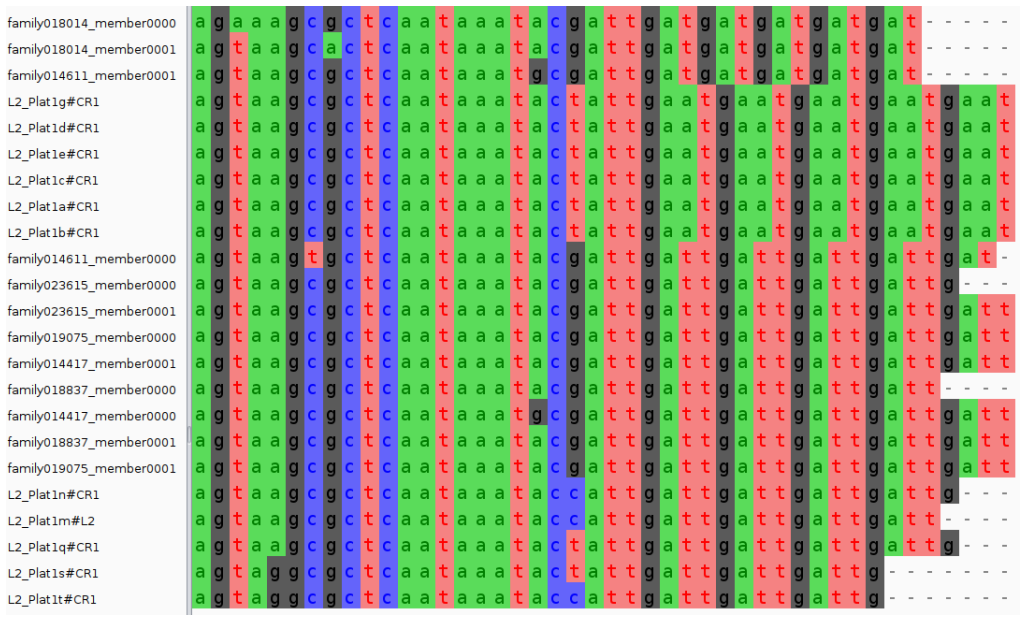


Figure 4.2: Alignment of CARP (Zeng et al., 2018) derived echidna consensus sequences containing potentially active L2s (labelled `family.....`), along with RepBase platypus L2 consensus sequences (labelled `L2_Plat1.`), zoomed to the 3' end.

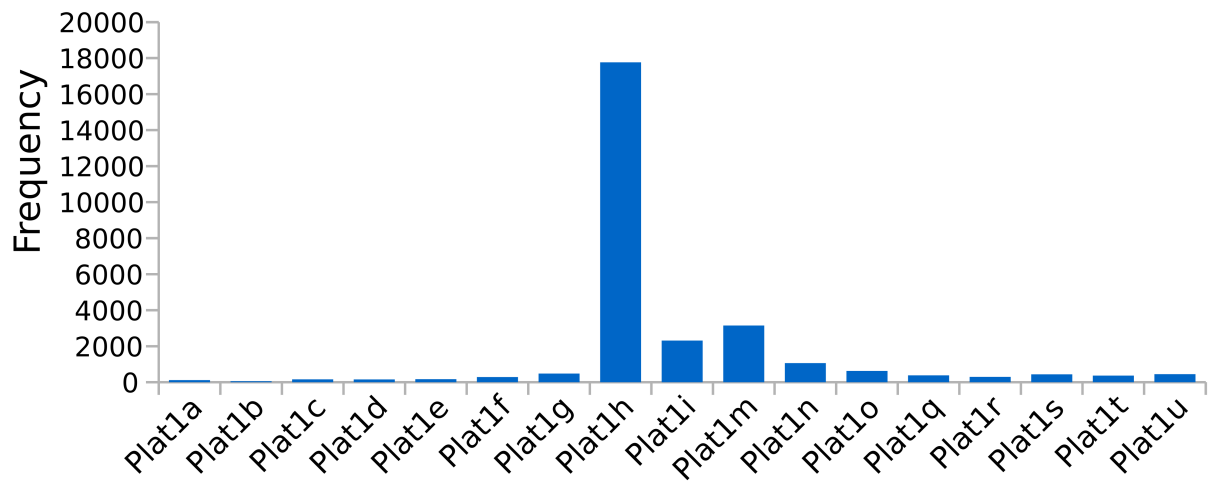


Figure 4.1: Distribution of primary annotations in echidna dispersed repeat families by RepBase derived platypus LINE-2 elements.

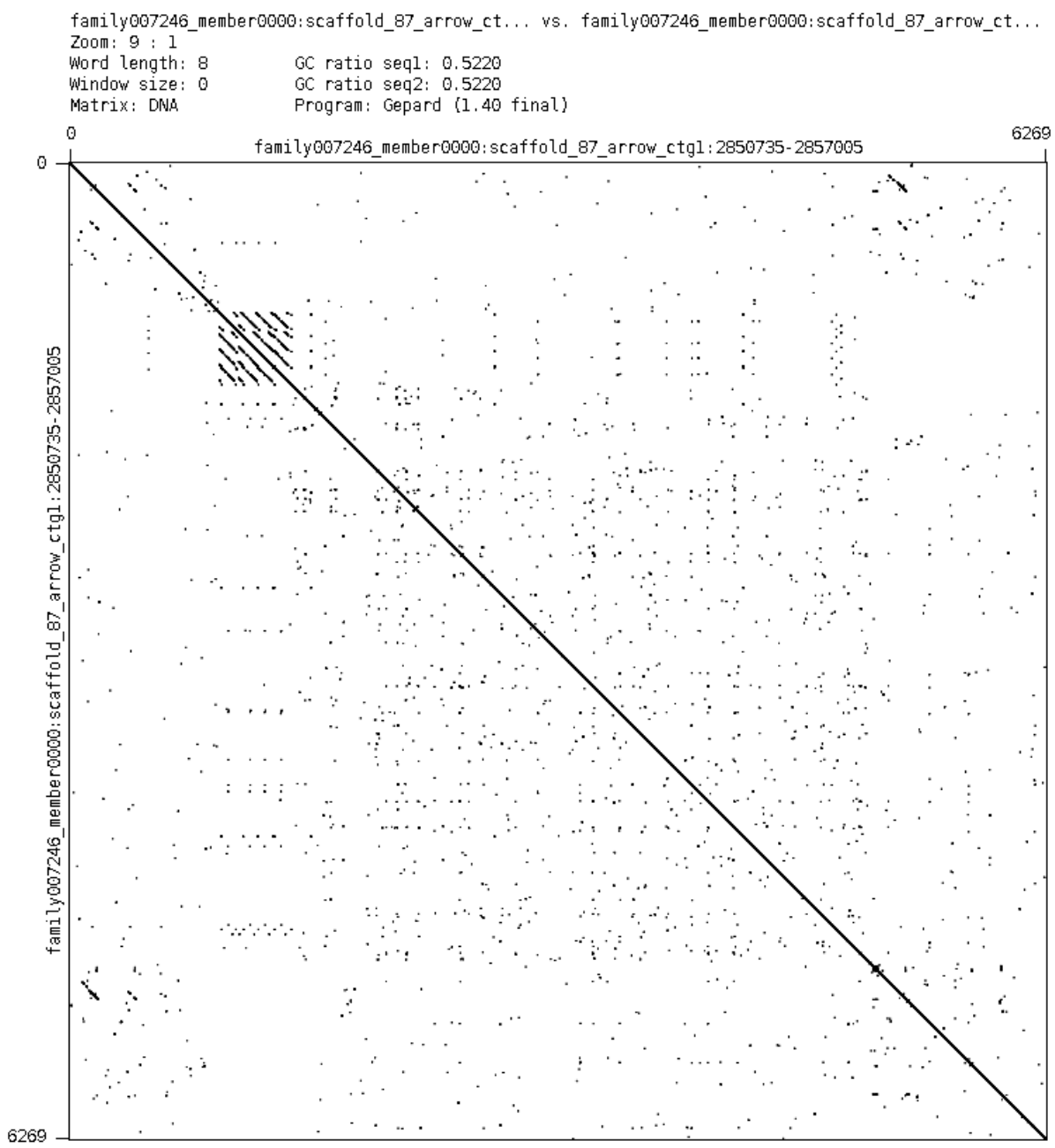


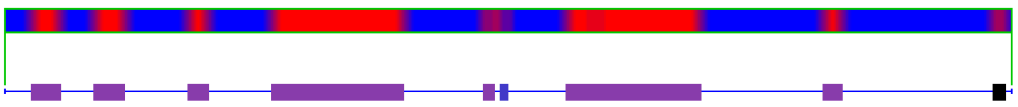
Figure 4.3: Stranded self-alignment of a potentially active LINE-2 sequence identified in the echidna, extended by 1000 base pairs in the 5' and 3' directions, visualised using the tool `gepard`. Note the repeating 103 base pair sequence at the 5' end of the LINE-2, which repeats four times. Additional dotplots showing the variability of this repeat number can be seen in Appendix D.

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
Tacu_ERV1	3	6966	PlatERVK2b	3	6986	ERV/ERV2	d	0.6409	2.0550	8777

(a) Tacu_ERV1

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
Tacu_ERV2	6	1468	ERV3-1_OAn-I	2	1504	ERV/ERV3	d	0.6850	1.7946	3453
Tacu_ERV2	1499	1542	COP7_I_MT	737	780	LTR/Copia	d	0.8667	1.3333	273
Tacu_ERV2	1766	1818	Submariner_APa	749	801	DNA/Mariner	c	0.7778	2.0000	221
Tacu_ERV2	2069	2163	HERVFH19I	2458	2552	ERV	d	0.7188	1.6923	259
Tacu_ERV2	2195	5079	PlatERVR1	937	3809	ERV/ERV1	d	0.6674	1.9934	6821
Tacu_ERV2	5247	5316	ERV15_MD_I	5991	6057	ERV/ERV1	d	0.7647	2.3333	240
Tacu_ERV2	5323	5562	PlatERVR2b	4241	4513	ERV/ERV1	d	0.6898	1.7368	565
Tacu_ERV2	5644	5777	Gypsy-13_CySi-I	7343	7466	LTR/Gypsy	c	0.7236	1.5294	265
Tacu_ERV2	6095	6153	L1-51_MT	2115	2169	NonLTR/L1	c	0.8214	1.4000	266
Tacu_ERV2	6623	7112	ERV3-1_OAn-I	1659	2149	ERV/ERV3	d	0.6157	1.7835	780
Tacu_ERV2	7319	7412	ERV3-1_OAn-I	2255	2344	ERV/ERV3	d	0.7447	1.3571	351

(b) Tacu_ERV2



Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
Tacu_ERV3b	155	317	PlatERVK1	143	299	ERV/ERV2	d	0.7215	2.5294	530
Tacu_ERV3b	516	690	ERV2-18_MM-I	1096	1270	ERV/ERV2	d	0.6761	1.4516	331
Tacu_ERV3b	1051	1163	ERV2-5_StHa-I	531	643	ERV/ERV2	d	0.6579	1.6818	292
Tacu_ERV3b	1526	2278	PlatERVK1	2527	3279	ERV/ERV2	d	0.6770	2.0396	1323
Tacu_ERV3b	2734	2787	ERVH_OC_I	4160	4213	ERV/ERV2	d	0.7407	2.0000	222
Tacu_ERV3b	2831	2870	Tx1-11_LMi	523	562	NonLTR/Tx1	c	0.8250	2.3333	232
Tacu_ERV3b	3202	3362	EqERV-beta1	6243	6402	ERV	d	0.6993	1.5714	237
Tacu_ERV3b	3374	3969	PlatERVK1	4345	4940	ERV/ERV2	d	0.6852	1.8889	1462
Tacu_ERV3b	4668	4774	LTR71C_TS	45	147	ERV	d	0.7115	1.3750	234
Tacu_ERV3b	5635	5698	REP-4_DR	9575	9639	Interspersed_Repeat	d	0.7538	1.0909	275

(c) Tacu_ERV3b

Figure 4.4: Repeat annotation of platypus-like endogenous retroviruses identified *de novo* in the echidna. Annotation was generated by [www.girinst.org/censor/], using the RepBase repeat library.

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
Tacu_ERV4	1	25	MPERV1_INT	1	25	ERV/ERV1	d	1.0000	99.0000	208
Tacu_ERV4	469	780	ERV1-1_PVa-I	317	624	ERV/ERV1	d	0.6711	1.7660	480
Tacu_ERV4	1090	5470	ERV1X2-I_ML	1162	5559	ERV/ERV1	d	0.6545	1.9071	7637
Tacu_ERV4	5611	7148	ERV2-5_Pca-I	5703	7231	ERV/ERV2	d	0.6218	1.7475	2252

(a) Tacu_ERV4

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
Tacu_ERV5	1	141	ERV1-1B_PVa-I	1	129	ERV/ERV1	d	0.7045	1.7059	254
Tacu_ERV5	698	916	ERV1-1_ET-I	482	709	ERV/ERV1	d	0.6802	1.8485	419
Tacu_ERV5	1248	1281	Copia-15_AT-I	131	164	LTR/Copia	d	0.8529	1.2500	226
Tacu_ERV5	1341	1560	PlatERVP1	679	896	ERV/ERV1	d	0.6301	1.9500	394
Tacu_ERV5	1581	1625	Gypsy-14_ATr-LTR	1845	1888	LTR/Gypsy	c	0.8444	1.6667	235
Tacu_ERV5	1902	1973	IS3EU-N1_EL	1041	1115	DNA/IS3EU	d	0.7067	1.4615	224
Tacu_ERV5	2017	2231	HERVP71A_I	2009	2217	ERV	d	0.6540	1.5000	367
Tacu_ERV5	2362	3345	ERV24_MD_I	4618	5601	ERV/ERV1	d	0.6454	1.9689	1821
Tacu_ERV5	3481	3706	PTERV2b	5494	5740	ERV/ERV1	d	0.6812	1.9118	417
Tacu_ERV5	4539	4663	ERV1A3_CPo-I	6477	6601	ERV/ERV1	d	0.6772	1.6087	323
Tacu_ERV5	4875	4943	Gypsy-36_AIp-I	4704	4769	LTR/Gypsy	c	0.7246	1.5000	218
Tacu_ERV5	5016	5491	ERV1-3B_TSy-I	4649	5122	ERV/ERV1	d	0.6512	1.9114	834

(b) Tacu_ERV5

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
Tacu_ERV6	20	96	PRIMA41	7656	7731	ERV	c	0.8205	2.7500	388
Tacu_ERV6	356	5465	CarERV2	920	5967	ERV/ERV1	c	0.6429	1.8731	4204
Tacu_ERV6	6019	6063	GYMAG2_I	4606	4653	LTR/Gypsy	d	0.7660	1.1250	205
Tacu_ERV6	6117	6192	Gypsy-23_ADe-I	2760	2830	LTR/Gypsy	c	0.7945	1.8000	216
Tacu_ERV6	6307	6398	ERV2-5_CPo-I	1	96	ERV/ERV2	c	0.7667	1.7500	207

(c) Tacu_ERV6

Figure 4.5: Repeat annotation of novel endogenous retroviruses identified *de novo* in the echidna. Annotation was generated by [www.girinst.org/censor/], using the RepBase repeat library.

Chapter 5: Secondary *ab initio* Repeat Annotation and Analysis

5.1 Results

5.1.1 Comparing `RepeatMasker` to `ins`

Upon describing the five L2 and six ERV retrotransposons specific to the echidna, the CARP pipeline was run again on both the echidna and platypus genomes, with these sequences incorporated into the RepBase vertebrate library. Upon reflection based on the initial CARP paper, a decision was made to use the tool `ins` for repeat classification, instead of `RepeatMasker`. It should be noted that the Adelaide University HPC experienced a total shutdown for over 3 weeks due to a security incident during this analysis, cancelling current jobs and preventing parallelised genome annotations from occurring; having to run each job individually on the laboratory computer pushed these results back by several weeks. Table 5.1 shows final CARP classifications of the 27,192 echidna consensus sequences, and differences between their annotations when run using `RepeatMasker` and `ins`.

Table 5.1: Differences in the classification of dispersed repeat consensus in echidna, identified using `krishna`, annotated using `RepeatMasker` or `ins`, and classified via CARP (Zeng et al., 2018).

	RepeatMasker	**Ins**
>90% Coverage	5984	1987
<90% Coverage	12996	21817
Chimeric	5521	1838
Retroviral	79	175
Unclassified	2612	1330

From this, it was decided that `ins` would be used for repeat classification in place of `RepeatMasker`, to increase stringency and reduce the rate of false positives. These consensus

annotations were manually curated to only include sequences mapping at 75% identity or higher, to ensure that only consensus sequences that represented transposon/retrotransposon derived dispersed families were included.

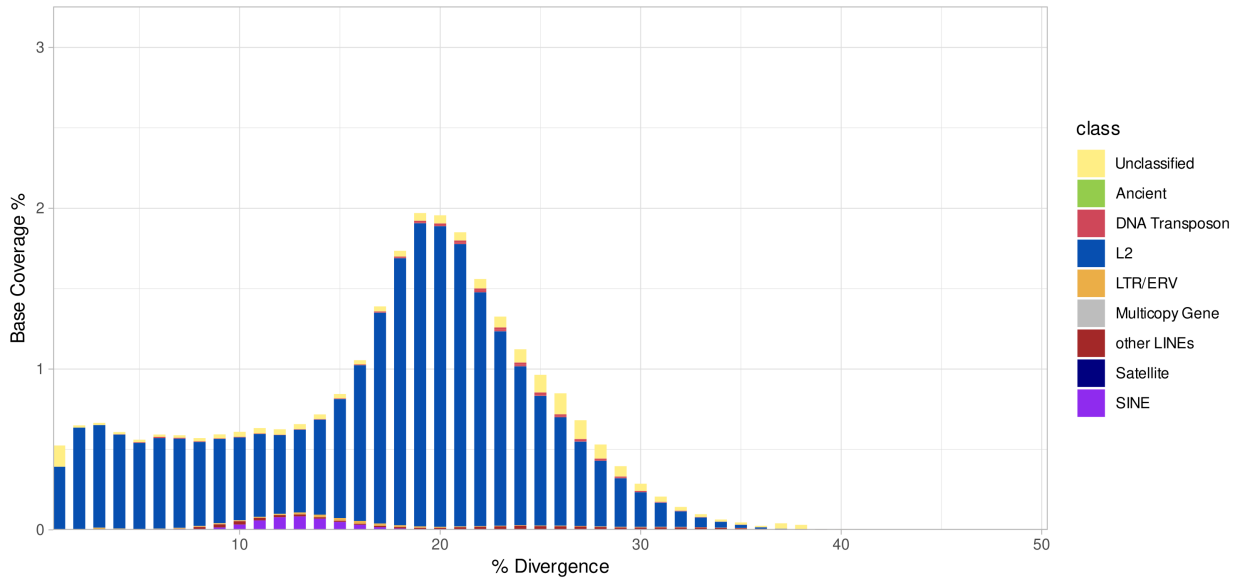
In addition to this, the majority of dispersed repeat families identified in the echidna now mapped to the newly described echidna L2s, as was expected.

While an initial attempt was made to use `ins` for annotation of the whole genome, the computational requirements of this task were extremely high. Instead, `RepeatMasker` was used for the final genome annotation of both the echidna and platypus, using the `ins` classified combined repeat library.

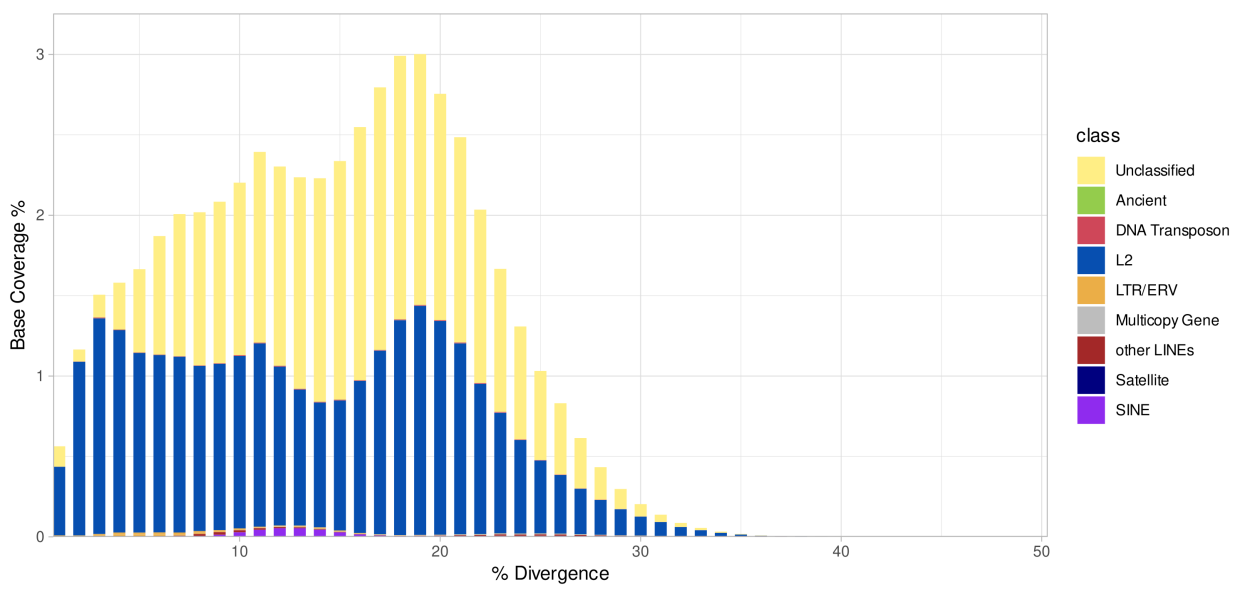
5.1.2 Quantifying Repeat Composition and Divergence

The perl script `parseRM.pl` [github.com/4ureliek/Parsing-RepeatMasker-Outputs] was used to calculate the % repeat coverage at both the class and individual repeat level. The overall composition of repeats in the platypus and echidna can be seen in Tables 5.2 and 5.3 respectively.

The custom R scripts `divergence.R` and `L2_subfamily_divergence.R` (see Appendix B.3) were used to produce graphs showing the coverage and percent identity of `RepeatMasker` masked repeats, in the echidna and platypus, for repeat classes and monotreme specific L2 subfamilies respectively. The whole genome graphs for repeat classes and L2 subclasses can be seen in Figures 5.1 and 5.2, with associated per chromosome data in Appendices F-I.

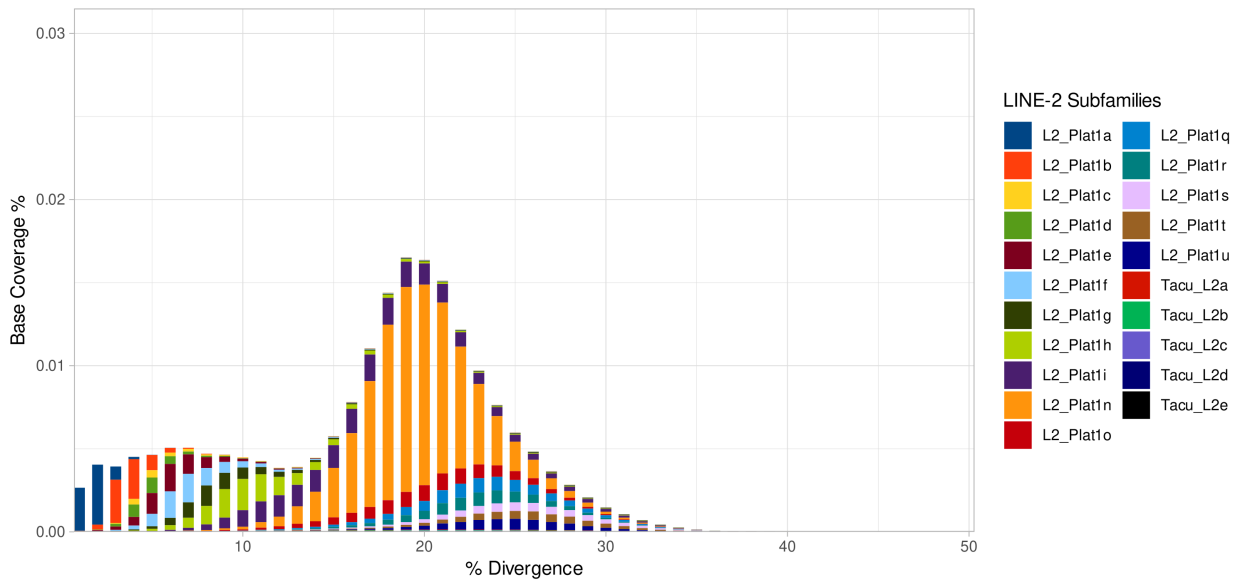


(a) Platypus

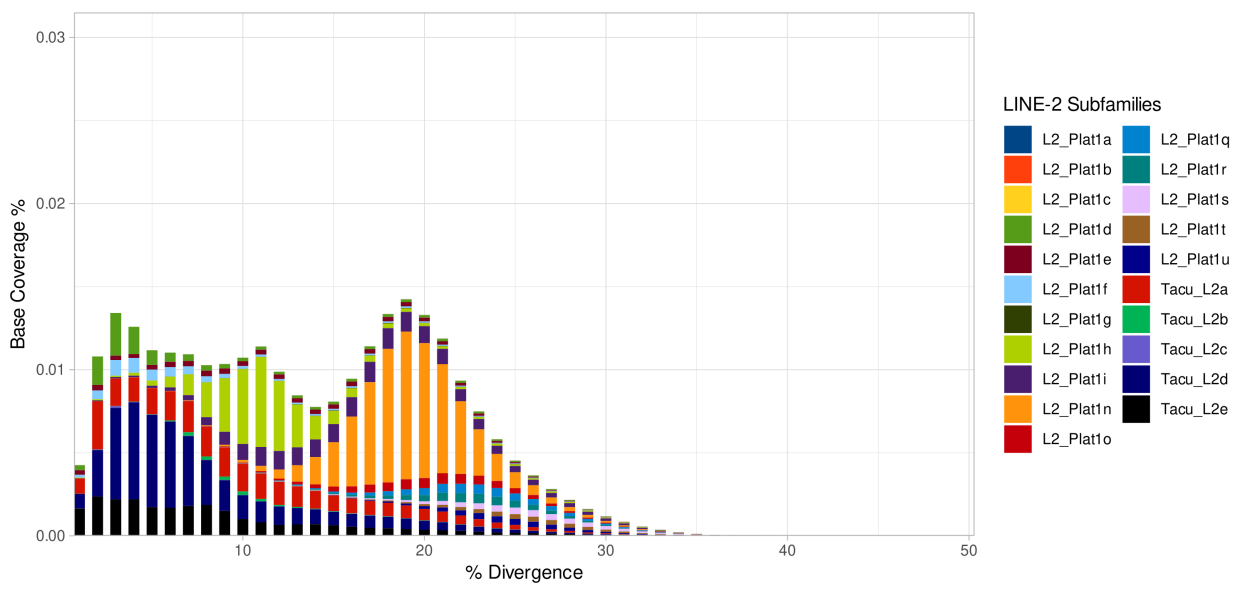


(b) Echidna

Figure 5.1: Graphs generated using the R script `divergence.R`, showing coverage of the genome by repeats identified using `RepeatMasker` in the platypus and echidna. Repeats have been divided into bins, based on % divergence from their corresponding query. Base coverage % refers to the percent of the entire genome covered by a specific element.



(a) Platypus



(b) Echidna

Figure 5.2: Graphs generated using the R script `L2_subfamily_divergence.R`, showing coverage of the genome by monotreme specific LINE-2 elements, identified using `RepeatMasker`. Dispersed repeat families identified through CARP Zeng et al. (2018) as LINE-2 elements have been included. Repeats have been divided into bins, based on % divergence from their corresponding query. Base coverage % refers to the percent of the entire genome covered by a specific element.

5.2 Discussion

5.2.1 Quantifying Repeat Composition and Divergence

Based on the output of the `RepeatMasker` whole genome annotation, summarised in Tables 5.2 and 5.3, it appears that the echidna genome is significantly more repetitive than that of the platypus, with 53% repeat coverage compared to 28%. This result matches the prediction made in Chapter 4, that the increased number of `igor` consensus sequences was due to a higher repeat content.

Interestingly, the echidna genome was annotated as having significant coverage of "unclassified" repeats. Unclassified repeats are defined by CARP as consensus sequences with less than 50% coverage by known repeats during annotation. This essentially means that 25% of the echidna genome is covered by dispersed repeat families that could not be classified using the existing RepBase library. Additionally, the base coverage of these unclassified repeats appears to peak around the same point as the L2s, which may be an indication of coupled activity. There are multiple reasons why this high rate of unclassified sequences may have been observed; it may be an artifact of the assembly process, where the combination of long read sequencing and the assembly of a haploid genome may have produced haplotype blocks. In the recently published tuatara genome paper (Gemmell et al., 2020), which also used CARP to classify repeats, a similarly high number of unclassified repeats were found, a large proportion of which were determined to be segmental duplications. It is also possible that these sequences represent something entirely new, potentially a new group of transposable element. Initially looking into the length and copy number of these repeats would give an indication as to their nature, but this was not possible in the timeframe of this project. Overall, this result warrants further analysis.

It should also be noted that, despite 6 potentially active ERV families being detected in the echidna genome, LTR retrotransposons only displayed a coverage of ∼0.25%. While this is significantly lower than coverage by L2s, this pattern is not uncommon in mammals

(Platt et al., 2018), and still represents over 5 million base pairs.

The two R scripts `divergence.R` and `L2_subfamily_divergence.R` were specifically written to parse and view repeat coverage at the chromosome level. This decision was made through consultation with the Grützner lab (grutznerlab.weebly.com), to visualise repeat dynamics in the monotreme sex chromosomes. Interestingly, it appears that both the X and Y chromosomes harbour an increased number of transposons, relative to autosomes. While this pattern is typical in the heteromorphic sex chromosome, which is unable to recombine (Chapolin, 2015), the X chromosomes should behave the same as autosomes in terms of long term repeat accumulation. The reasons behind this discrepancy are unknown, and should be explored in the future.

Figure 5.2 shows the composition, coverage and divergence of LINE-2 elements within the echidna and platypus genome. Waves of repeat expansion can quite clearly be seen within both species. Interestingly, the inactive repeat L2_Plat1n appears to show a similar pattern of coverage and divergence in both monotreme species; this could be considered evidence of an expansion of L2s, prior to the platypus-echidna split. L2s with the lowest levels of divergence only appear within a single species, giving evidence for the idea that these repeats have been evolving seperately after the species split. It can also clearly be seen that the echidna has had a more recent burst of L2 activity, compared to the platypus.

A brief note on SINEs

The coverage by SINEs obtained by `RepeatMasker` was 0.48 and 0.30 percent for the platypus and echidna respectively. This result was unexpected; coverage like this is significantly lower than the average % SINE coverage in mammals. While writing this thesis, I discovered that inclusion of the `RepeatMasker` flag `-norna` was causing this discrepancy. Due to the way that SINEs were described in the fasta description line of my custom repeat library, `RepeatMasker` was not specifically identifying these sequences as SINEs; the `-norna` flag leaves sequences not mapping to what `RepeatMasker` is told is a SINE, but that map to small pol III transcribed RNAs (which would include the majority of SINEs) as unmasked,

and thus undetected in subsequent analysis (this can be seen in the `RepeatMasker` documentation, www.animalgenome.org/bioinfo/resources/manuals/RepeatMasker.html). Rerunning `RepeatMasker` without this flag on chromosome 1 of the echidna showed that this was indeed the issue, shown in Figure 5.3.

Additionally, the SINEs identified without the `-norna` flag show divergence levels between 5 and 25%, indicating it is likely there are no active SINEs in my combined repeat library. It would be interesting to perform a *de novo* search for SINEs, and subsequent tests for potential activity.

While this means that the SINE data obtained for this analysis is not accurate, it should not affect any other classifications, and does not directly relate to my hypothesis; although rerunning `RepeatMasker` for the entire echidna and platypus genome was not achievable due to time pressures, this would be performed for future analyses.

5.3 Closing Thoughts on CARP

Although `RepeatMasker` and `ins` are both supported by CARP, the outputs of these repeat identification programs are vastly different. While `RepeatMasker` seems to be better at finding dispersed families, the specialised `rmblast` search algorithm is unclear with its edges, which may cause false positives. Additionally, the % divergence statistic given by `RepeatMasker` is not implemented into the CARP java `GenerateAnnotatedLibraries`; thus a high coverage result may be misleading. The results of `ins` are more transparent and stringent, which should give fewer false positives, but may miss undescribed or underrepresented repeats.

Additionally, the classification portion of CARP does not work with LTR retrotransposons, and will always classify full length copies as chimeric. Some sort of separate LTR detection pipeline must be implemented in parallel, to compensate for this.

Additionally, while several SINEs were detected and classified through the dispersed

repeat family identification portion of CARP, the minimum dispersed repeat family length of 400bp (which was chosen for memory use purposes) means that many SINEs will be missed. To remedy this, I would propose the integration of a separate SINE identification pipeline to CARP.

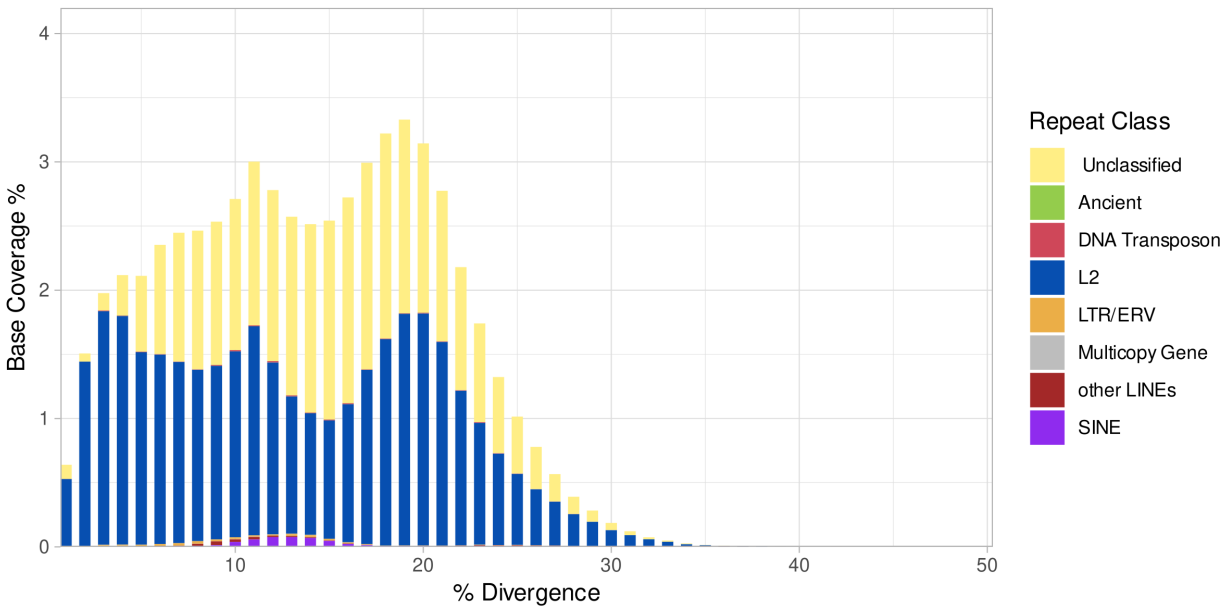
In summary, the comprehensive *ab initio* pipeline appears effective at initial dispersed family identification, but requires major overhauls to its classification system; namely in the more effective integration of `RepeatMasker`, and LTR retrotransposon detection.

Table 5.2: Platypus genome coverage by interspersed repeats

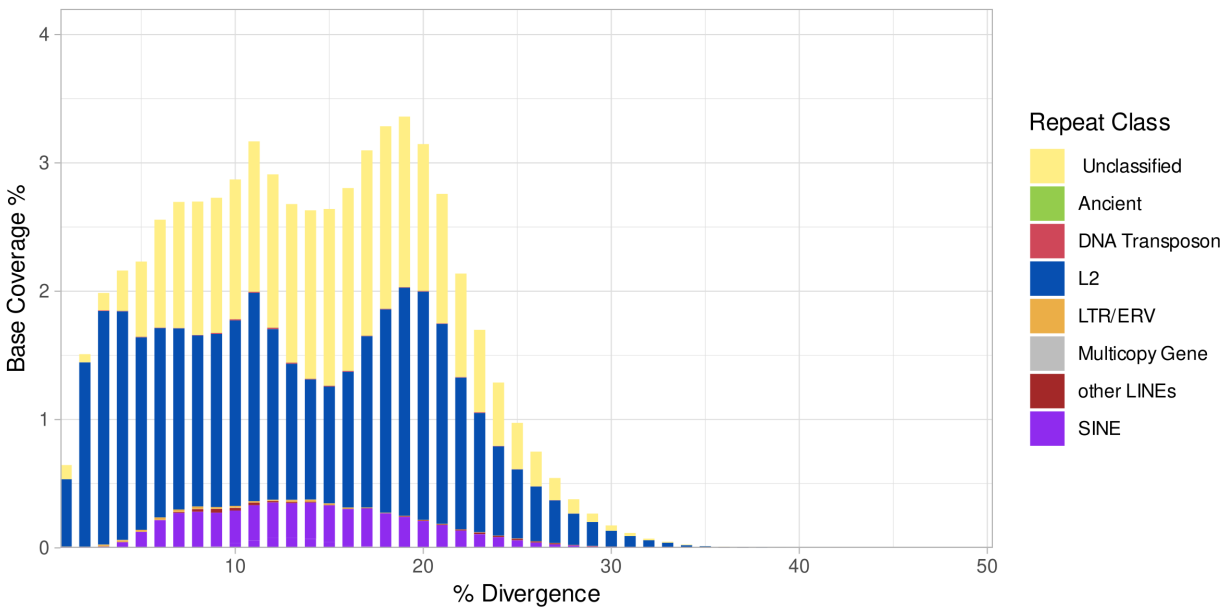
Class	Subclass	Coverage (bp)	Coverage (%)
Non-LTR			
	LINE-2	459388323	24.708
	Other	177178	0.010
	SINE	9046546	0.487
LTR		4171971	0.224
DNA Transposon		3844599	0.207
Other		19379	0.001
Unclassified		30237960	1.626
Total		**506885956**	**27.262**

Table 5.3: Echidna genome coverage by interspersed repeats

Class	Subclass	Coverage (bp)	Coverage (%)
Non-LTR			
	LINE-2	539192130	26.565
	Other	4170536	0.205
	SINE	6205302	0.306
LTR		4983886	0.246
DNA Transposon		2419412	0.119
Other		1981078	0.098
Unclassified		527188796	25.973
Total		**1086141140**	**53.511**



(a) With the `Repeatmasker -norna` flag



(b) Without the `Repeatmasker -norna` flag

Figure 5.3: Graphs generated using the R script `divergence.R`, showing coverage of the genome by repeats identified using `RepeatMasker` for Chromosome 1 of the Echidna. Repeats have been divided into bins, based on % divergence from their corresponding query. Base coverage % refers to the percent of the entire genome covered by a specific element. In these two graphs, the difference that the `Repeatmasker -norna` flag makes in SINE detection can be clearly seen.

Chapter 6: Comparison of L2s to L1s

6.1 Results

Initially, gene annotation and repeat data were sourced from the UCSC browser for analysis of the placental mammal species investigated; upon further inspection, genes did not appear to be clearly designated as protein coding, and there did not seem to be any easy way to parse out pseudogenes and non-coding RNA. Although NCBI does not have the same repeat annotation standards as UCSC across its entire platform, the human, horse and dog genomes all have curated, species-specific repeat annotations uploaded to NCBI. Thus, all analysis presented uses data either generated *de novo*, or downloaded from the NCBI genome browser.

Overlaps between L1s/L2s and exons, the 5' UTR and 3' UTR associated with protein coding genes were determined using a combination of `GIGGLE` and `bedtools` as described in the methods, and shown in Appendix B.6. The total coverage values from this analysis can be seen in Appendix D.5, while a normalised plot is shown in Figure 6.1.

6.2 Discussion

The three therian mammal species used in this analysis were human, horse and dog, chosen due to the quality of their genome assemblies, and their evolutionary distance from each other. Although a species representative from all four eutherian superorders would have been ideal, only two were used, as there there does not appear to be a high quality, chromosome level assembly for members of *Afrotheria* or *Xenarthra* with species specific repeat annotation. The inclusion of a metatherian (marsupial) would have also been preferred, but no species with both repeat annotation and gene modelling of high enough quality could be found.

Figure 6.1 shows the coverage of either L2s or L1s in exons, as a percentage of

total exon coverage. While the echidna and platypus have the lowest % LINE coverage, Supplementary Figure S7 shows that the trend of repeat coverage seems to correlate with total exon coverage (in bp). Based on this, it is difficult to draw any strong conclusions.

Supplementary figures S8 and S9 show that the two monotreme species display significantly lower 5' UTR and 3' UTR coverage, compared to the therians investigated. It is difficult to determine whether this is actually a real result, or an artefact of gene model annotation. The fact that the most well annotated genome, the human, consistently displays the highest coverage in all categories leads me to believe it is at least partially due to an artifact.

To correct for this, a number of adjustments could be made, to reduce bias caused by differences in annotation quality. First, to normalise for identified proteins, only proteins with homologues between all 5 species could be looked at. Additionally, looking at the age of these insertions (inferred by the % divergence) would give a timeline; this would require specific research into the repeat analysis methods of the placental mammals chosen however. Although this additional analysis could not be completed in the timeframe of this honours project, the high quality repeat annotations obtained for the platypus and echidna lay a solid foundation for a more extensive project.

Overall, preliminary analysis of L2 and L1 insertions into exonic, 5' UTR and 3' UTR regions of protein coding genes does not appear to show a significant difference.

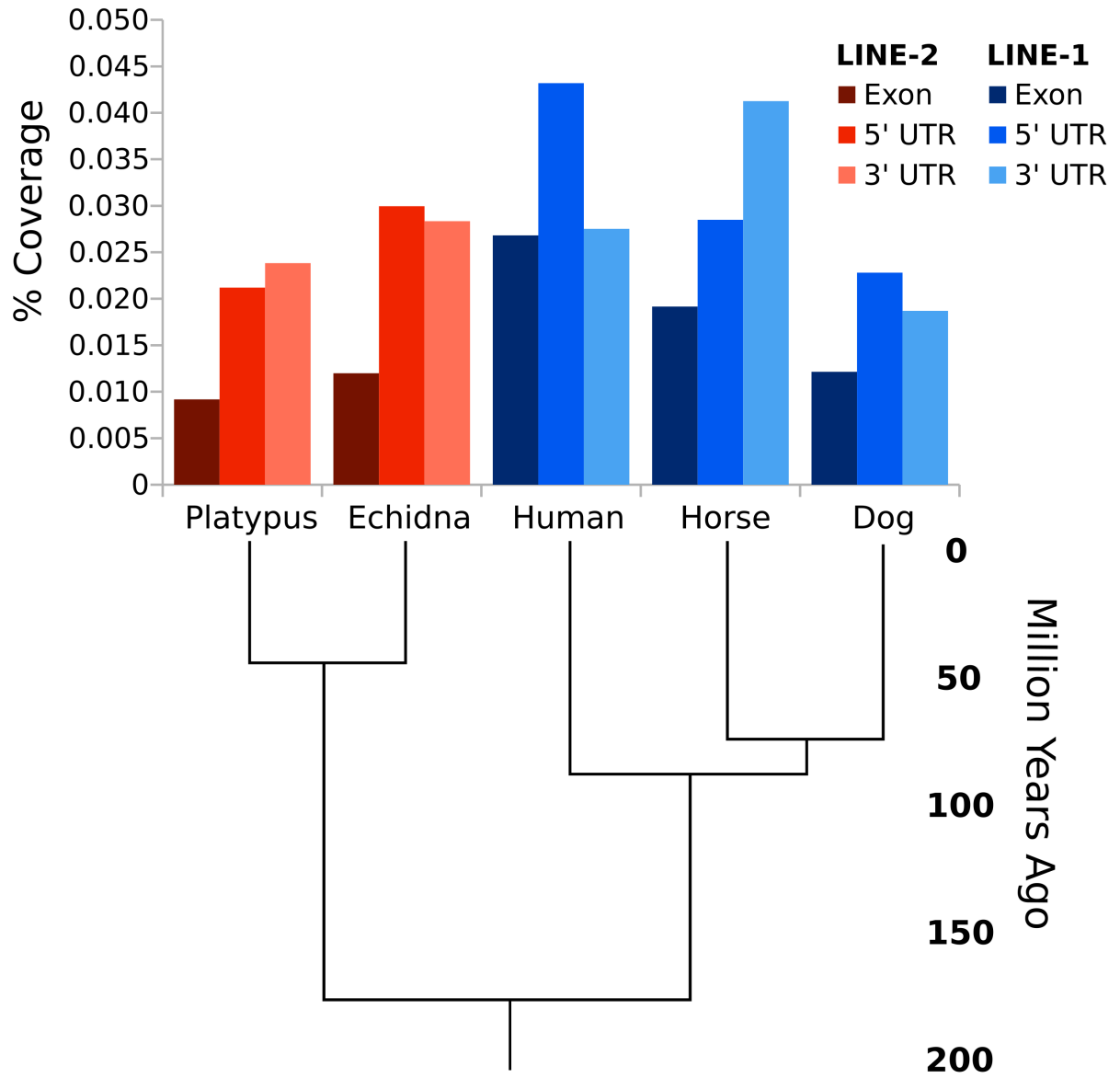


Figure 6.1: Plot showing coverage of LINE-2 or LINE-1 elements within exons, 5' untranslated regions and 3' untranslated regions, as a percentage of the total coverage of those regions. Exons and UTRs are associated with protein coding genes. Divergence times for phylogenetic tree are based on inferences by Upham et al. (2019).

Chapter 7: Conclusion

Through the completion of this project, five new subclasses of LINE-2 elements with potential activity were identified in the echidna. Additionally, a pipeline for the detection of LTR retrotransposons was developed, identifying six endogenous retroviruses with potential activity within the echidna. Seed alignments of these sequences, along with LINE-2s found in the platypus, were generated for submission to the open source transposon repository, Dfam. This is the first species-specific identification of active repeats in the echidna, and will shortly represent the first monotreme derived transposons ever submitted to Dfam.

The Comprehensive *ab initio* Repeat Pipeline was used to generate dispersed libraries of repeats in both the echidna and platypus, which were subsequenly classified using the tools `RepeatMasker` and `ins`. Outputs from these two programs were evaluated and compared, where it was determined that `ins` is the more stringent option, but that modifications to CARP would allow the use of `RepeatMasker`. With `ins` as the classifier, CARP was used to classify dispersed repeat families *ab initio*, which were in turn used to produce whole genome annotations using `RepeatMasker` for both monotreme species.

The composition and divergence of repeats within the echidna and platypus were analysed, where it was found that the echidna genome has significantly higher repeat content than the platypus overall, along with more recent L2 activity. Additionally, a high proportion of the echidna genome was identified as repetitive but unclassified, warranting further analysis.

This repeat coverage data was used to determine instances where L2s overlap with exons and untranslated regions associated with protein coding genes. This same analysis was carried out on the human, horse and dog genomes, using preexisting gene annotations. The differences in gene annotation quality made a direct comparison between the monotremes and therians investigated not possible, preventing a solid conclusion from being made. Despite this, several methods to correct for this issue have been suggested, which are now possible with the existence of high quality repeat annotations generated for the echidna and platypus.

References

Adelson, D. L., Buckley, R. M., Ivancevic, A. M., Qu, Z., and Zeng, L. (2015). Retrotransposons: Genomic and trans-genomic agents of change. In *Evolutionary Biology: Biodiversification from Genotype to Phenotype*, pages 55–75. Springer International Publishing.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):1–6.

Batzer, M. A. and Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics 2002 3:5*, 3(5):370–379.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology 2000 42:1*, 42(1):251–269.

Boeke, J. and Stoye, J. (1997). Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. *Retroviruses*.

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits.

Cosby, R. L., Judd, J., Zhang, R., Zhong, A., Garry, N., Pritham, E. J., and Feschotte, C. (2021). Recurrent evolution of vertebrate transcription factors by transposase capture. *Science*, 371(6531):eabc6405.

Daniel Kortschak, R., Bleecher Snyder, J., Maragkakis, M., and L Adelson, D. (2017). bíogo: a simple high-performance bioinformatics toolkit for the Go language. *The Journal of Open Source Software*, 2(10):167.

Deininger, P. (2011). Alu elements: Know the SINEs.

Deininger, P. L., Moran, J. V., Batzer, M. A., and Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution.

Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603.

Eickbush, T. H. and Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Research*, 134(1-2):221–234.

Eric M. Ostertag and Kazazian, H. H. (2001). Biology of Mammalian L1 Retrotransposons. *http://dx.doi.org/10.1146/annurev.genet.35.102401.091032*, 35:501–538.

Feng, Q., Moran, J. V., Kazazian, H. H., and Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, 87(5):905–916.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics 2008 9:5*, 9(5):397–405.

Filshtein, T., Sirobhushanam, S., Tiwari, K. B., and Borchert, G. M. (2012). OrbId: Origin-based identification of microRNA target s.

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, 5(C):103–107.

Gebert, D. and Rosenkranz, D. (2015). RNA-based regulation of transposon expression. *Wiley Interdisciplinary Reviews: RNA*, 6(6):687–708.

Gemmell, N. J., Rutherford, K., Prost, S., Tollis, M., Winter, D., Macey, J. R., Adelson, D. L., Suh, A., Bertozzi, T., Grau, J. H., Organ, C., Gardner, P. P., Muffato, M., Patricio, M., Billis, K., Martin, F. J., Flicek, P., Petersen, B., Kang, L., Michalak, P., Buckley, T. R., Wilson, M., Cheng, Y., Miller, H., Schott, R. K., Jordan, M. D., Newcomb, R. D., Arroyo, J. I., Valenzuela, N., Hore, T. A., Renart, J., Peona, V., Peart, C. R., Warmuth, V. M., Zeng, L., Kortschak, R. D., Raison, J. M., Zapata, V. V., Wu, Z., Santesmasses, D., Mariotti, M., Guigó, R., Rupp, S. M., Twort, V. G., Dussex, N., Taylor, H., Abe, H., Bond, D. M., Paterson, J. M., Mulcahy, D. G., Gonzalez, V. L., Barbieri, C. G., DeMeo, D. P., Pabinger, S., Van Stijn, T., Clarke, S., Ryder, O., Edwards, S. V., Salzberg, S. L., Anderson, L., Nelson, N., Stone, C., Stone, C., Smillie, J., and Edmonds, H. (2020). The tuatara genome reveals ancient features of amniote evolution. *Nature*, 584(7821):403–409.

Grossnickle, D. M., Smith, S. M., and Wilson, G. P. (2019). Untangling the Multiple Ecological Radiations of Early Mammals.

Grützner, F., Rens, W., Tsend-Ayush, E., El-Mogharbel, N., O'Brien, P. C., Jones, R. C., Ferguson-Smith, M. A., and Marshall Graves, J. A. (2004). In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature*, 432(7019):913–917.

Havecker, E. R., Gao, X., and Voytas, D. F. (2004). The diversity of LTR retrotransposons. *Genome Biology 2004 5:6*, 5(6):1–6.

Hua-Van, A., Le Rouzic, A., Maisonhaute, C., and Capy, P. (2005). Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenetic and Genome Research*, 110(1-4):426–440.

Huang, C. R. L., Burns, K. H., and Boeke, J. D. (2012). Active Transposition in Genomes. *Annual Review of Genetics*, 46(1):651–675.

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., and Adelson, D. L. (2016). LINEs between species: Evolutionary dynamics of LINE-1 retrotransposons across the eukaryotic tree of life. *Genome Biology and Evolution*, 8(11):3301–3322.

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):85.

Ivancevic, A. M., Walsh, A. M., Kortschak, R. D., and Adelson, D. L. (2013). Jumping the fine LINE between species: Horizontal transfer of transposable elements in animals catalyses genome evolution. *BioEssays*, 35(12):1071–1082.

Jangam, D., Feschotte, C., and Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts.

Joly-Lopez, Z. and Bureau, T. E. (2018). Exaptation of transposable element coding sequences.

Kramerov, D. A. and Vassetzky, N. S. (2011). Origin and evolution of SINEs in eukaryotic genomes. *Heredity 2011 107:6*, 107(6):487–495.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M. L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp,

D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Layer, R. M., Pedersen, B. S., DiSera, T., Marth, G. T., Gertz, J., and Quinlan, A. R. (2018). GIGGLE: a search engine for large-scale integrated genome analysis. *Nature methods*, 15(2):123.

Lovšin, N., Gubenšek, F., and Kordi, D. (2001). Evolutionary Dynamics in a Novel L2 Clade of Non-LTR Retrotransposons in Deuterostomia. *Molecular Biology and Evolution*, 18(12):2213–2224.

Luo, Z. X. (2007). Transformation and diversification in early mammal evolution.

Malik, H. S., Burke, W. D., and Eickbush, T. H. (1999). The Age and Evolution of Non-LTR Retrotransposable Elements. *Mol. Biol. Evol*, 16(6):793–805.

Musser, A. M. (2003). Review of the monotreme fossil record and comparison of palaeontological and molecular data. *Comparative Biochemistry and Physiology Part A: Molecular and Integrative Physiology*, 136(4):927–942.

Ogino, S., Nosho, K., Kirkner, G. J., Kawasaki, T., Chan, A. T., Schernhammer, E. S., Giovannucci, E. L., and Fuchs, C. S. (2008). A Cohort Study of Tumoral LINE-1 Hypomethylation and Prognosis in Colon Cancer. *JNCI: Journal of the National Cancer Institute*, 100(23):1734–1738.

Okada, N. (1991). SINEs: Short interspersed repeated elements of the eukaryotic genome. *Trends in Ecology and Evolution*, 6(11):358–361.

Okada, N., Hamada, M., Ogiwara, I., and Ohshima, K. (1997). SINEs and LINEs share common 3' sequences: a review. *Gene*, 205(1-2):229–243.

Payer, L. M. and Burns, K. H. (2019). Transposable elements in human genetic disease.

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H.-Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C.-S., Guo, Y., Paxinos, E. E., Korbel, J. O., Darnell, R. B., McCombie, W. R., Kwok, P.-Y., Mason, C. E., Schadt, E. E., and Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods 2015 12:8*, 12(8):780–786.

Petri, R., Brattås, P. L., Sharma, Y., Jonsson, M. E., Pircs, K., Bengzon, J., and Jakobsson, J. (2019). LINE-2 transposable elements are a source of functional human microRNAs and target sites. *PLoS Genetics*, 15(3).

Piriyapongsa, J., Mariño-Ramírez, L., and Jordan, I. K. (2007). Origin and evolution of human microRNAs from transposable elements. *Genetics*, 176(2):1323–1337.

Piskurek, O. and Jackson, D. J. (2012). Transposable Elements: From DNA Parasites to Architects of Metazoan Evolution. *Genes*, 3(3):409–422.

Platt, R. N., Vandewege, M. W., and Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research*, 26(1-2):25–43.

Saleh, A., Macia, A., and Muotri, A. R. (2019). Transposable Elements, Inflammation, and Neurological Disease. *Frontiers in Neurology*, 0(AUG):894.

SanMiguel, P., Tikhonov, A., Bennetzen, J. L., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., EDWARDS, K. J., LEE, M., and AVRAMOVA, Z. (1996). Nested Retrotransposons in the Intergenic Regions of the Maize Genome. *Science (American Association for the Advancement of Science)*, 274(5288):765–768.

Shen, S., Lin, L., Cai, J. J., Jiang, P., Kenkel, E. J., Stroik, M. R., Sato, S., Davidson, B. L., and Xing, Y. (2011). Widespread establishment and regulatory impact of Alu exons in human genes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7):2837–2842.

Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., Foster, J. W., Frischauf, A.-M., Lovell-Badge, R., and Goodfellow, P. N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature 1990 346:6281*, 346(6281):240–244.

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes.

Spengler, R. M., Oakley, C. K., and Davidson, B. L. (2014). Functional microRNAs and target sites are created by lineage-specific transposition. *Human Molecular Genetics*, 23(7):1783–1793.

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 2021 12:1*, 12(1):1–14.

Suh, A., Churakov, G., Ramakodi, M. P., Platt, R. N., Jurka, J., Kojima, K. K., Caballero, J., Smit, A. F., Vliet, K. A., Hoffmann, F. G., Brosius, J., Green, R. E., Braun, E. L., Ray, D. A., and Schmitz, J. (2015). Multiple Lineages of Ancient CR1 Retroposons Shaped the Early Genome Evolution of Amniotes. *Genome Biology and Evolution*, 7(1):205–217.

Upham, N. S., Esselstyn, J. A., and Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biology*, 17(12):e3000494.

Walsh, A. M., Kortschak, R. D., Gardner, M. G., Bertozzi, T., and Adelson, D. L. (2013). Widespread horizontal transfer of retrotransposons. *Proceedings of the National Academy of Sciences of the United States of America*, 110(3):1012–1016.

Warren, W. C., Hillier, L. D. W., Marshall Graves, J. A., Birney, E., Ponting, C. P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A. T., Yang, S. P., Heger, A., Locke, D. P., Miethke, P., Waters, P. D., Veyrunes, F., Fulton, L., Fulton, B., Graves, T., Wallis, J., Puente, X. S., López-Otín, C., Ordóñez, G. R., Eichler, E. E., Chen, L., Cheng, Z., Deakin, J. E., Alsop, A., Thompson, K., Kirby, P., Papenfuss, A. T., Wakefield, M. J., Olender, T., Lancet, D., Huttley, G. A., Smit, A. F., Pask, A., Temple-Smith, P., Batzer, M. A., Walker, J. A., Konkel,

M. K., Harris, R. S., Whittington, C. M., Wong, E. S., Gemmell, N. J., Buschiazzo, E., Vargas Jentzsch, I. M., Merkel, A., Schmitz, J., Zemann, A., Churakov, G., Ole Kriegs, J., Brosius, J., Murchison, E. P., Sachidanandam, R., Smith, C., Hannon, G. J., Tsend-Ayush, E., McMillan, D., Attenborough, R., Rens, W., Ferguson-Smith, M., Lefèvre, C. M., Sharp, J. A., Nicholas, K. R., Ray, D. A., Kube, M., Reinhardt, R., Pringle, T. H., Taylor, J., Jones, R. C., Nixon, B., Dacheux, J. L., Niwa, H., Sekita, Y., Huang, X., Stark, A., Kheradpour, P., Kellis, M., Flicek, P., Chen, Y., Webber, C., Hardison, R., Nelson, J., Hallsworth-Pepin, K., Delehaunty, K., Markovic, C., Minx, P., Feng, Y., Kremitzki, C., Mitreva, M., Glasscock, J., Wylie, T., Wohldmann, P., Thiru, P., Nhan, M. N., Pohl, C. S., Smith, S. M., Hou, S., Renfree, M. B., Mardis, E. R., and Wilson, R. K. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175–183.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics 2007 8:12*, 8(12):973–982.

Yoder, J. A., Walsh, C. P., and Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340.

Zeng, L., Kortschak, R. D., Raison, J. M., Bertozzi, T., and Adelson, D. L. (2018). Superior ab initio identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. *PLOS ONE*, 13(3):e0193588.

Zhou, Y., Shearwin-Whyatt, L., Li, J., Song, Z., Hayakawa, T., Stevens, D., Fenelon, J. C., Peel, E., Cheng, Y., Pajpach, F., Bradley, N., Suzuki, H., Nikaido, M., Damas, J., Daish, T., Perry, T., Zhu, Z., Geng, Y., Rhie, A., Sims, Y., Wood, J., Haase, B., Mountcastle, J., Fedrigo, O., Li, Q., Yang, H., Wang, J., Johnston, S. D., Phillippy, A. M., Howe, K., Jarvis, E. D., Ryder, O. A., Kaessmann, H., Donnelly, P., Korlach, J., Lewin, H. A., Graves, J., Belov, K., Renfree, M. B., Grutzner, F., Zhou, Q., and Zhang, G. (2021). Platypus and echidna genomes reveal mammalian biology and evolution. *Nature*, pages 1–7.

Appendices

A Description of programs used

Table S1: Names of programs used, along with version, accession and usage descriptions

Program (version)	Description of Usage & URL
AliView (1.27)	Nucleotide/protein alignment viewer and editor. `https://github.com/AliView/AliView`
bedtools (2.30.0)	Used to perform `intersect`, `coverage`, `genomecov`, `sort`, `merge`, `cluster`, `getfasta` and `overlap` operations on BED, GTF and GFF files. `https://github.com/arq5x/bedtools2`
Blast+ (2.10.1)	Local alignment search tool, to find regions of similarity between sequences, through `blastn` and blastx. `https://github.com/ncbi/blast_plus_docs`
bundle (b0da10f)	Split multiple MFA sequence files into a collection of MFA files which are smaller than a specified bundle size. Author: Daniel Kortschak. `https://github.com/biogo/examples/blob/master/bundle/`
cd-hit (4.8.1)	Used for clustering non-redundant nucleotide sequences at a specific % identity. `https://github.com/weizhongli/cdhit`
Exonerate (2.4.0)	Selection of utilities for performing simple manipulations on fasta files, including `fastalength`, `fastafetch`, `fastaoverlap` and `fastasort`. `https://github.com/nathanweeks/exonerate`
fasplit (3.85)	Used to split fasta files based on length and sequence name. `hgdownload.cse.ucsc.edu/admin/exe/`
Gepard (2.1.0)	GUI compatible application, for the creation of dotplots. Used for self-alignments. `https://github.com/univieCUBE/gepard`
gffer (b7a3754)	Converts the JSON output of `igor` to GFF. Author: Dan Kortschak. `https://github.com/biogo/examples/blob/master/igor/gffer/`

Continued on next page...

Program (version)	Description of Usage & URL
GIGGLE (v0.6.3)	Genomic search engine used to identify shared genomic loci between query features and genome interval files. Author: Ryan Layer. `https://github.com/ryanlayer/giggle`
hmmbuild (3.3.2)	Build hmm profile from multiple sequence alignment. `https://github.com/EddyRivasLab/hmmer`
hmmemit (3.3.2)	Used to generate a consensus sequence from a hmm profile. `https://github.com/EddyRivasLab/hmmer`
hmmfetch (3.3.2)	Extract profile(s) from a profile file. `https://github.com/EddyRivasLab/hmmer`
hmmscan (3.3.2)	Search for hmm profile from within a profile database. `https://github.com/EddyRivasLab/hmmer`
hmmsearch (3.3.2)	Used to search hmm profile(s) against a sequence database. `https://github.com/EddyRivasLab/hmmer`
igor (fde48e)	Clusters result coordinates output by `krishna` through single-linkage clustering. Author: Dan Kortschak. `https://github.com/biogo/examples/tree/master/igor`
ins (6eceafc)	Repeat identification/annotation tool. Uses BLAST+ to find instances of repeats (based on a repeat library) within a query. In this project, ins was used to classify dispersed repeat families detected *ab initio* through CARP. Author: Daniel Kortschak. `https://github.com/kortschak/ins`
IQ-TREE (2.1.4)	Stochastic algorithm to infer phylogenetic trees by maximum likelihood. `https://github.com/Cibiv/IQ-TREE`
krishna (88c7c8f)	Performs all to all alignment of fasta files, and reports coordinates matching a minimum length and percentage identity. Written by Dan Kortschak. `https://github.com/biogo/examples/tree/master/igor`
LTRharvest (1.2.1)	Software to detect potential LTR retrotransposon sequences *de novo*. `https://github.com/genometools/genometools`
LTRretriever (v2.9.0)	Used for identification and classification of potential LTR retrotransposons, generated by LTRharvest. `https://github.com/oushujun/LTR_retriever`

Continued on next page...

Program (version)	Description of Usage & URL
MAFFT (v7.453)	Program used to perform multiple sequence alignments of amino acid and protein sequences. `https://github.com/GSLBiotech/mafft`
MrBayes (3.2.7)	Program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models. `https://github.com/NBISweden/MrBayes`
ORFfinder	Detect open reading frames from a queried DNA sequence. `ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/`
RepeatMasker (4.0.7)	Detect and annotate interspersed repeats within a DNA sequence, based on homology to a queried repeat library. Used for the classification of dispersed repeat families detected *ab initio* through CARP, and to perform whole genome annotations with species-specific libraries. `https://github.com/rmhubley/RepeatMasker`
seqer (b7a3754)	Return multiple fasta sequences corresponding to feature intervals. Author: Dan Kortschak. `https://github.com/biogo/examples/tree/master/igor/seqer`
seqsplit (a5b6f8c)	Split contig sequences above a minimum cut-ff length to generate fragments. `https://github.com/biogo/examples/blob/master/seqsplit/`
seqkit (c03b347)	Used to remove fasta sequences from a MFA file under a certain length. `https://github.com/shenwei356/seqkit`
Suffixerator (1.2.1)	Used to compute enhanced suffix array, for input to LTRharvest. `https://github.com/genometools/genometools`

B Programs/Scripts Written

B.1 `dfamgenerator.sh`

```
#!/usr/bin/env
# Used to generate HMM models for repetitive elements previously
# identified within a genome. Only tested on L2s and ERVs, alterations
# may be needed for other transposon types.


REPEAT=name

# Find query matches in the genome
blastn -query $REPEAT\_rep.fa -db ~/Taculeatus_newL2s/LTRstuff/LTRanalysis/potentiallyactive/blastdb
    -outfmt "7 length sacc sstart send" -task blastn | sort -nr > $REPEAT\_allblast.txt

# Change cutoff length for blast results
awk '$1>1000 {print}' $REPEAT\_allblast.txt > $REPEAT\_allblast_over1k.txt

# Specific to Echidna
sed -E 's/(scaffold_\d*_)/\1arrow_ctg/g' $REPEAT\_allblast\_over1k.txt > $REPEAT\_over1k_bed
sed -Ei 's/^\S*\t//g' $REPEAT.bed

# Orient bed file
awk '{if ($2 > $3) print $1"\t"$3"\t"$2; else print $1"\t"$2"\t"$3}' $REPEAT.bed > $REPEAT\_stranded

# Extract from genome
bedtools getfasta -fi ~/Taculeatus_ERVs/original/mTacAcu1.pri.cur.20201026.fasta -bed $REPEAT\_stran

# Align
mafft --adjustdirection --maxiterate 300 $REPEAT.fa > $REPEAT.ali

# Build hmm profile
hmmbuild -o $REPEAT\_hmmbuild.log -O $REPEAT.stk -n Tacu_$REPEAT $REPEAT.hmm $REPEAT\_blastn.ali

# Extract consensus from profile
hmmemit -c ERV3a.hmm
```

B.2 L2_CR1_extractor.sh

```
#!/usr/bin/env
## Extracts the RVT domain from L2 and CR1 sequences
## Requires the RVT_1 .hmm profile, available at [http://pfam.xfam.org/family/pf00078]

SPECIES="Lchalumnae"

# Extract L2s and CR1s and place into combined file
awk -v search="L2" -f ../fasta_search.awk ../originals/"$SPECIES".fa > L2"$SPECIES".fa
awk -v search="CR1" -f ../fasta_search.awk ../originals/"$SPECIES".fa > CR1"$SPECIES".fa
cat CR1"$SPECIES".fa L2"$SPECIES".fa > "$SPECIES"CR1L2.fa

# Extract ORFs over 1500bp as both aa and nt
ORFfinder -in "$SPECIES"CR1L2.fa -out ORF"$SPECIES"CR1L2.fa -ml 1500 -s 2
ORFfinder -in "$SPECIES"CR1L2.fa -out ORF"$SPECIES"CR1L2_nucl.fa -ml 1500 -s 2 -outfmt 1

# hmmsearch for the RVT domain in aa ORF file
hmmsearch --domtblout ORF"$SPECIES"CR1L2.domtblout ../RVT.hmm ORF"$SPECIES"CR1L2.fa

# Extract coordinates from hmm domtblout file
awk '($1!~"#"){print $1, $18, $19 }' ORF"$SPECIES"CR1L2.domtblout >
    allRVT"$SPECIES"CR1L2_prot.bed
# Assume coordinates for aa are 3x that for nt
awk '($1!~"#"){print $1, $2 * 3, $3 * 3 }' allRVT"$SPECIES"CR1L2_prot.bed >
    allRVT"$SPECIES"CR1L2_nucl.bed

# Format tabs and sequence names to match for bedtools
sed -i 's/ /\t/g' allRVT"$SPECIES"CR1L2_prot.bed
sed -i 's/ /\t/g' allRVT"$SPECIES"CR1L2_nucl.bed
sed -Ei 's/:.*:....//g' allRVT"$SPECIES"CR1L2_nucl.bed
sed -Ei 's/:.*:....//g' ORF"$SPECIES"CR1L2_nucl.fa
sed -i 's/|ORF1_/|/g' allRVT"$SPECIES"CR1L2_nucl.bed

# Extract RVT matching sequence from ****CR1L2 file
bedtools getfasta -fi ORF"$SPECIES"CR1L2.fa -bed allRVT"$SPECIES"CR1L2_prot.bed -fo
    allRVT"$SPECIES"CR1L2_prot.fa.out
bedtools getfasta -fi ORF"$SPECIES"CR1L2_nucl.fa -bed allRVT"$SPECIES"CR1L2_nucl.bed -fo
    allRVT"$SPECIES"CR1L2_nucl.fa.out
```

```
# Remove sequences below 200 aa and 600 bp respectively
awk 'BEGIN {RS = ">" ; ORS = ""} length($2) >= 200 {print ">"$0}'
    allRVT"$SPECIES"CR1L2_prot.fa.out > RVT"$SPECIES"CR1L2_prot.fa.out
awk 'BEGIN {RS = ">" ; ORS = ""} length($2) >= 600 {print ">"$0}'
    allRVT"$SPECIES"CR1L2_nucl.fa.out > RVT"$SPECIES"CR1L2_nucl.fa.out
```

B.3 divergence.R

```
## Parse *.align.landscape.Div.Rname.tab files generated by parseRM.pl
## [https://github.com/4ureliek/Parsing-RepeatMasker-Outputs], and generate
## graphs from the command line
##
## Usage: Rscript divergence.R "folderpath" "filename" "figuretitle"
library(tidyverse)
library(dplyr)
library(purrr)
library(svglite)

args <- commandArgs(trailingOnly = TRUE)
folderpath <- args[1]
filename <- args[2]
figuretitle <- args[3]

div_table_spaces <- paste("~/L2_L1_analysis/", args[1], "/", args[2],
 ".fa.align.landscape.Div.Rname.tab")
div_summary_spaces <- paste("~/L2_L1_analysis/", args[1], "/", args[2],
 ".fa.align.parseRM.summary.tab")

div_summary <- gsub(" ", "", div_summary_spaces, fixed = TRUE)
div_table <- gsub(" ", "", div_table_spaces, fixed = TRUE)
print(div_table)

chr_size_table <- read.table(file = div_summary, header = FALSE)
chr_size <- chr_size_table[1, "V1"]
chr_size <- as.numeric(chr_size)

# Add headers
```

```
repNoFamilyID <- read.table(file = div_table, header = FALSE, skip = 3,
col.names = c("name", "remove", "class", "1", "2", "3", "4", "5", "6",
"7", "8", "9", "10", "11", "12", "13", "14", "15", "16", "17","18",
"19", "20", "21", "22", "23", "24", "25", "26", "27", "28", "29", "30",
"31", "32", "33", "34", "35", "36", "37", "38", "39", "40", "41", "42",
 "43", "44", "45", "46", "47", "48", "49", "50"))

valuesNormalised = repNoFamilyID[, c(4:53)]
Name <- repNoFamilyID$name
Class <- repNoFamilyID$class
valuesNormalised  <- valuesNormalised /chr_size

## Normalised Analysis ##
# Separate into class and name
repClassNormalised <- bind_cols(... = Class, valuesNormalised)
colnames(repClassNormalised)[1] <- "class"

repNameNormalised <- bind_cols(... = Name, valuesNormalised)
colnames(repNameNormalised)[1] <- "name"

# Sum variables with the same name
summaryClassNormalised <- aggregate(.~class, data = repClassNormalised, FUN = sum)
summaryNameNormalised <- aggregate(.~name, data = repNameNormalised, FUN = sum)

# Merge subclasses into classes (LINE-2s are kept separate in this case)
summaryClassNormalisedReduced <- data.frame((lapply(summaryClassNormalised,
 function(x) {
  gsub("Unclassified", "Unclassified", x)
    })))
summaryClassNormalisedReduced <- data.frame((lapply(summaryClassNormalisedReduced,
function(x) {
  gsub("Penelope|R4|L1|Non-LTR|Nimb|Tx1|Vingi|CR1|RTE|RTEX", "other LINEs", x)
})))
summaryClassNormalisedReduced <- data.frame((lapply(summaryClassNormalisedReduced,
function(x) {
  gsub("DIRS|Ginger1|Gypsy|LTR-775_Gav_odd|Endogenous|ERV|ERV1|ERV2|ERV3|
  ERV4|Retrovirus_like|LTR|Rex1", "LTR/ERV", x)
})))
summaryClassNormalisedReduced <- data.frame((lapply(summaryClassNormalisedReduced,
function(x) {
```

```
  gsub("CACTA|Crypton|CryptonV|Dada|DNA|Harbinger|hAT|Helitron|IS3EU|Kolobok|
  MuDR|piggyBac|Polinton|Tc1|Transposable|Zisupton", "DNA Transposon", x)
}))
summaryClassNormalisedReduced <- data.frame((lapply(summaryClassNormalisedReduced,
function(x) {
  gsub("5S|BEL|conserved|MARE8|MSRBMI|Nonautonomous|REP-24_CPB|Repetitive|Simple|
  Unspecified", "Ancient", x)
}))
summaryClassNormalisedReduced <- data.frame((lapply(summaryClassNormalisedReduced,
function(x) {
  gsub("SINE|SINE3", "SINE", x)
}))
summaryClassNormalisedReduced <- data.frame((lapply(summaryClassNormalisedReduced,
function(x) {
  gsub("Multicopy", "Multicopy Gene", x)
}))
summaryClassNormalisedReduced <- data.frame((lapply(summaryClassNormalisedReduced,
function(x) {
  gsub("SAT|Satellite|satellite", "Satellite", x)
}))

# Change to longform
summaryClassNormalised_long <- summaryClassNormalisedReduced %>%
  gather(X, value, X1:X50)

# Remove "X" from column
summaryClassNormalised_long <- data.frame((lapply(summaryClassNormalised_long,
function(x) {gsub("^X", "", x)
}))

# Make numbers numeric
summaryClassNormalised_long$value <- as.numeric(summaryClassNormalised_long$value)
summaryClassNormalised_long$value <- as.numeric(summaryClassNormalised_long$value*100)
summaryClassNormalised_long$X <- as.numeric(summaryClassNormalised_long$X)

summaryClassNormalisedTrimmed_long <- summaryClassNormalised_long

graph <- ggplot(data=summaryClassNormalisedTrimmed_long, aes(x = X,
                y = value, fill = class, ordered = TRUE)) +
  geom_bar(stat = "identity", width = 0.6) +
```

```
scale_y_continuous("Base Coverage %", expand=expansion(mult=c(0,0.05)),
    limits=c(0, 3.1)) +
scale_x_continuous("% Divergence", expand=expansion(mult=c(0,0))) +
ggtitle(figuretitle) +
theme_light() +
theme(plot.title = element_text(hjust = 0.5)) +
scale_fill_manual(values=c("Unclassified" = "#FFEE86", "Ancient" = "#94CC4D",
                  "DNA Transposon" = "#d1495b", "L2" = "#0951B0", "LTR/ERV" = "#edae49",
                  "Multicopy Gene" = "grey", "other LINEs" = "brown", "RTE" = "navy",
                  "Satellite" = "purple2", "SINE" = "black"))

ggsave(path = "/home/alex/L2_L1_analysis/AllRepeats_labelled_Tacu",
               file = figuretitle, device = svg, plot=graph, width=10, height=5)
```

B.4 L2_subfamily_divergence.R

```
## Parse *.align.landscape.Div.Rname.tab files generated by parseRM.pl
## [https://github.com/4ureliek/Parsing-RepeatMasker-Outputs], into LINE-2
## subfamilies, and generate graphs from the command line
##
## Usage: Rscript L2_subfamily_divergence.R "folderpath" "filename" "figuretitle"
library(tidyverse)
library(dplyr)
library(purrr)
library(svglite)


# Setup args
args <- commandArgs(trailingOnly = TRUE)
folderpath <- args[1]
filename <- args[2]
figuretitle <- args[3]


#change for echidna and platypus
div_table_spaces <- paste("~/L2_L1_analysis/", args[1], "/", args[2],
```

```
".fa.align.landscape.Div.Rname.tab")
div_summary_spaces <- paste("~/L2_L1_analysis/", args[1], "/", args[2],
".fa.align.parseRM.summary.tab")

div_summary <- gsub(" ", "", div_summary_spaces, fixed = TRUE)
div_table <- gsub(" ", "", div_table_spaces, fixed = TRUE)
chr_size_table <- read.table(file = div_summary, header = FALSE)
chr_size <- chr_size_table[1, "V1"]
chr_size <- as.numeric(chr_size)

# Add headers
repNoFamilyID <- read.table(file = div_table, header = FALSE, skip = 3,
col.names = c("name", "remove", "class", "1", "2", "3", "4", "5", "6",
    "7", "8", "9", "10", "11", "12", "13", "14", "15", "16", "17","18",
    "19", "20", "21", "22", "23", "24", "25", "26", "27", "28", "29", "30",
    "31", "32", "33", "34", "35", "36", "37", "38", "39", "40", "41", "42",
    "43", "44", "45", "46", "47", "48", "49", "50"))

valuesNormalised = repNoFamilyID[, c(4:53)]
Name <- repNoFamilyID$name
Class <- repNoFamilyID$class
valuesNormalised  <- valuesNormalised /chr_size

## Normalised Analysis ##
# Separate into class and name
repClassNormalised <- bind_cols(... = Class, valuesNormalised)
colnames(repClassNormalised)[1] <- "class"

repNameNormalised <- bind_cols(... = Name, valuesNormalised)
colnames(repNameNormalised)[1] <- "name"

# Sum variables with the same name
summaryClassNormalised <- aggregate(.~class, data = repClassNormalised, FUN = sum)
summaryNameNormalised <- aggregate(.~name, data = repNameNormalised, FUN = sum)

# Extract L2 subfamilies
L2Families <- summaryNameNormalised %>%
    filter_all(any_vars(.=="L2_Plat1a" | .=="L2_Plat1b" | .=="L2_Plat1c" | .=="L2_Plat1d"
                    | .=="L2_Plat1e" | .=="L2_Plat1f" | .=="L2_Plat1g" | .=="L2_Plat1h"
```

```
                    | .=="L2_Plat1i" | .=="L2_Plat1n" | .=="L2_Plat1o" | .=="L2_Plat1q"
                    | .=="L2_Plat1r" | .=="L2_Plat1s" | .=="L2_Plat1t" | .=="L2_Plat1u"
                    | .=="Tacu_L2a" | .=="Tacu_L2b" | .=="Tacu_L2c" | .=="Tacu_L2d"
                    | .=="Tacu_L2e"))

# Change to longform
L2Families_long <- L2Families %>%
    gather(X, value, X1:X50)

# Remove "X" from column
L2Families_long <- data.frame((lapply(L2Families_long, function(x) {
    gsub("^X", "", x)
    })))

# Make numbers numeric
L2Families_long$value <- as.numeric(L2Families_long$value)
L2Families_long$X <- as.numeric(L2Families_long$X)

graph <- ggplot(data=L2Families_long, aes(x = X, y = value, fill = name, ordered = TRUE)) +
    geom_bar(stat = "identity", width = 0.6) +
    scale_y_continuous("Base Coverage %", expand=expansion(mult=c(0,0.05)),
                       limits=c(0, 0.03)
    ) +
    scale_x_continuous("% Divergence", expand=expansion(mult=c(0,0))) +
    ggtitle(figuretitle) +
    scale_fill_manual(values=c('#004586', '#FF420E', '#FFD320', '#579D1C', '#7E0021',
                    '#83CAFF','#314004', '#AECF00', '#4B1F6F', '#FF950E', '#C5000B',
                    '#0084D1', '#008080', '#e6beff', '#9a6324', 'darkblue', '#D41500',
                    '#00B355', 'slateblue', '#000075', '#000000')) +
  theme_light() +
  theme(plot.title = element_text(hjust = 0.5))

graph <- graph + guides(fill=guide_legend(title="LINE-2 Subfamilies"))

ggsave(path = "/home/alex/L2_L1_analysis/Tacu_L2_Divergence",
       file = figuretitle, device = svg, plot=graph, width=10, height=5)
```

B.5 GIGGLE indexing

```
giggle index -i "RMSK/*.gz" -o new_Ecaballus_WholeGenome_index/ -f -s
```

B.6 `Overlap_Extractor.sh`

```
#!/usr/bin/env
# Script for finding "$REGION" overlaps with a repeat class
# "$REGION" is exon, 5UTR or 3UTR
#
# Requires giggle index to already exist, as "$SPECIES"_WholeGenome_Index/
#
# Requires directory structure:
# "$SPECIES"
# |
# |-- gtf_files
# |    |
# |    |-- "$SPECIES".gff
# |-- exon
# |-- 5UTR
# |-- 3UTR
# |-- "$SPECIES"_WholeGenome_Index
# |-- "$SPECIES"_genome.txt

SPECIES=Hsapiens
REGION=exon

# Convert gff file to .bed
awk -F'\t|;' '{print $1"\t"$4"\t"$5"\t"$13}' "$SPECIES"_"$REGION"_coverage.gff >
        "$SPECIES"_"$REGION"_coverage.bed

# Sort, bgzip and index bed file
bedtools sort -i "$SPECIES"_"$REGION"_coverage.bed | bedtools merge >
        "$SPECIES"_"$REGION"_coverage_SM.bed
bash ~/giggle/scripts/sort_bed ""$SPECIES"_"$REGION"_coverage.bed" ./
cd ../
giggle search -i "$SPECIES"_WholeGenome_index/ -q
```

```
    "$REGION"/"$SPECIES"_"$REGION"_coverage.bed.gz -v -o >
    "$REGION"/"$SPECIES"_"$REGION"_repeat_overlap.txt
cd -

# Find L1 overlaps, extract intersections with "$REGION"
grep "LINE/L1" "$SPECIES"_"$REGION"_repeat_overlap.txt > "$SPECIES"_"$REGION"_L1_overlap.txt
bedtools sort -i "$SPECIES"_"$REGION"_L1_overlap.txt |
        bedtools merge > "$SPECIES"_"$REGION"_L1_overlap_SM.txt
bedtools intersect -a "$SPECIES"_"$REGION"_L1_overlap_SM.txt -b
        "$SPECIES"_"$REGION"_coverage_SM.bed >
    "$SPECIES"_"$REGION"_L1_overlap_SM_intersect.bed

# Only include chromosomal regions
grep "NC" "$SPECIES"_"$REGION"_coverage_SM.bed >
    "$SPECIES"_"$REGION"_coverage_SM_chromosomes.bed
grep "NC" "$SPECIES"_"$REGION"_L1_overlap_SM_intersect.bed >
    "$SPECIES"_"$REGION"_L1_overlap_SM_intersect_chromosomes.bed

# Determine coverage as a % of the genome
bedtools genomecov -i "$SPECIES"_"$REGION"_L1_overlap_SM_intersect_chromosomes.bed
        -g ../"$SPECIES"_genome.txt > "$SPECIES"_"$REGION"_L1_coverage.txt
bedtools genomecov -i "$SPECIES"_"$REGION"_coverage_SM_chromosomes.bed
        -g ../"$SPECIES"_genome.txt > "$SPECIES"_"$REGION"_coverage.txt
```

B.7 Repeat annotation tools

Typical RepeatMasker usage

```
RepeatMasker -pa 4 -a -nolow -norna -dir ./ -lib combined_library.fa ConsensusSequences.fa
```

```
-pa
Number of processors to use in parallel

-a
Produce -align file (needed for ParseRM.pl)

-nolow
Removes simple repeats from the output

-norna
Does not mask small RNA genomes
```

Typical Ins usage

```
ins -lib ../combined_library.fa -query
    ConsensusSequences.fa >ConsensusSequences.gtf 2>ConsensusSequencesIns.log
```

C Programs/Scripts Used

C.1 Software used in LTR discovery pipeline

```
# Suffixerator is used to  generate an enhanced suffix array from a multi-fasta file
gt suffixerator \
    -db Taculeatus_genome.fa \
    -indexname genome.fa \
     -tis -suf -lcp -des -ssp -sds -dna

#  LTRharvest predicts LTR retrotransposons
```

```
gt ltrharvest \
    -index Taculeatus_genome.fa \
    -similar 90 -vic 10 -seed 20 -seqids yes \
    -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 \
    -motif TGCA -motifmis 1 > Taculeatus_genome.harvest.scn

# LTRretriever identifies non-redundant LTR retrotransposons from
# the output of LTRharvest
LTR_retriever -genome Taculeatus_genome.fa -inharvest Taculeatus_genome.harvest.scn

# seqkit to remove sequences under 2kb
seqkit seq -m 2000 Taculeatus_genome.LTRlib.fa > Tacu_LTRs_over2k.fa

# RepeatMasker search to identify non-LTR retrotransposons
RepeatMasker -pa 1 -a -nolow -norna  -dir ./ -lib FormattedVertebrata_TacuL2s.fa
    Tacu_LTRs_over2k.fa

# Manually remove sequences mapping to non-LTR retrotransposons

# Search for protein domains with hmmsearch
hmmsearch --domtblout Taculeatus_LTRs.domtblout
    ~/databases/Pfam-A.hmm Tacu_LTRs_over2k_nonLTRsremoved.fa

# Manually identify domains as gag, pol or env associated, extract associated sequences
# Cluster sequences at 90% identity for subsequent alignment
cd-hit -i Tacu_LTRs_potentiallyactive.fa -o Tacu_LTRs_potentiallyactive_090.fa -c 0.9
```

C.2 Indexing with GIGGLE

```
giggle index -i "RMSK/*.gz" -o Taculeatus_WholeGenome_index/ -f -s
```

D Supplementary Figures

D.1 Unknown 103bp repeat sequence

```
>Tacu_5primerepeat
GCCCGCCCGCCCGCCCCCATCACACCGCGTGGCCTTACCGCTCCCGGAAGGCTCCTCTCCTCAGAGCGCCGCCATTGACGCCCGGGACTC
TGTTCCCGAGAGG
```

D.2 Coverage by the unknown 103 bp repeat specific to potentially active echidna LINE-2 elements, by chromosome

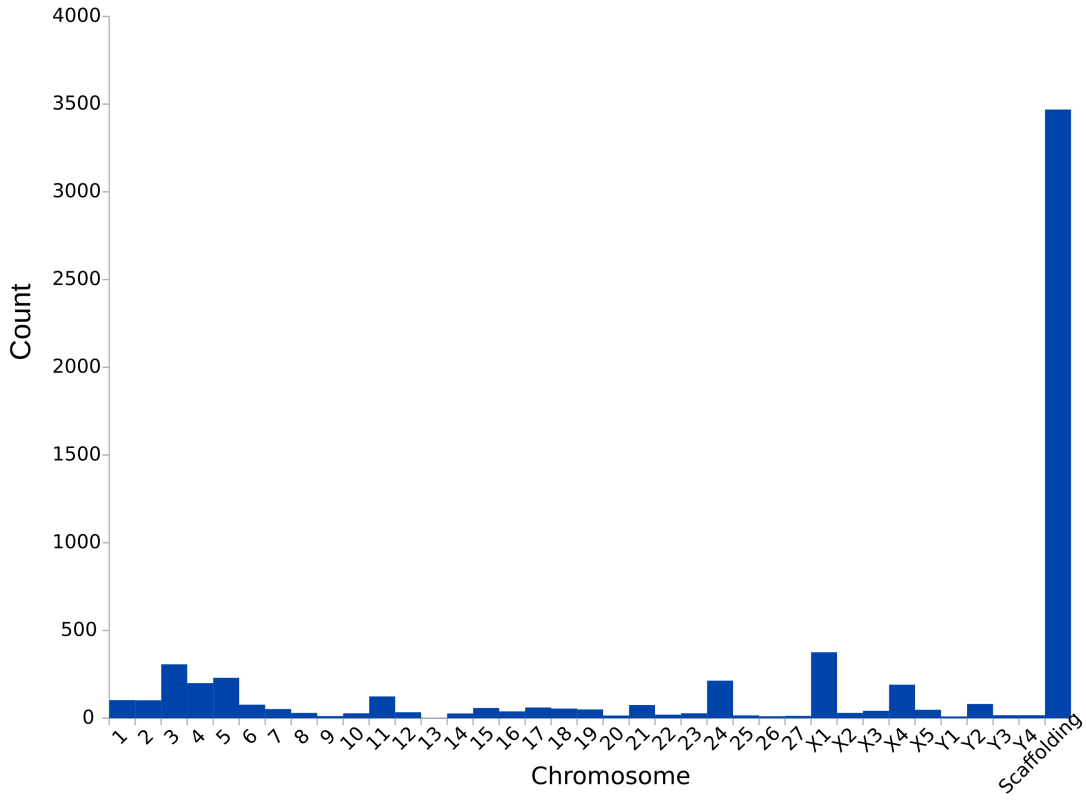


Figure S1: Number of `blastn` hits of 50bp or greater, mapping to the 5' UTR 103bp repeat, by chromosome, in the echidna.

D.3 Self aligments of potentially echidna L2s

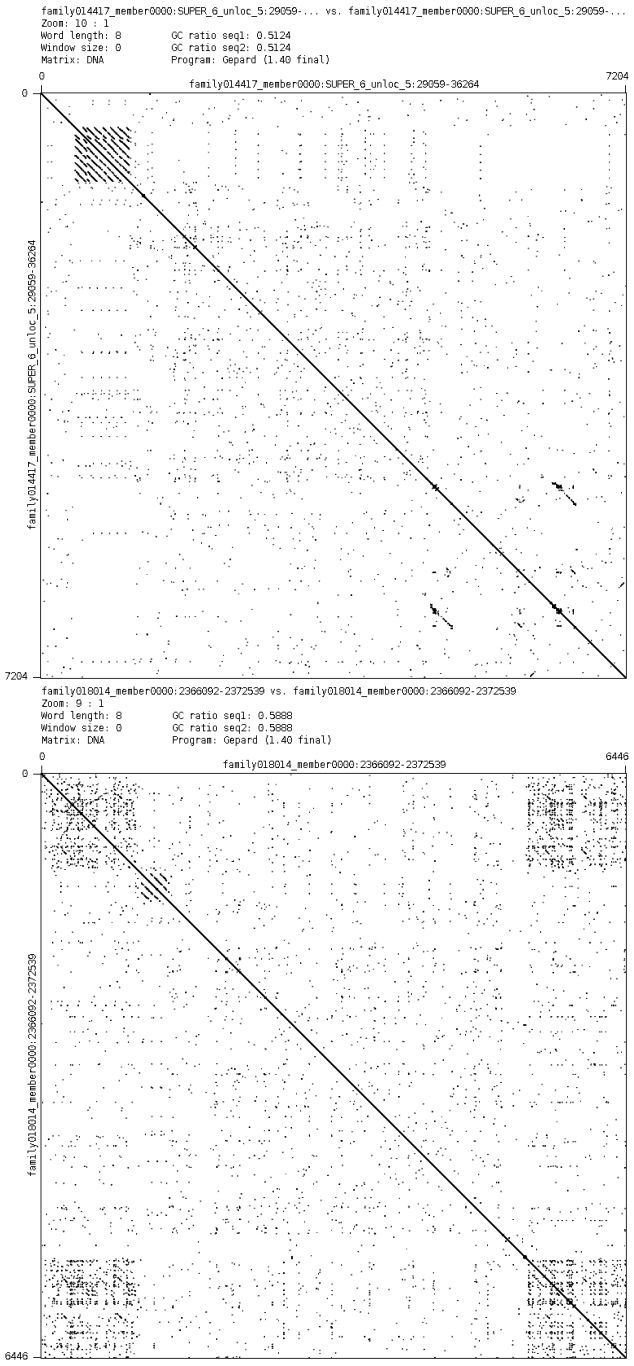
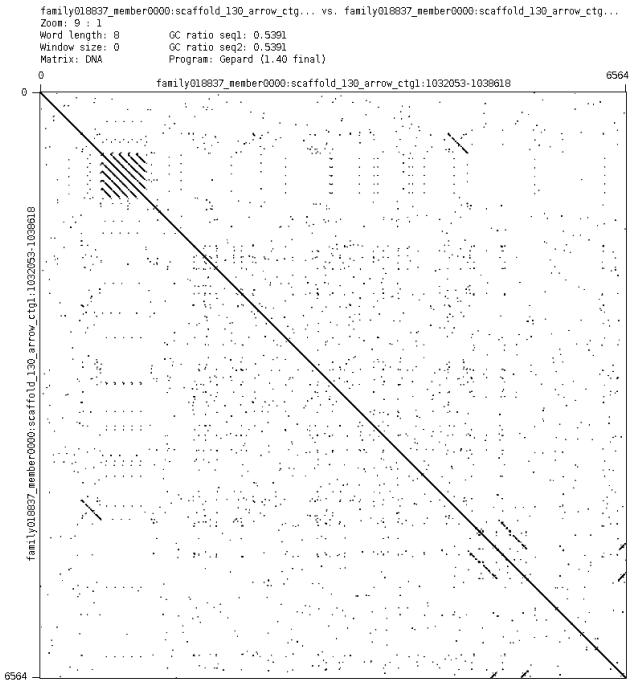
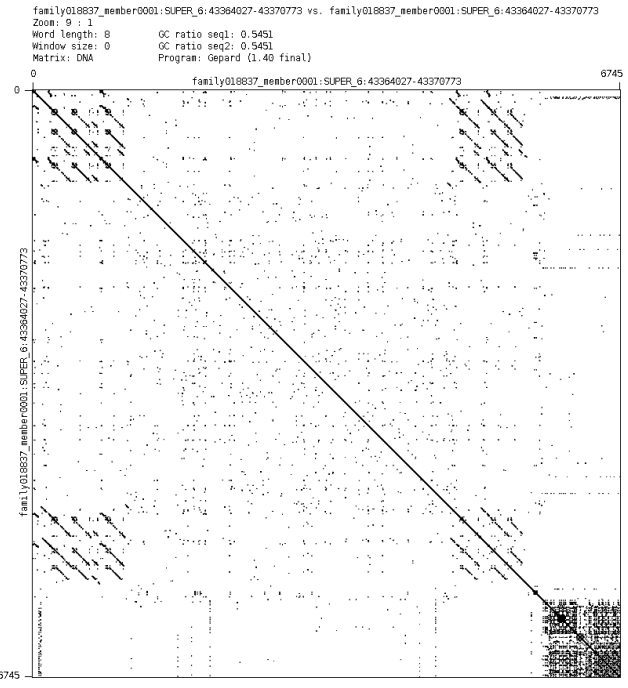
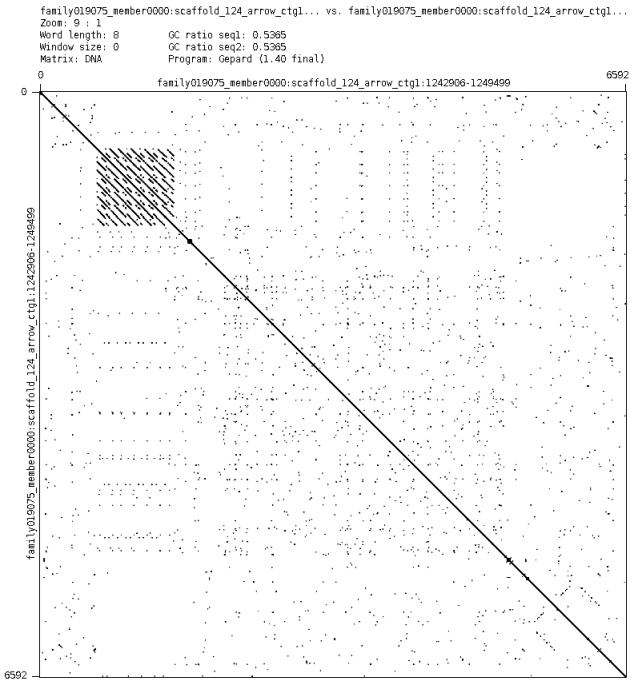


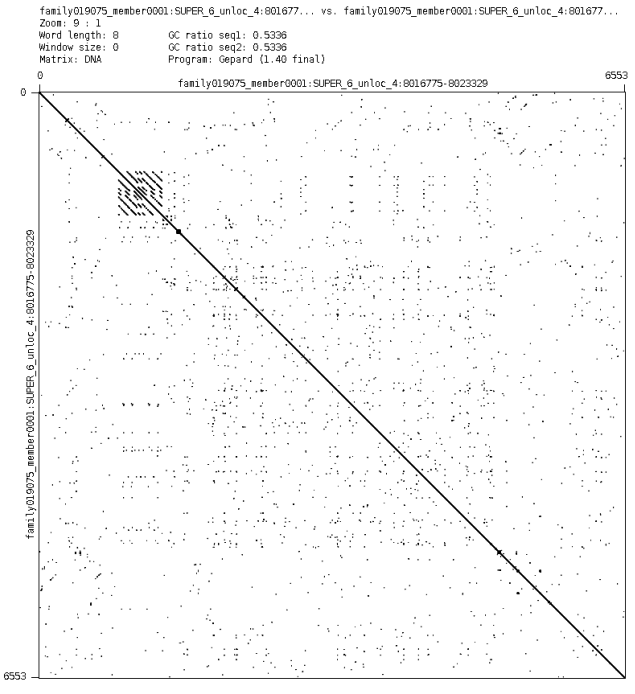
Figure S2: Stranded self alignments of potentially active LINE-2 sequence identified in the echidna, extended by 1000 base pairs in the 5' and 3' directions, visualised using the tool gepard. Note the presence of the repeating sequence at the 5' end

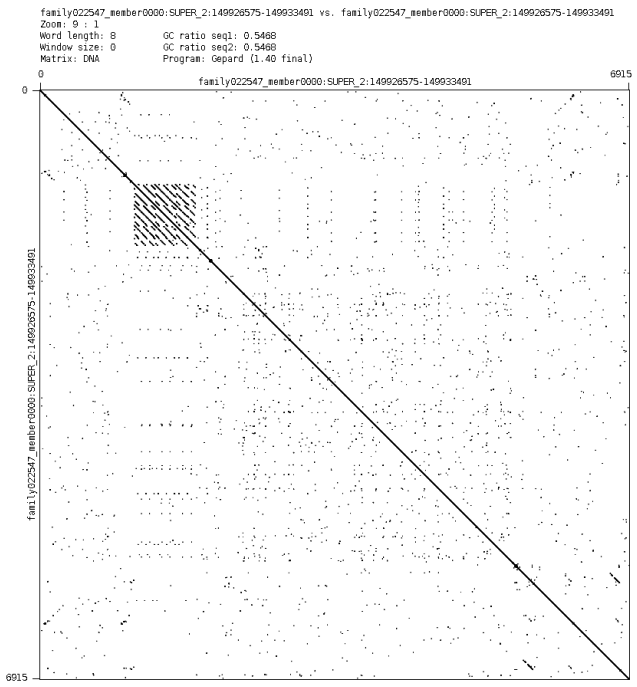


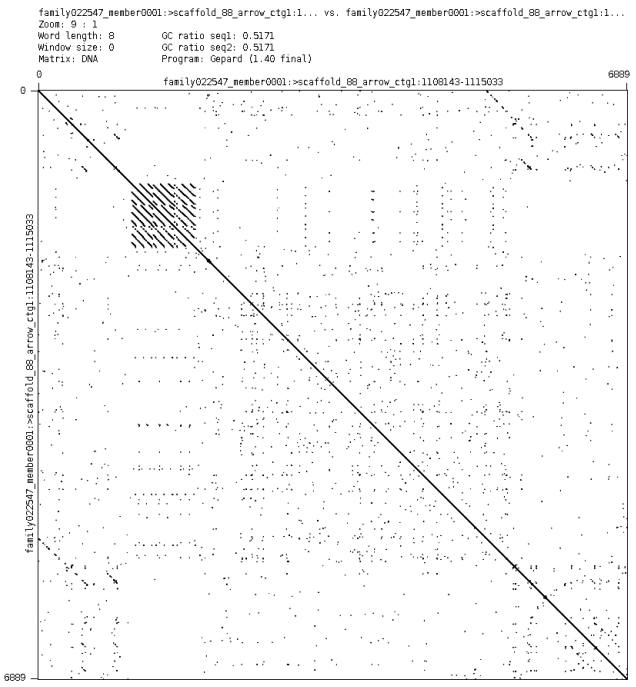


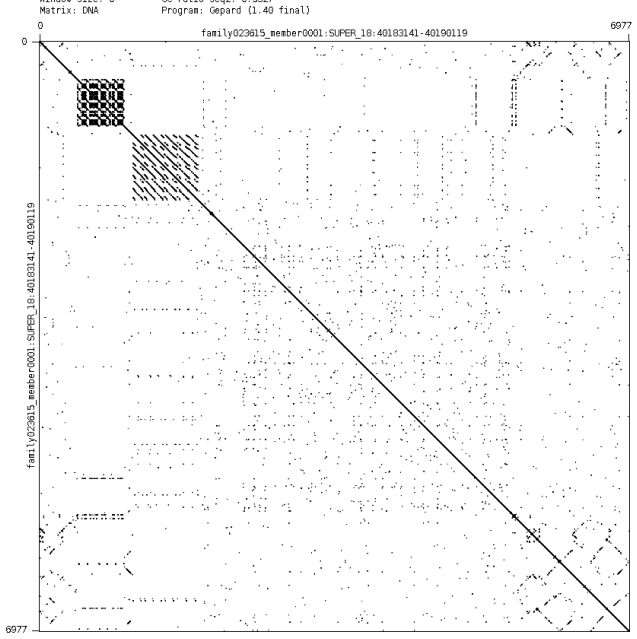
(Note the absence of the typical repeat block, as the 103bp repeat only occurs once here)

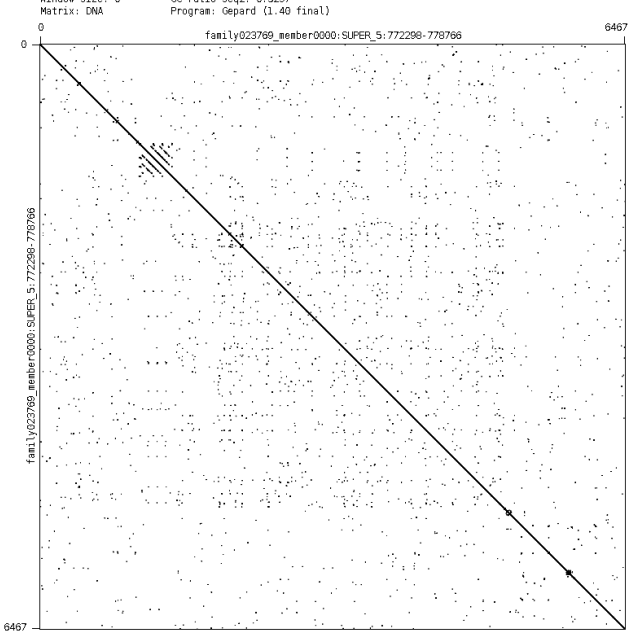


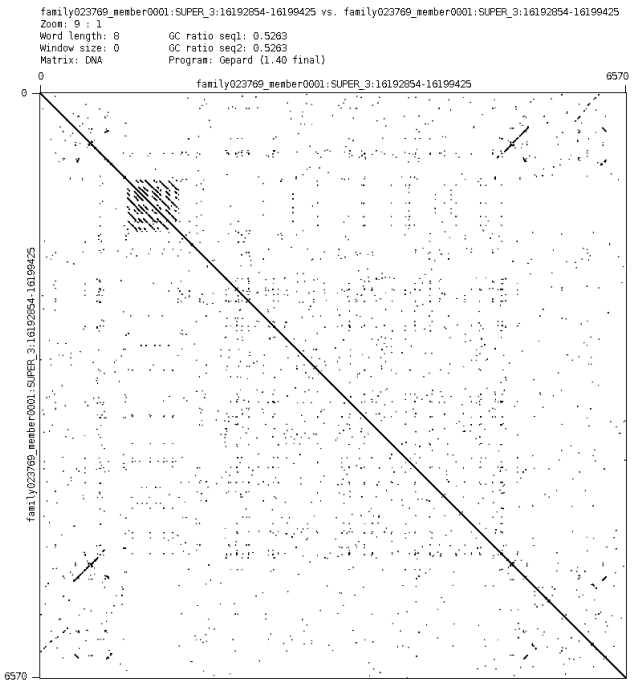


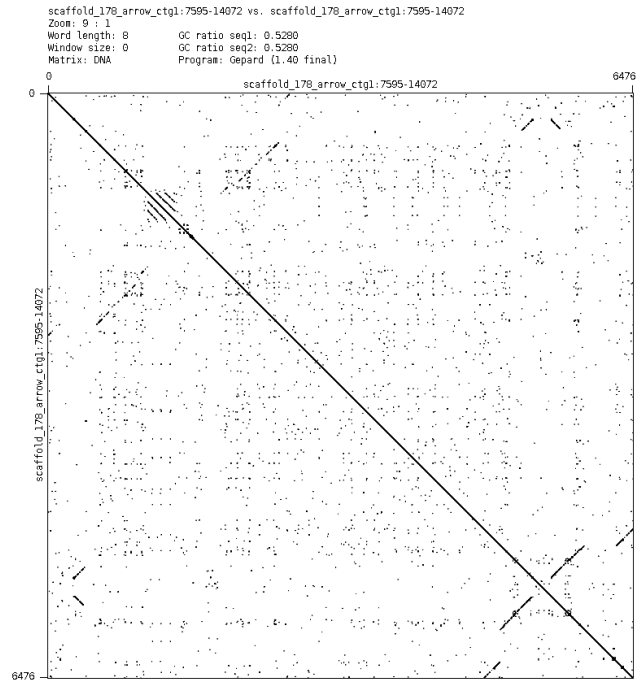

family022547_member0000:SUPER_2:149926575-149933491 vs. family022547_member0000:SUPER_2:149926575-149933491
Zoom: 9 : 1
Word length: 8
Window size: 0
Matrix: DNA
GC ratio seq1: 0.5468
GC ratio seq2: 0.5468
Program: Gepard (1.40 final)
0
6915
family022547_member0000:SUPER_2:149926575-149933491
family022547_member0000:SUPER_2:149926575-149933491


family022547_member0001:>scaffold_88_arrow_ctg1:1... vs. family022547_member0001:>scaffold_88_arrow_ctg1:1...
Zoom: 9 : 1
Word length: 8
Window size: 0
Matrix: DNA
GC ratio seq1: 0.5171
GC ratio seq2: 0.5171
Program: Gepard (1.40 final)
0
6889
family022547_member0001:>scaffold_88_arrow_ctg1:1108143-1115033
family022547_member0001:>scaffold_88_arrow_ctg1:1108143-1115033


family023615_member0001:SUPER_18:40183141-40190119 vs. family023615_member0001:SUPER_18:40183141-40190119
Zoom: 9 : 1
Word length: 8
Window size: 0
Matrix: DNA
GC ratio seq1: 0.5327
GC ratio seq2: 0.5327
Program: Gepard (1.40 final)
0
6977
family023615_member0001:SUPER_18:40183141-40190119
family023615_member0001:SUPER_18:40183141-40190119


family023769_member0000:SUPER_5:772298-778766 vs. family023769_member0000:SUPER_5:772298-778766
Zoom: 9 : 1
Word length: 8
Window size: 0
Matrix: DNA
GC ratio seq1: 0.5297
GC ratio seq2: 0.5297
Program: Gepard (1.40 final)
0
6467
family023769_member0000:SUPER_5:772298-778766
family023769_member0000:SUPER_5:772298-778766


family023769_member0001:SUPER_3:16192854-16199425 vs. family023769_member0001:SUPER_3:16192854-16199425
Zoom: 9 : 1
Word length: 8
Window size: 0
Matrix: DNA
GC ratio seq1: 0.5263
GC ratio seq2: 0.5263
Program: Gepard (1.40 final)
0
family023769_member0001:SUPER_3:16192854-16199425
6570
0
family023769_member0001:SUPER_3:16192854-16199425
6570


scaffold_178_arrow_ctg1:7595-14072 vs. scaffold_178_arrow_ctg1:7595-14072
Zoom: 9 : 1
Word length: 8
Window size: 0
Matrix: DNA
GC ratio seq1: 0.5280
GC ratio seq2: 0.5280
Program: Gepard (1.40 final)
0
scaffold_178_arrow_ctg1:7595-14072
6476
0
scaffold_178_arrow_ctg1:7595-14072
6476

D.4 Comparing L2 elements to CR1s in vertebrates

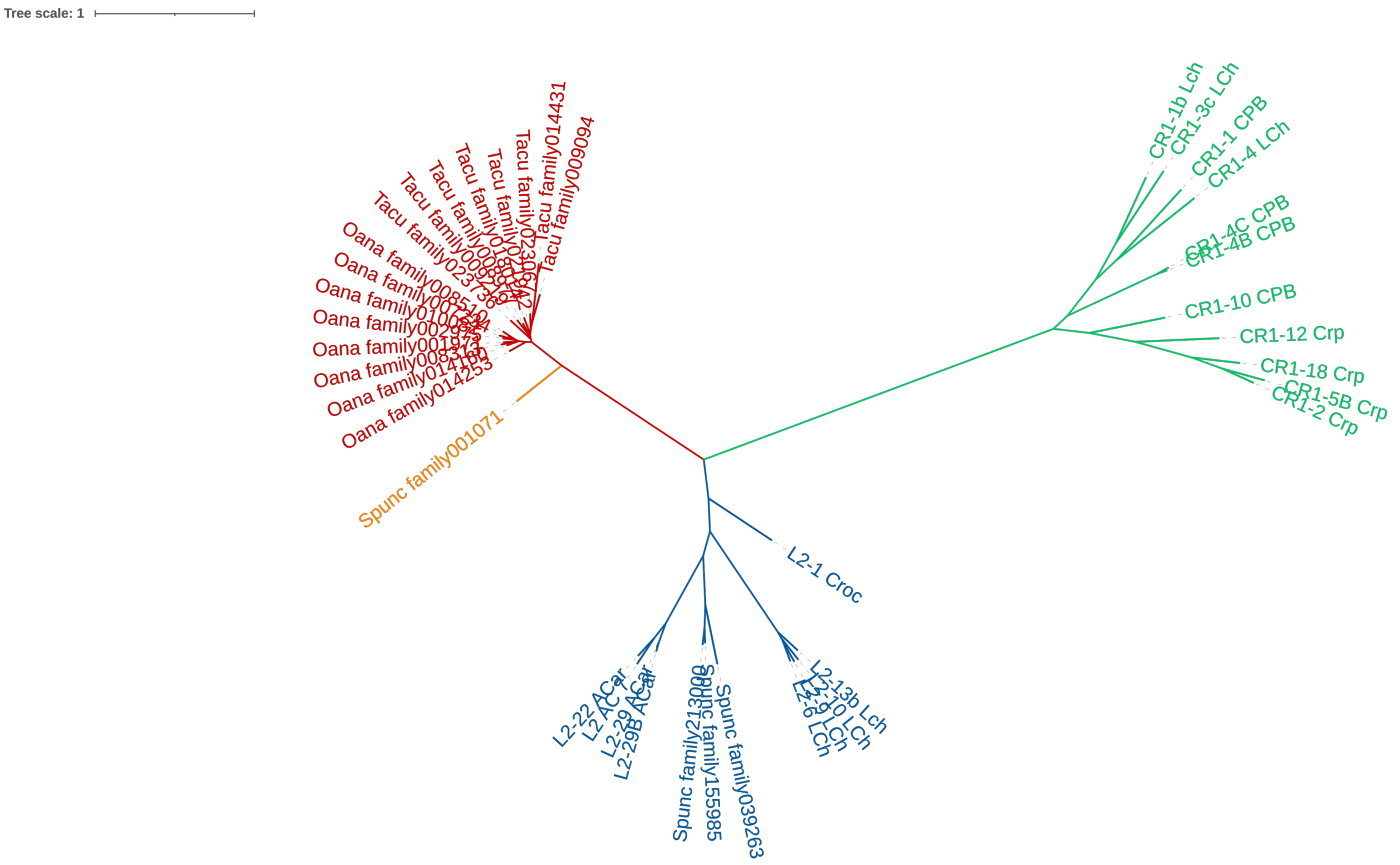


Figure S6: Maximum likelihood tree, showing the phylogeny of the reverse transcriptase domain contained within ORF2 of L2 and CR1 retrotransposons, found in vertebrates. Species included have full length LINE-2 and/or CR1 elements, which are presumed active. L2s are shown in red, orange and blue, while CR1s are shown in green.

D.5 Total bp coverage by protein coding genes, and associated UTRs

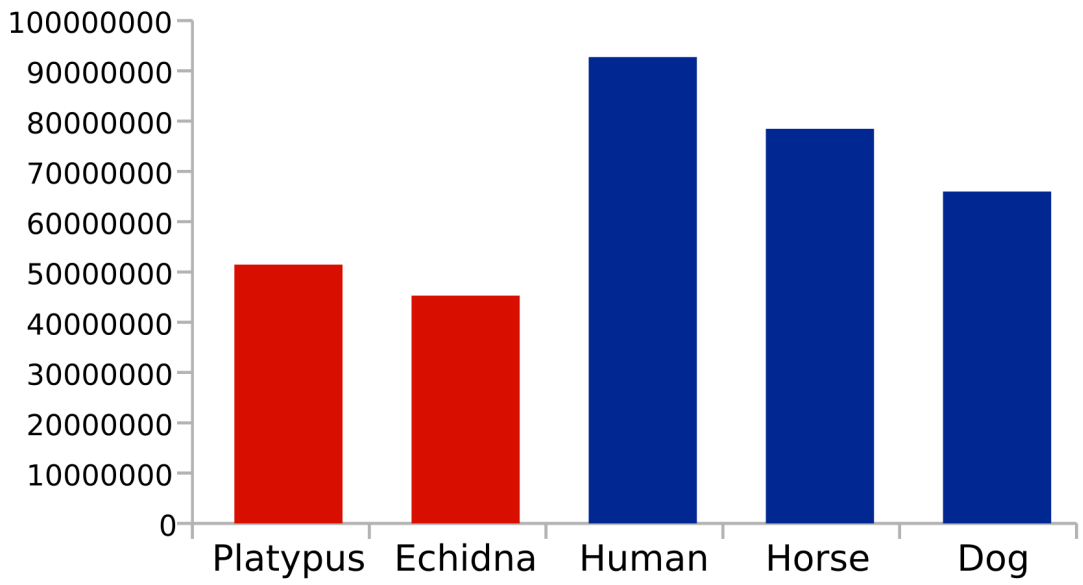


Figure S7: Total coverage by exons associated with genes identified as "protein_coding", as annotated by NCBI, by LINE-2 (red) and LINE-1 (blue) elements. NCBI accession numbers for these species can be found in Subsection 1.3.1

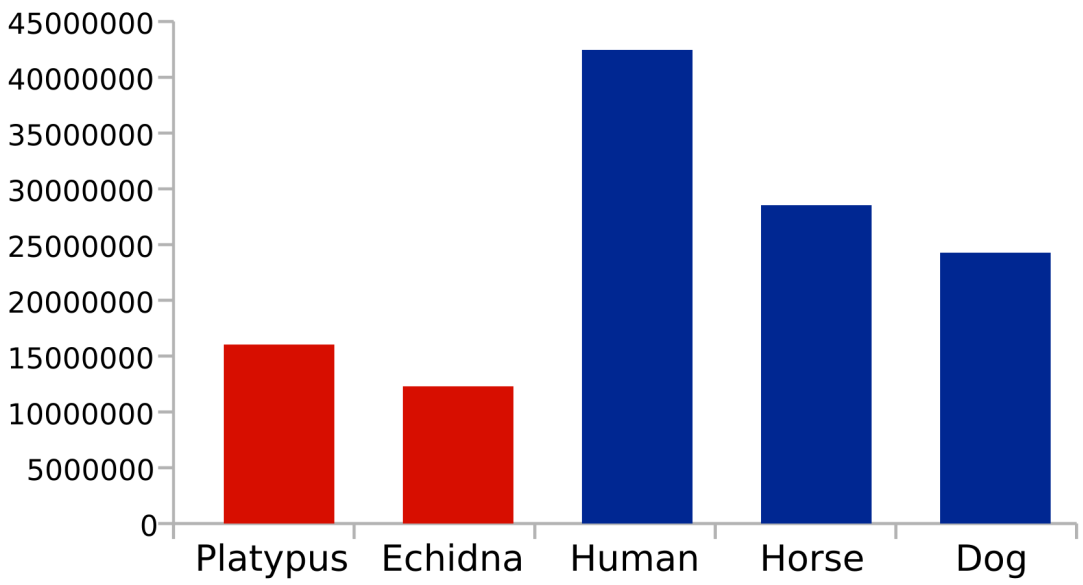


Figure S8: Total coverage by the thee prime untranslated region of genes identified as "protein_coding", as annotated by NCBI, by LINE-2 (red) and LINE-1 (blue) elements. NCBI accession numbers for these species can be found in Subsection 1.3.1

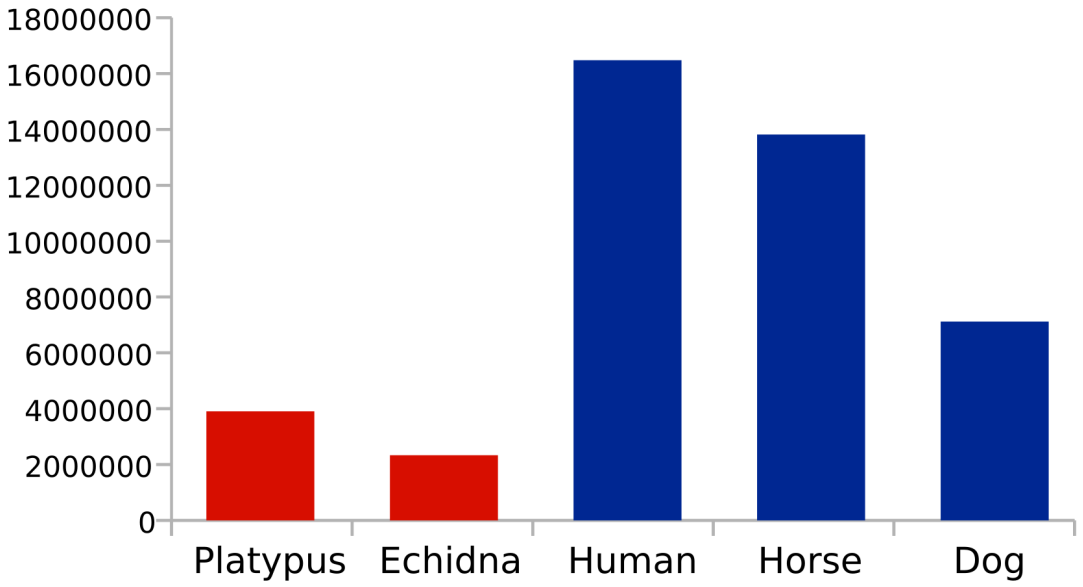


Figure S9: Total coverage by the five prime untranslated region of genes identified as "protein_coding", as annotated by NCBI, by LINE-2 (red) and LINE-1 (blue) elements. NCBI accession numbers for these species can be found in Subsection 1.3.1

Table S2: Table showing summary information about the masking and coverage of repeats as generated by `RepeatMasker`, by class and by family, in the platypus. Parsing conducted by the perl script `ParseRM.pl`. nt_masked_double corresponds to DNA fragments that are masked by several elements; the element with the lower % identity is subtracted when calculating the %_masked statistic.

#TOTAL:			
#nt_total_in_genome	**nt_masked-minus-double**	**%_masked**	**FYI:nt_masked_double**
2029738745	1086141140	53.51137	1045027566
#BY CLASS			
#class	**nt_masked-minus-double**	**%_masked**	**FYI:nt_masked_double**
LTR	708716	0.03492	
Mariner	369111	0.01819	6310
ERV	1610240	0.07933	99625
MuDR	47	0.00000	38
ERV3	563818	0.02778	16438
ERV2	32241	0.00159	14157
Unclassified	527188796	25.97323	328350756
CR1	808634	0.03984	31635
Transposable	1827836	0.09005	220876
satellite	30	0.00000	16
Multicopy	1197403	0.05899	365222
R4	26461	0.00130	74
RTEX	7	0.00000	7
Rex1	423	0.00002	386
Gypsy	1372	0.00007	1236
L2	539192130	26.56461	222100304
CryptonV	20463	0.00101	17822
piggyBac	139	0.00001	0
DIRS	320	0.00002	100
EnSpm	3149	0.00016	2450
BEL	13	0.00000	0
ERV1	123609	0.00609	68594
DNA	2127	0.00010	957
hAT	143559	0.00707	118350
Penelope	158	0.00001	0
Endogenous	15276	0.00075	0
Zisupton	33121	0.00163	26154

Continued on next page...

#BY CLASS			
#class	**nt_masked-minus-double**	**%_masked**	**FYI:nt_masked_double**
Helitron	22138	0.00109	16474
SINE	6205302	0.30572	132689
Kolobok	616	0.00003	541
L1	79	0.00000	59
Retrovirus_like	1925145	0.09485	730585
SAT	488275	0.02406	393880
Polinton	218	0.00001	46
Unspecified	295357	0.01455	117431
Harbinger	12	0.00000	9
RTE	3334774	0.16430	102920

Table S3: Table showing summary information about the masking and coverage of repeats as generated by `RepeatMasker`, by class and by family in the echidna. Parsing conducted by the perl script `ParseRM.pl`. nt_masked_double corresponds to DNA fragments that are masked by several elements; the element with the lower % identity is subtracted when calculating the %_masked statistic.

#TOTAL:			
#nt_total_in_genome	**nt_masked-minus-double**	**%_masked**	**FYI:nt_masked_double**
1859281927	516528621	27.78108	254629704
#BY CLASS			
#class	**nt_masked-minus-double**	**%_masked**	**FYI:nt_masked_double**
BEL	129	0.00001	66
Tx1	6122	0.00033	5099
ERV3	1591938	0.08562	112606
DIRS	37	0.00000	0
Harbinger	1161	0.00006	112
Nonautonomous	764	0.00004	239
CryptonV	521	0.00003	304
Polinton	10102	0.00054	7318
EnSpm	13447	0.00072	9069
SAT	9751	0.00052	7199
Nimb	867	0.00005	0
REP-24_CPB	309	0.00002	0
ERV1	617708	0.03322	33337
ERV	14089	0.00076	642
MARE8	142	0.00001	0
R4	124982	0.00672	204
L1	989	0.00005	538
Endogenous	3160	0.00017	0
Non-LTR	664	0.00004	0
Dada	88	0.00000	0
Ginger1	468	0.00003	114
Mariner	1955279	0.10516	32531
LTR-775_Gav_odd	9571	0.00051	0
SINE	9040788	0.48625	145016
ERV4	516	0.00003	0
Gypsy	51126	0.00275	31478

continued...

#BY CLASS			
#class	nt_masked-minus-double	%_masked	FYI:nt_masked_double
piggyBac	497	0.00003	0
MSRBMI	849	0.00005	0
IS3EU	1044	0.00006	1028
Penelope	18422	0.00099	13115
Simple	2066	0.00011	1784
Unclassified	30237960	1.62632	12215634
SINE3	5758	0.00031	0
Repetitive	4999	0.00027	204
Helitron	57490	0.00309	41270
Transposable	2961514	0.15928	296789
ERV2	168882	0.00908	839
DNA	38623	0.00208	6542
Kolobok	9589	0.00052	1989
Crypton	5511	0.00030	323
Retrovirus_like	1717636	0.09238	1029615
CR1	72874223	3.91948	22791328
conserved	370	0.00002	0
Zisupton	52151	0.00280	32107
hAT	706057	0.03797	56988
L2	386514100	20.78835	102264144
Multicopy	215505	0.01159	16458
Vingi	6710	0.00036	0
RTE	7473696	0.40197	170250
MuDR	251	0.00001	128

E Retrotransposons identified *de novo* in *Tachyglossus aculeatus*

E.1 LINE-2 retrotransposons

Fasta headers are in the format >RepeatName#RepeatClass

>Tacu_L2a#L2

GCCCGCCCGCCCGCCCCCATCACACCGCGTGGCCTTACCGCTCCCGGAAGGCTCCTCTCCTCAGAGCGCCGCCATTGACGCCCGGGACTC
TGTTCCCGAGAGGCCCGGCCGCCCACCCTCCCACCATCTCAGCGCCCCCACGGACACCCCCCCTGCTTCCACCCCCCCACTTCGGAGGCC
TTCGGTAAAGTGCCCCTCATCCCCCCCCACCTTCCCGGTCCGGTTCTGACTGACTCTCCCCCCGCCCGCCCCCAGGTAAAAAGCCCTTGT
TAGCACCTTGTTAGCACCATGTTAGCACCATGTTACATGTTACATTAACGACAACCTCATTCACACCTCCTCAGCCCTCCCATGCTAGTT
GACTGTTCGCTCCCCTCCCCAGGTATCTCTGCTCCCTGACCCCCCTTAACTGGCCCACCCTACTCCAGCTCTTAATAGTTTGTATCTGCT
CTCTGTGGCATCATTGTCTGCCAATTCTATTATTGCCATAGCATATTGATTGCTCAACTACACCCTGTGATGCCATCGCCTACTTTTATC
ATTGTTACTGTTTTATTGCTTCTGCATCTCAATTGTTAACCCCCCGCGGTACTGTCACTCGCCCTGCTGACCTCACCATGAACTATCAGT
TCCCTTGCTCCCAACTGCCAGTCCCATTCCTCACCTTCCCCTTCCCCTCTCCCACCCAGCTGTTTCCCTCCCTCTCACCCACCAAGACCG
CTCCCCCTCAACCCCTACCAAGGATTCCTCTGTACCAGCGCAGGTCCTCCGCTTCTCCCTCCCGGCCCCCCTCCTCCCCTTCTCCCCGCC
CCCACACCATCCCAGTTCTCCTTTCCCATTGCCACCCCCCTTCCCACTCTCCCCGCCCAGGCCCCCGCCAACTCATCCCAATCCAAACCC
TCCCCACCCCTCGCACCCTTCCCCCTCCCTCCCCACCCTCGACAGCTGATGCCAAGTGTGGCCTCTGGAACCCCCGCTCCGTTTTAAGTA
AGATCCCTTTCATCCTGGACCTTTTCCTTTCCAGTTCTCTACTCCTCCTCGCCCTAACTGAAACATGGCTGTCGCCGGACGACACGGTCT
CTTCTGCTGCTCTCTGCAGTGCAGGCCTCTTCTTCTCCCACTCCCCCAGACTCACCGGAAAAGGAGGAGGTGTCGGTTTCCTTCTCGCCC
CCCAATGTCGCTTTCGCACTATCCCTCCTCCCCCTTCCCTTTCCTTCCCTTCCTTTGAAGCCCACATTATTCGCCTCTACCACCCCCTCC
AGATTCTTGTAGCCGTCATCTACCGCCCTCCGGGCCCCACTTCCAACTTCTTTAATGACTTTGACCCCTTCCTCACATTCCTTCTCTCCT
TCTCCATGCCCACTCTGATCCTCGGAGACTTCAATATCCACATGGATATCCCTAACGACTCCTCTGCCGCCCGCCTTCTATCTCTCCTTG
ACGCTGCCAACCTCTTCCTCCACCCCACCTCACCCACTCACCAACTTGGTCATACCCTCGACCTCATCATCTCCTACCGCTGCACTGTGT
CCACCCTCACCAACTCTGTGATCCCTCTCTCTGATCATAATCTTCTCACCTGCCTCCTCACTCACACTCCTTTCCCCTGTAAATCCGTTT
TACTCCCTCACAGAGATCTCCGCTCTCTGGACCCCACCCATCTTTCGGAGCGCCTCACACCCCACCTCGACGCCCTCTCCTCTCTACCCA
GTCTTGATGATCAGATTACTGCTCTCAACTCTACCCTTTCTACTCAGCTAGACTCGCTCGCTCCCCTTTCCCTTCGCCGCTCTCGCACCA
CTAACCCACAGCCCTGGATCACTGCCACTGTCCGCCTCCTTCGCTCTTATGCTCGAGCTGCCGAACGCTGCTGGCGAAAGTCTAAACACC
ATGCCAACCTCGTTCACTTCAAGTTTATCCTTTCCTGCCTTAACTCAGCCCTCTCTTCTGCCAGACAAAACTATTTCTCCTCCCTTATTG
ACACCCATGCCCATCACCCCCGCCAGCTCTTCCGTACATTCAACTCCCTTCTCAGGCCCCCGGTTCCTCCCCCTCCTCCTTCCCTCACCC
CCAACGATCTGGCCTCCTACTTCATTAACAAAATTAAATCCATCAGGTCCGACCTCCCCAAAGTCTCTTCCCCCCTTTCTCCAACCCCCC
GGCTCTCAACACTCTCTGCTACTCTCCCATCCTTCCCAGCGGTATCCTCAGAGGAACTCTCCTCCCTCCTCTCAAGTGCTACTCCGGCCA
CCTGTGCTTCTGACCCCATTCCCTCTCATCTTATGAAATCTCTCGCTCCATCCCTTCTCCCCTCCTTAACTTCCATCTTCAACCGCTCAC
TCTCCACTGGTTCCTTCCCCTCTGCCTTCAAACATGCCCATGTCTCTCCCATCCTAAAAAAACCCTCTCTTGACCCCACCTCACCTTCTA
GTTATCGTCCCATATCCCTCCTACCATTCCTTTCCAAACTCCTTGAACGAGTTGTCTACACGCGCTGTCTAGAATTCCTCAACAACAACT
CTCTCCTCGACCCCCTCCAGTCTGGCTTCCGTCCCCTTCATTCCACGGAAACTGCGCTCTCAAAGGTCACCAATGACCTCCTGCTTGCCA
AATCCAACGGCTCATACTCTGTCTTAATCCTCCTCGACCTCTCAGCTGCCTTTGACACTGTGGACCACCCCCTTCTCCTCAACACGTTAT

CTGACCTTGGCTTCACTGACTCCGTCCTCTCCTGGTTCTCCTCTTATCTCTCCGGTCGTTCTTTCTCAGTCTCTTTTGCAGGCTCCTCCT
CCCCCTCCCATCCTCTTACTGTGGGGGTTCCCCAAGGTTCAGTGCTTGGTCCCCTTCTGTTCTCAATCTACACTCACTCCCTTGGTGGCC
TCATTCGCTCCCACGGCTTCAACTATCATCTCTACGCTGATGACACCCAGATCTACATCTCTGCCCCTGCTCTCTCCCCCTCCCTCCAGG
CTCGCATCTCCTCCTGCCTTCAGGACATCTCCATCTGGATGTCCGCCCGCCACCTAAAGCTCAACATGTCGAAGACTGAGCTCCTTGTCT
TCCCTCCCAAACCTTGTCCTCTCCCTGACTTTCCCATCTCTGTTGACGGCACTACCATCCTTCCCGTCTCACAAGCCCGCAACCTCGGTG
TCATCCTCGACTCCGCTCTCTCATTTACCCCTCACATCCAAGCCGTCACCAAAACCTGCCGGTCTCAGCTCCGCAACATTGCCAAGATCC
GCCCTTTCCTCTCCATCCAAACTGCTACCCTGCTCATTCAAGCTCTCATCCTATCCCGTCTGGACTACTGCACCAACCTTCTCTCTGATC
TCCCATCCTCGTGTCTCTCTCCACTTCAATCCATACTTCATGCTGCTGCCCGGATTATCTTTGTCCAGAAACGCTCTGGACATATTACTC
CCCTCCTCAAAAACCTCCAATGGCTACCGATCAATCTGCGCATCAAGCAGAAACTCCTCACCCTGGGCTTCAAGGCTCTCCATCACCTCG
CCCCCTCCTACCTCACCTCCCTTCTCTCCTTCTACTGCCCAGCCCGAACCCTCCGCTCCTCCACCACTAATCTCCTCACCGTACCTCGCT
CTCGCCTGTCCCGCCATCGACCCCCGGCCCACGTCATCCCACGGGCCTGGAATGCCCTCCCTCTGCCCCTCCGCCAAGCTAGCTCTCTTC
CTCCCTTCAAGGCCCTGCTGAGAGCTCACCTCCTCCAGGAGGCCTTCCCAGACTGAGCCCCTTCTTTCCTCTCCCCTCGTCCCCCTCTCC
ATCCCCCCGTCTTACCTCCTTCCCTTCCCCACAGCACCTGTATATATGTATATATGGTTGTACATATTTATTACTCTATTTATTTATTTA
TTTATTTATTTTACTTGTACATTTCTATCCTACTTATTTTATTTTGTTGGTATGTTTGGTTCTGTTCTCTGTCTCCCCCTTTTAGACTGT
GAGCCCACTGTTGGGTAGGGACTGTCTCTATGTGATGCCAATTTGTACTTCCCAAGCGCTTAGTACAGTGCTCTGCACATAGTAAGCGCT
CAATAAATACGATTGATTGATTGATTGATT

>Tacu_L2b#L2
TCCCGCCCGCCCCCATCACACCGCGCGGCCTTGCCGCTCCCGGAAGGCTCCTCTCCTCAGAGCGCCGCCATTGATGCCCGGGACTCCGTT
CCCGAGAGGCCCAGCCGCCCACCCTCCCGCCATCTCAGCGCCCCCACGGACGCCCTCCCTGCTTCCACCCCCCCACTTCGGAGGTCTTCC
GTAAAGTGCCCCTCATCCCCCCCCACTTCCCGGTTCGGTCCTGACTGAGTCTCCCCCCGCCCCCAGGTAAAAAGCCCTTGTTAGCACCTT
GTTAGCACCATGTTACATGTTACACTAACGACAACCTCATTCACACCTCCTCAGCCCTCCCATGCTAGTTGACTGTTGGCTCCCCTCCCC
AGGTATCTCTGCTCCCTGACCCCCCTTAACTGGCCCACCCTACTCCAGCTCTTAATAGTTTGTATCTGCTCTCTGTGGCATCATTGTCTG
CCAATTCTATTATTGCCATAGCATATTGATTGCTCAACTACACCCTGTGATGCCATCGCCTACTTTTATCATTGTTACTGTTTTATTGCT
TCTGCATCTCAATTGTTAACCCCCCGCGGTACTGTCACTCGCCCTGCTGACCTCACCATGAACTATCAGTTCCCTTGCTCCCAACTGCCA
GTCCCATTCCTCACCTTCCCCTTCCCCTCTCCCACCCAGCTGTTTCCCTCCCTCTCACCCACCAAGACCGCTCCCCCTCAACCCCTACCA
AGGATTCCTCTGTACCAGCGCAGGTCCTCCGCTTCTCCCTCCCGGCCCCCCTCCTCCCCTTCTCCCCGCCCCCACCCCATCCCAGTTCTC
CTTTCCCATCGCCACCCCCCTTCCCACTCTCCCCGCCCAGGCCCCCGCCAACTCATCCCAATCCAAACCCTCCCCACCCCTCGCACCCTT
CCCCCTCCCTCCCCACCCTCGACAGCTGATGCCAAGTGTGGCCTCTGGAACCCCCGCTCCGTTTTAAGTAAGATCCCTTTCATCCTGGAC
CTTTTCCTTTCCAGTTCTCTACTCCTCCTCGCCCTAACTGAAACATGGCTGTCGCCGGACGACACGGTCTCTTCTGCTGCTCTCTGCAGT
GCAGGCCTCTTCTTCTCCCACTCCCCCAGACTCACCGGAAAAGGAGGAGGTGTCGGTTTCCTTCTCGCCCCCCAATGTCGCTTTCGCACT
ATCCCTCCTCCCCCTTCCCTTTCCTTCCCTTCCTTTGAAGCCCACATTATTCGCCTCTACCACCCCCTCCAGATTCTTGTAGCCGTCATC
TACCGCCCTCCGGGCCCCACTTCCAACTTCTTTAATGACTTTGACCCCTTCCTCACATTCCTTCTCTCCTTCTCCATGCCCACTCTGATC
CTCGGAGACTTCAATATCCACATGGATATCCCTAACGACTCCTCTGCCGCCCGCCTTCTATCTCTCCTTGACGCTGCCAACCTCTTCCTC
CACCCCACCTCACCCACTCACCAACTTGGTCATACCCTCGACCTCATCATCTCCTACCGCTGCACTGTGTCCACCCTCACCAACTCTGTG
ATCCCTCTCTCTGATCATAATCTTCTCACCTGCCTCCTCACTCACACTCCTTTCCCCTGTAAATCCGTTTTACTCCCTCACAGAGATCTC
CGCTCTCTGGACCCCACCCATCTTTCGGAGCGCCTCACACCCCACCTCGCCGCCCTCTCCTCTCTACCCAGTCTTGATGATCAGATTACT
GCTCTCAACTCTACCCTTTCTACTCAGCTAGACTCGCTCGCTCCCCTTTCCCTTCGCCGCTCTCGCACCACTAACCCACAGCCCTGGATC
ACTGCCACTGTCCGCCTCCTTCGCTCTTATGCTCGAGCTGCCGAACGCTGCTGGCGAAAGTCTAAACACCATGCCAACCTCGTTCACTTC

AAGTTTATCCTTTCCTGCCTTAACTCAGCCCTCTCTTCTGCCAGACAAAACTATTTCTCCTCCCTTATTGACACCCATGCCCATCACCCC
CACCAGCTCTTCCGTACATTCAACTCCCTTCTCAGGCCCCCGGTTCCTCCCCCTCCTCCTTCCCTCACCCCCAACGATCTGGCCTCCTAC
TTCATTAACAAAATTAAATCCATCAGGTCCGACCTCCCCAAAGTCTCTTCCCCCCTTTCTCCAACCCCCCGGCTCTCAACACTCTCTGCT
ACTCTCCCATCCTTCCCAGCGGTATCCTCAGAGGAACTCTCCTCCCTCCTCTCAAGTGCTACTCCGGCCACCTGTGCTTCTGACCCCATT
CCCTCTCATCTTATGAAATCTCTCGCTCCATCCCTTCTCCCCTCCTTAACTTCCATCTTCAACCGCTCACTCTCCACTGGTTCCTTCCCC
TCTGCCTTCAAACATGCCCATGTCTCTCCCATCCTAAAAAAACCCTCTCTTGACCCCACCTCACCTTCTAGTTATCGTCCCATATCCCTC
CTACCATTCCTTTCCAAACTCCTTGAACGAGTTGTCTACACGCGCTGCCTAGAATTCCTCAACAACAACTCTCTCCTCGACCCCCTCCAG
TCTGGCTTCCGTCCCCTTCATTCCACGGAAACTGCGCTCTCAAAGGTCACCAATGACCTCCTGCTTGCCAAATCCAACGGCTCATACTCT
GTCCTAATCCTCCTCGACCTCTCAGCTGCCTTTGACACTGTGGACCACCCCCTTCTCCTCAACACGTTATCTGACCTTGGCTTCACTGAC
TCCGTCCTCTCCTGGTTCTCCTCTTATCTCTCCGGTCGTTCTTTCTCAGTCTCTTTTGCAGGCTCCTCCTCCCCCTCCCATCCTCTTACT
GTGGGGGTTCCCCAAGGTTCAGTGCTTGGTCCCCTTCTGTTCTCAATCTACACTCACTCCCTTGGTGGCCTCATTCGCTCCCACGGCTTC
AACTATCATCTCTACGCTGATGACACCCAGATCTACATCTCTGCCCCTGCTCTCTCCCCCTCCCTCCAGACTTGCATCTCCTCCTGCCTT
CAGGACATCTCCATCTGGATGTCTGCCCGCCACCTAAAACTCAACATGTCTAAGACTGAACTCCTTGTCTTCCCTCCCAAACCCTGCCCT
CTCCCTGACTTTCCCATCACTGTTGACAGCACTACCATCCTTCCCATCTCACAAGCCCACAACCTTGGTGTCATCCTCGACTCTGCTCTC
TCGTTCGCCCCTCACATCCAAGCCATCACCAAAATCTGCCGGTCTCAGCTCCGCAACATTGCCAAGATTTGCCGTTTCCTCTCCATCCAA
ACCACTATCCTGCTCGCTGAAGCTCTCATCCTATCCCGTCTGGACTACTGTATCAGCCTCCTATCCGATCTCCCATCTTCTTGTCTCTCC
CCACTTCAATCCATACTTCACACCACTGCCTGGATTGTCTTTGTCCAGAATCGCTCTGGGCATGATACTCCCCTCCTCAAAAATCTCCAG
TGGCTACCAATCAACCTACACATCAGGCAGAAACTCCTCACCCTCAGCTTCAAGGCTGTCCATCACCTCGCCCCCTCCTACCTCTCCTCC
CTTCTCTCCTTCTACAGCCCAGCCCACACCCTCCACTCCTCTGCCGCTAATCTCCTCACCGTGCCTCGTTCTCACCTGTCCCGCCGACGA
CCCCCGGCCCACGTCATCCCCCTGGCCTGGAATGCCCTCCCTCCCCACATCCACCAAGCTAGCTCTCTTCCTCCCTTCAAGGCCCTACTG
AGAGCTCACCTCCTCCAGGAGGTCTTTCTAGACTGAACCCCCTCCTTCCTCTCCCCCTCCTCCCCCCTCTCCATCCACCCTGCCTTACCTC
TTTCCCTTCCCCACAGCACCTGTATATATGTATATATGTTTGCACGGATTTATTACTCTATTTATTTTACTTGTATGTATTTATTCTATT
CTATTTATTTTATCTTGTTAATATGTTTTGTTTTAGTCTCTGTCTCCCCCTTCTAGACTGTGAGCCCAATGTTGGGTAGGGACCATCTCT
ATATGTTGCCAACTTGTACTTACCAAGCACTTAGTACAGTGCTCTGCACACAGTAAGTGCTCAATAAATATGATTGATTGATTGATT

>Tacu_L2c#L2
GCCCGCCCGCCCGCCATCACACCGCGCGGCCTTGCTGCTCCCGGAAGGATCCTCTCCTCAGAGCGCCGCCATAGACGCCCGGGACTCCGT
TCCCGAGAGGCCCGGCCGCCCACCCACCCGCCATCTCAGCGCCCCCACGGACGCCCCCCCTGCTTCCACCCCCCCACTTAGGAGGCCTTC
CGTAAAGTGCCCCTCGTCCCCCCGCCTTCCCGGTCCGGTCCCGACTGACTCTCCCCCCGCCCCCAGGTAAAAAGCCCTTGTTAGGACCTT
GTTAGCACCATGTTACATGTTACATTAAAGACAACCTCATTCACACCTACTCAGCCCTCCCATGCTAGTTGACTGTTCGCTCCCCTCCCC
AGGTATCTCTGCTCCCTGACCCCCCTTAACTGGCCCAGCCTACTCCAGCTCTTAATAGTTTGTATCTGCTCTCTGTGGCATCATTATCTG
CCAATTCTATTATTGCCATAGCATATTGATTGCTTAACTACACCCTGTGATGCCATCACCTACTTTTATCATTGCTACTGTTTTATTGCT
TCTGCATCTCAATTGTTAACCTCCCGCTGTACTGTCACTCGCCCTGCTGACCTCACCATGAACTTTCAGTTCCCTTGCTCCCAACTGCCA
TTCCCATTCCTCACCTTCCCCTTCCCCTCTCCCACCCAGCTGTTTCCCTCCCTCTCACCCACCAAGACCACTCCCCCTCACCCCCTACCA
AGGATTCCTCTGTACCAGCGCAGGTCCTCCGCTTCTCCCTCCCGGCCCCCTTCCTCCCCTTCTCCCCGCCCCCACCTCATCCCAGTTCTC
CTTTCCCATCGCCACCCCCCTTCCCACTCTCCCCGCCCAGGCCCCCGCCAACTCAACCCAATCCAAACCCTCCCCACCCCTCGCACCCTT
CCCCCTCCCTCCCCTCCATCGACAGCTGCTGCCAAGTGTGGCCTCTGGAACCCCCGCTCCGTTTTAAGTAAGACCCCTTTCATCCGGGAC
CTTTTCCTTTCCAGTTCTCTAGTCCTCCTCGCCCTAACTGAAACATGGCTGTCGCCGGACGACACAGTCTCTTCTGCTGCTCTCTGCAGT
GCAGGCCTCTTCTTCTCCCACTCCCCCAGACTCACGGGAAAAGGAGGAGGTGTCGGTTTCCTTCTCGCCCCCCAATGTCGCTTTCGCACT

ATCCCTCCTCCCCCTTCCCTTTCCTTCCCTTCCTTTGAAGCCCACATCATTCGCCTCTACGACCCCCTCCAGATTCCTGTAGCCGTCATC
TACCGCCCTCCCGGCCCCACTTCCAACTTCTTTAACGACTTTGACCCCTTCCTCACCTTCCTTCTCTCCTTCTCCATGCCCACTCTGATC
CTCGGAGACTTCAATATCCACATGGATATTCCTAACGACTCCTCTGCCGCCCGCCTTCTATCTCTCCTTGACGCTGCCAACCTCTTCCTC
CACCCCACCTCGCCCACTCACCAACTTGGTCATACCCTCGACCTCATCATCTCCTACCGCTGCACTGTGTCCACCCTCACCAACTCTGTA
ATCCCTCTCTCTGATCATAATCTTCTCACCTGCCTCCTCACTCACACTCCTTTCCCCTGTAAATCCGTATTACTCCCTCACAGAGATCTC
CGCTCTCTGGACCCTACCCATCTTTCGGAGCGCCTCACACCCCACCTCGCCGCCCTCTCCTCTCTACCCAGTCTTGATGATCAGATTACT
GCTCTCAACTCTACCCTTTCTACTCAGCTAGACTCACTCGCTCCCCTTTCCCTTCGCCGCTCTCGTACCACTAACCCACAGCCCTGGATC
ACTGACACTGTCCGCCTCCTTCGCTCTTATGCTCGAGCTGCCGAACGCTGCTGGCGAAAGTCTAAACACCATGCCAACCTCGTTCACTTT
AAGTTTATCCTTTCCTGCCTTAACTCAGCCCTCTCTTCTGCCAGACAAAACTATTTCTCCTCCCTTATTGACACCCATGCCCATCACCCC
CACCAGCTCTTCCGTACATTCAACTCCCTTCTCAGGCCTCCGGTTCCTCCCCCTCCTCCTTCCCTCACCCCCAACGATCTGGCCTCCTAC
TTCATTAACAAAATTATATCCATCAGGTCCGACCTCCCCAAAGTCTCTTCCCCCCTTTCTCCATCCCCCCGGCTCTCAACACTCTCTGCT
ACTCTCCCATCCTTCCCAGCGGTATCCTCAGAGGAACTCTCCTCCCTCCTCTCAAGTGCTACTCCGGCCACCTGTGCTTCTGACCCCATT
CCCTCTCATCTTATGAAATCTCTCGCTCCATCCCTTCTCCCCTCCTTAGCTTCCATCTTCAACCGCTCACTCTCCACTGGGTCCTTCCCC
TCTGTCTTCAAACATGCCCATGTCTCTCCCATCCTAAAAAAACCCTCTCTTGACCCCACCTCACCTTCTAGTTATCGTCCCATATCCCTC
CTACCATTCCTTTCCAAACTCCTTGAACGAGTTGTCTACACGCGCTGCCTAGAATTCCTCAACAACAACTCTCTCCTCGACCCCCTCCAG
TCTGGCTTCCGTCCCCTTCTTTCCACGGAAACTGCTCTCTCAAAGGTCACCAATGACCTCCTGCTTGCCAAATCCAACGGCTCATATTCT
GTCCTAATCCTCCTCGACCTCTCAGCTGCCTTTGACACTGTGGACCACCCCCTTCTCCTCAACACGCTATCTGACCTTGGCTTCACAGAC
TCCGCCCTCTCCTGGTTCTCCTCTTATCTCTCCAGTCGTTCTTTCTCAGTCTCTTTTGCAGGCTCCTCCTCCCCCTCCCATCCTCTTACT
GTGGGGGTTCCCCAAGGTTCAGTGCTTGGTCCCCTTCTGTTCTCAATCTACACTCACTCCCTTGGTGATCTCATTCGCTCCCACGGCTTC
AACTATCATCTCTACGCTGATGATACCCAGATCTACATCTCTGCCCCTGCTCTCTCCCCCTCCCTCCAGGCTCGCATCTCCTCCTGCCTT
CAGGACATCTCCATCTGGATGTCCGCCCGCCACCTAAAGCTCAACATGTCGAAGACTGAGCTCCTTGTCTTCCCTCCCAAAACTTGTCCT
CTCCCTGACTTTCCCATCTCTGTTGACGGCACTACCATCCTTCCCGTCTCACAAGCCCGCAACCTTGGTGTCATCCTCAACTCCGCTCTC
TCATTCACCCCTCACATCCAAGCCGTCACCAAAACCTGCCGGTCTCAGATCCGCAACATTGCCAAGATCCGCCCTTTCCTCTCCAATCAA
AATGCTACCCTGCTCATTCAAGCTCTCATCCTATCCCGTCTGGACTGCTGCACCAGCCTTCTCTCTGATCGCCCATCCTCGTGTCTCTCT
CCACTTCAATCCATACTTCATGCTGCTGCCCGGATTATCTTTGTCCAGAAACGCTCTGGGCATATCACTCCCCTCCTCAAAAATCTCCAG
TGGCTACCAATCAATCTGCGCATCAGGCAGAAACTCCTCACCCTGGGCTTCAAGGCTCTCCATCACCTCGCCCCCTCCTACCTCACCTCC
CTTCTCTCCTTCTACTGCCCAGCCCGCAACCTCCGCTCCTCCACCGCTAATCTCCTCACTGTACCTTGTTCTCGCCTGTCCCGCCGTCGA
CCCCCGGCCCACGTCATCCCCCGGGCCTGAAATGCCCTCCCTCTGCCCCTCCGCCAAGCTAGCTCTCTTCCTCCCTTCAAGGCCCTGCTG
AGAGCTCACCTCCTCCAGGAGGCCTTCCCAGACTGAGCCCGTTCCTTCTTCTCCCCCTCGTCCCCCCTCTCCATCCCCCCATCTTACCTC
CTTCCCTTCCCCACAGCACCTGTATATATGTATATATGGTTGTACATCTTTATTACTCTATTTATTTATTTATTTATCTTACTTGT
ACATTTCTATCCTATTTATTTTATTTTGTTGGTATGTTTTGTTCTGTTCTCTGTCTCCCCCTTTTAGACTGTGAGCCCACTGTTGGGTAG
GGACTGTTTCTATGTGTTGCCAATTTGTATTTCCCAAGCGCTTAGTACAGTGCTCTGCACATAGTAAGCGCTCAATAAATACGATTGATT
GATTGATTGATT

>Tacu_L2d#L2
GCCCGCCCGCCCGCCCGCCATCACACCAAGCGGCCTTGCTGGTCCCGGAAGGCTCCTCTCCTCAGAGCGCCGCCATTGACGCCCGGGACT
CCGTTCCCGAGAGGCCCGGCCGCCCACCCACCCGCCATCTCAGCGCCCCCACGGACGCCCCCCCTGCTTCCACCCCCCCCACTTAGGAGG
CCTTCCGTAAAGTGCCCCTCATCCCCCCCGCTTCCCGGTCCGGTCCTGACTGACTCTCCCCCCGCCCCCAGGTAAAAAGCCCTTGTTAGC
ACCATGTTACATTAACGACAACCTCATTCACACCTACTCAGCCCTCCCATGCTAGTTGACTGTTCGCTCCCCTCCCCAGGTATCTCTGCT

CCCTGACCCCCCTTAACTGGCCCACCCTACTCCAGCTCTTAATAGTTTGTATCTGCTCTCTGTGACATCATTATCTGCCAATTCTATTAT
TGCCATAGCATATTGATTGCTTAACTACACCCTGTGATGCCATCGCCTACTTTTATCATTGCTACTGTTTTATTGCTTCTGCATCTCAAC
TGTTAACCCCCCACTGTATTGTCACTCGCCGTGCTGACCTCACCATGAACTTTCAGTTCCCTTGCTCCCAACTGCCATTCCCATTCCTCA
CCTTCCCCTTCCCCTCTCCCACCCAGCTGTTTCCCTCCCTCTCACCCACCAAGACCACTCCCCCTCACCCCCTACCAAGGATTCCTCTGT
ACCAGCGCAGGTCCTCTGCTTCTCCCTCCCGGCCCCCTCCCTCCCCTTCTCCCCGCCCCCACCCCATCCCAGTTCTCCTTTCCCATTGCC
ACCCCCCTTCCCACTCTCCCCGCCCAGGCCCCCGCCAACTCAACCCAATCCAAACCCTCCCCACCCCTCGCACCCTTCCCCCTCCCTCCC
CACCCTCGACAGCTGATGCCAAGTGTGGCCTCTGGAACCCCCGCTCCGTTTTAAGTAAGATCCCTTTCATCCTGGACCTTTTCCTTTCCA
GTTCTCTACTCCTCCTCGCCCTAACTGAAACATGGCTGTCGCCGGACGACACGGTCTCTTCTGCTGCTCTGCAGTGCAGGCCTCTTCTTC
TCCCACTCCCCCAGACTCACCGGAAAAGGAGGAGGTGTCGGTTTCCTTCTCGCCCCCCAATGTCGCTTTCGCACTATCCCTCCTCCCCCT
TCCCTTTCCTTCCCTTCATTTGAAGCCCACATTATTCGCCTCTACCACCCCCTCCAGATTCTTGTAGCCGTCATCTACCGCCCTCCCGGC
CCCACTTCCAACTTCTTTAACGACTTTGACCCCTTCCTCACCTTCCTTCTCTCCTTCTCCATGCCCACTCTGATCCTCGGAGACTTCAAT
ATCCACATGGATATCCCTAACGACTCCTCTGCCGCCCGCCTTCTATCTCTCCTTGACGCTGCCAACCTCTTCCTCCACCCCACCTCACCC
ACTCACCAACTTGGTCATACCCTCGACCTCATCATCTCCTACCGCTGCACTGTGTCCACCCTCACCAACTCTGTAATCCCTCTCTCTAAT
CATAATCTTCTCACCTGCCTCCTCACTCACACTCCTTTCCCCTGTAAATCTGTACTACTCCCTCAAAGAGACCTCCGCTCTCTGGACCCC
ACCCATCTTTCGGAGCGCCTCACACCCCACCTCGCCGCCCTCTCCTCTCTACCCAGTCTTGATGATCAGATTACTGCTCTCAACTCTACC
CTTTCTACTCAGCTAGACTCACTCGCTCCCCTTTCCCTTCGCCGCTCTCGCACCACTAACCCACAGCCCTGGATCACTGCCACTGCCCGC
CTCCTTTGCTCTTATGCTCGAGCTGCCGAATGCTGCTGGCGAAAGTCTAAACACCATGCCAACCTCGTTCACTTCAAGTTTATCCTTTCC
TGCCTTAACTCAGCCCTCTCTTCTGCCAGACAAAACTATTTCTCCTCCCTTATTGACACCCATGCCCATCACCCCCGCCAGCTCTTCCGT
ACATTCAACTCCCTTCTCAGGCCCCCGGTTCCTCCCCATCCTCCTTCCCTCACCCCCAACGATCTGGCCTCCTACTTCATTAACAAAATT
CAATCCATCAGGTCCGACCTCCCCAAAGTCTCTTCCCCCCTTTCTCCAACCCCCCGGCTCTCAACACTCTCTGCTACTCTCCCATCCTTC
CCAGCGGTATCCTCAGAGGAACTCTCCTCCCTCCTCTCAAGTGCTACTCCGGCCACCTGTGCTTCTGACCCCATTCCCTCTCATCTTATG
AAATCTCTCGCTCCATCCCTTCTCCCCTCCTTAAATTCCATCTTCAACCACTCACTCTCCACTGGTTCCTTCCCCTCTGCCTTCAAACAT
GCCCATGTCTCTCCCATCCTAAAAAAACCCTCTCTTGACCCCACCTCACCTTCTAGTTATCGTCCCATATCCCTCCTACCATTCCTTTCC
AAACTCCTTGAACGAGTTTTCTACACACGCTGCCTAGAATTCCTCAACAACTATTCTCTCCTCGACCCCCTCCAGTCTGGCTTCCGTCCC
CTTCATTCCACGGAAACTGCCCTCTCAAAGGTCACCAATGACCTCCTGCTTGCCAAATCCAACGGCTCATACTCTGTCCTAATCCTCCTC
GACCTCTCAGCTGCCTTTGACACTGTGGACTACGCCCTTCTCCTCAACATGCTATCTGACCTTGGCTTCACAGACTCCGTCCTCTCCTGG
TTCTCCTCTTATCTCTCCGGTCGTTCTTTCTCAGTCTCTTTTGCAGGCTCCTCCTCCCCCTCCCATCCTCTTACTGTGGGGGTTCCCCAA
GGTTCAGTGCTTGGTCCCCTTCTGTTCTCAATCTACACTCACTCCCTTGGTGACCTCATTCGCTCCCACGGCTTCAACTATCATCTCTAC
GCTGATGACACCCAGATCTACATCTCTGCCCCTGCTCTCTCCCCCTCCCTCCAGGCTCGCATCTCCTCCTGCCTTCAGGACATCTCCATC
TGGATGTCCGCCCGCCACCTAGAGCTCAACATGTCGAAGACTGAGCTCCTTGTCTTCCCTCCCAAACCTTGTCCTCTCCCTGACTTTCCC
ATCTCTGTTGACGGCACTACCATCCTTCCCGTCTCACAAGCCCGCAACCTTGGTGTCGTCCTCGACTCCGCTCTCTCATTCACCCCTCAC
ATCCAAGCCGTCACCAAAACCTGCCGGTCTCAGCTCCACAACATTGCCAAGATCCGCCCTTTCCTCTCCATCCAAACCGCTACCCTGCTC
ATTCAAGCTCTCATCCTATCCCGTCTGGAGTACTGCACCAGCCTTCTCTCTGATCTCCCATCCTCGTGTCTCTCTCCACTTCAATCCATA
CTTCATGCTGCTGCCCGGATTATCTTTGTCCAGAGACGCTCTGGGCATATTACTCCCCTCCTCAAAAACCTCCAGTGGCTACCAATCAAT
CTGCGCATCAGGCAGAAACTCCTCACCCTGGGCTTCAAGGCTGTCCATCACCTCGCCCCCTCCTACCTCACCTCCCTTCTCTCCTTCTAC
TGCCCAGCCCGCACCCTCTGCTCCTCCACCGCTAATCTCCTCACTGTACCTCGTTCTCGCCTGTCCCGCCATCGACCCCCGCCCCACGTC
ATCCCCCGGGCCTGGAATGCCCTCCCTCTGCCCATCCGCCAAGCTAGCTCTCTTCCTCCCTTCAAGGCCCTGCTGAGAGCTCACCTCCTC
CAGGAGGCCTTCCCAGACTGAGCCCCTTCCTTCCTCTCCCCCTCGTCCCCCTCTCCATCCCCCAATCTTACCTCCTTCCCTTCCCCACAG
CACCTGTATATTTGTATATATGGTTGTACATATTTATTACTCTATTTATTTATTTATTTATTTATTTATTTTACTTGTACATTTCT

ATCCTATTTATTTTATTTTGTTGGTATGTTTGGTTCTGTTCTCTGTCTCCCCCTTTTAGACTGTGAGCCCACTGTTGGGTAGGGACTGTC
TCTATGTGTTGCCAATTTGTACTTCCCTAGCGCTTAGTACAGTGCTCTGCACATAGTAAGTGCTCAATAAATACGATTGATTGATTGATT
GATTGATT

>Tacu_L2e#L2
GCCCGCCCGCCCGCCCGCCATCACACCGCGCGGCCTTGCTGCTCCCGGAAGGATCCTCTCCTCAGAGCGCCGCCACAGACGCCCGGGACT
CCGTTCCCGAGAGGCCCGCCCGCCATCACACTGCGTGGCCCCCGCTGCTCCCGGAAGTCCCCTCCTCCAAGCGCCCCCATAGACGCCCCC
TTCTTCCCCCTCCCACTTAAGCGACCTTCCTTAAAGTGCCCCCCATCCCCGCCCTTCCCGTCCGGTCCTGACTGACTCTCCCCCCCCCCA
CCCCGGGAAAAAGGCCCTTGTTAACACCATGTTACATTAACGACAACCTCATTCGCACCTCCTCAGCCCTCCCACGCTAGTTGGCTGTTC
GCTCCTCTCCCCAGGTATCTCTGCTCCCTGACCCCCCTTAACAGGCCCACCCTACTCCAGCACATAATAGTTTGTATCTGCTCTCTGTGA
CATCATTATCTGCCAATTTTATTACTCTGCCATAGCATATTGATTGTTTAACTACACCCTGTGATGCCACCGCCTACTTATATCATTGCT
ACTGTTTTATTGCTTCTGCATCTCAATTGTTAACCTCCCGCTGTGCTGTCACTGGCCCTGCTGACCTCACCATGAACTTTCAATTCCCTT
GCTCCCAACTGCCATTCCCATTCCTCACCTTCCCCTTCCCCTCTCCCACCCAGCTTTTTCCCTCCATCTCCCCCACCAAGACCACTACTT
CTCACCCCCTACCAAGGATCCCTCTGTACCAGCGCCAGTCCTCCACTCCTCCCTCCCGCCCCCCTCCCTCCATTTCTCTTCCGCACCCAC
CCAATCCCAGTTCTCCTTTCCCATCGCCACCCCTCTTCCCACTCTCCCTGCTCAGGCCCCCGCCAGCTCATCCCAATCCAAACCCTCCCC
ACCCCTCGCACCCTTCCCCCTCCCTCCCCTCCCTCGACAGCTGCTGCCAAGTGTGGCCTCTGGAACCCCCGCTCCGTTTTAAGTAAGATC
CCTTTCATCCTGGACCTATTTCTTTCCAGCTCTCTACTCCTCCTCGCCCTAACTGAAACATGGCTGTCTTAAGAAGACACGGTCTCTTCT
GCTGCTCTCTCCAGTGGAGGCCTCTTCTTCTCCCACTCCCCCAGACTCACCGGAAAAGGAGGAGGTGTCGGTTTCCTTCTCGCCCCCCAA
TGTCGCTTTCGCACTATCCCTACTCCCCCTTCCCTTTCCTTCCCTTCCTTTGAAGCCCACATTATTCGCCTCTACCACCCCCTCCAGATT
CTTGTAGCCGTCATCTACCGCCCTCCCGGCCCCACTTCCAACTTCTTTAACGACTTTGACCCCTTCCTCACCTTCCTTCTCTCCTTCTCC
ATGCCCACTCTGATCCTCGGAGACTTCAACATCCATATGGATATACCTGATGACTCCTCTGCCTCCCGCCTTCTATCTCTCCTCGACGCT
GCCAACCTCTTCCTCCACCCCACCTCACCCACTCACCAACTTGTTCATACCCTCGACCTCATCATCTCCTATCGCTGCAATCTGTCCACC
ATCACCAACTCTGAAATCCCTCTCTCTGATCATAATCGTCTCACCTGCCTCCTCACTCACACTCCTTTCCCCTGTAAATCCATATTACTC
CCTCACAGAGATCTCCGCTCTCTTGACCCCATCCATCTTTCGGAGCGCCTCACACCCCACCTCGCCGCCCTCTCCTCTCTACCGAGTCTT
GATGATCAGATTACTGCTCTCAACTCTACCCTTTCTACTCAGCTAGATTCACTCGCTCCCCTTTCCCTTCGACGCTCTCGTACCACTAAC
CCACAGCCCTGGATCACAGCCACTGTCCGCCTCCTTCGCTCTTATGCTCTAGCTGCCGAACGCTGCTGGCGAAAGTCTAAACACCATGCC
AACCTCGTTCACTACAAGTTTATCCTTTACTGCCTTAACTCAGCCCTCTCCTCTGCCAGACAAAACTATTTCTCCTACCTTATTGACACC
CATGCCCATCGCCCCCGCCAGCTCTTCCGTACATTCAACTCCCTTCTCATGCCCCCGGTTCCTCCCCCTCCCCCTTCCCTCACCCCCAAC
GATCTGGCCTCTTACTTCATTGACAAAATTAAATCCATCAGGTCCGACCTCCCCAAAGTCACTCCCCCGCCTTCCCCAACCCCCCGGCTC
TCAACACTCTCTGCTACTCTCCCATCCTTCCCAGCAGTATCCTCAGAGGAGCTCTCATCCCTCCTCTCAAGTGCTACTCCGGCCACCTGT
GCTTCTGACCCCATTCCCTCTCATCTCATGAAATCTCTCGCTCCATCCCTTCTCCCCTCCTTAACTTCCATCTTCAACCGCTCACTCTCC
ACTGGTTCCTTCCCCTCTGTCTTCAAACATGCCCATGTCTCTCCCATTCTAAAAAAACCCTCTCTTGACCCCACCTCACCTTCTAGTTAT
CGCCCCATATCCCTCCTACCTTTCCTTTCCAAACTCCTTGAACGTGTTGTCTACACGCGCTGCCTCGAATTCCTCAACAACATCACTCTC
CTCGACCCCCTCCAGTCTGGCTTCCGTCCCCTACATTTCACGGAAACTGCCCTCTCAAAGGTCACCAATGACCTCCTGCTTGCCAAATCC
AACGGCTCATACTCTATACGAATCCTCCTCGACCTCTCAGCTGCCTTCGACACTGTGGACCACCCCCTTCTCCTCAACACGCTATCTGAC
CTTGGCTTCACAGACTCCGTCCTCTCCTGGTTCTCCTCTTATCTCTCCGGTCGTTCTTTCTCAGTCCCTTTTGCAGGCTCCTCCTCCCCC
TCCCATCCCCTTACTGTGGGGGTTCCCCAATGTTCAGTGCTTGGTCCCCTTCTGTTCTCGATCTACACGCACTCCCTTGGTGACCTCATT
TGCTCCCACGGCTTCAACTACCATCTGTGCGCTGATGACATCCAGATCTACATCTCTGCCCCTGCTCTCTCCCCCTTTCTCCAGGCTCGC
ATCTCCTCCTGCCTTCAGGACATCTCCATCTGGATGTCTGCCCGCCATCTAAAACTCAAAATGTCCAAGACTGAACTCCTTGTCTTCCCT

CCCAGACCCTGCCCTCTCCCTGACTTTCCCATCTCTGTTGACGGCACTACCATCTTTCCCGTCTCACAAGCCCGCAACCTTGGTGTCATC
CTCGACTCCGCTCTCTCGTTCACCCCTCACATCCAAGCCGTCACCAAAACCCGCCGGTCTCAGCTCCGCAACATTGCCAAGATCCGCCCT
TTCCTCTCCATCCACACTGCTACCCTTCTCATTCAAGCTCTCATCCTATCCCGTCTGGACTACTGCATCAGCCTTCTCTCTGATCTCCCA
TCCTCGTGTTTTTCCCCACTTCAATCCATACTTCATGCTGCTGCCCGGATTGTCTTTGTCCAGAAACGCTCTGGGCATGTTACTCCCCTC
CTCAAAAATCTCCAGTGGCTACCAATCAATCTGCGCATCAGGCAGAAACTCCTCACCCTCGGCTTCAGGGCTCTCCATCACCTCGCCCCC
TCCTACCTCCCCTCCCCTCTCTCCTTCTACAGCCCACCCCGCACCCTCCGCTCCTCTGTCGCTAATCTCCTCACCGTACCTCGTTCTCGC
CTGTCCCGCCATCGACCCCCCAGACCACGTCATCCCCCGGGCCTGGAATGCCCTCCCTCTGCCCATCGGCCAAGCTAGCTCTCTTCCTCC
CTTCAAGGCCCTACTGAGAGCTCACCTACTCCAGGAGGTCTTCCCAGACTGAGCCCCTTCCTTCCTCTCCCACTCGTGCCCCTCTCCATC
CCCCAATCTTACCTTCTTCCCTTCCCCACCTCACCTGTATATATGTATATATGTTTGTACATATTTGATACTCTATTTATTTATTTATGT
ACTTTACTTGTTCATATCTATTCTATTTATTTTATTTTGTTAGTATATTTGGTTTTGTTCCCTGTCTCCCTCTCTTAGACTGTGAGCCCA
CTGTTGTGTAGGGACTGTCTCTATATGTTGCCAACTTGCACTTCCCAAGTGCTTAGTACTCTGCAAACAGAAAGCGCTCAATAAATACGA
TTGATGATGATGATGATGATGATGATGATG

E.2 LTR retrotransposons

```
>Tacu_ERV1#ERV
ACTGGCGCCCAACGTGGGGCTCGAGTACCCCCATAGGGGAAGATCAGGGAGAGTGCCCCGTAGGCCAGGTAGCTAAAGGTCAGGCGGGAA
GATGGGGACAGTACGATCGCTGCCCATATATAAGCCAGAAGATAAGGAATTGTACACCCGACACTGTTACAAGATATTGAAAAGTAAAGG
GGTAAAGATTGAGCTGAAAACAATAAGGGAATTTATAGGGAAGGTAGCTATGACCTCCCCGTGGATACTTGATTCCGGGATCACGGAGGA
GAGATGGGACGTTATTGGGGAACAGATGACGGCCTATGAAGACTCCCACCCGGGGGAGTTAAAGGACGTGGACTTTCTTATACACGGTAT
CCTCCGTGCGGCCTTTCAAGGGCCAGAGAGACTCGTTCGTAAGTTTGACGCGACCTGCCAGACTGACAAGGAGGAAGCGGGGCAGGAAGG
GGAGAACAGCCGACAACCTGAGGAGACTGGTCAGGCCGAGGCAGAGACACCCCCCCCACGGGCCCCGAGTTACCACCGCATCTATCCTGA
TCTCGGGCCTTACACCCAGAGTGAAACGCCCAACGCCGAGCGGGAAAGAGGAGATCAGGCTACACAGACGGTGGTTGAGAATGAGATAGG
GGATACGCGGGAACAGTTCAGGCAGATGGATGTGGGAAGCAAGAAACCGATCGTTAAAAGTGGGGAGGGGACGAGAGCGGTTTCAGCCCT
ACAGCGAGCTTTGGAAGAGGCGGTGGCGAGGGGAGAGGACGTCACGGGATGGGAGGTATTCCCCGTGATAGAAAGACCAGACGGGGGACG
AGGTTTCGCCCCGATACCTTGGGCGAAACTCAAAGAGTTGAAGGCGGCGTGTGTCGCCTATGGTCCCAGCTCCCCTTATGTGAGCCAGCT
CCTGGACACTATGTCCCTGGAAAGCGTTTTGACCCCGAATGATTGGAAATCCCTTGCCCGCGGGTGTTTGGATCCCGGACAGGGCCTTAT
ATGGATGTCTGAGTTTACCGCCGCTGCGAAAGAACTAATATGTAGACGGGGTTTCCCGAACCCCGCCGAGGCTTTTGCAGCTGTGACCGG
GACGGGACGATTCGAGACTCCTGAGATGCAGGTCAACTACGAGCCCGAGACGTATATGGTGATTGCCAGGGTTGCCCTAACTGCCTGGCA
AAAGGTTCCCGAGAAAGGGGACCATCGCTCCCCCTTAACGCAGATCAGACAGCGCCCGGACGAGGCGATTCAAGATTTCGTTTCACGCAT
GCAGTCCGCGGTCACCCGTATTATAGGGGATCGGGACGGCCCCGAAATTGTATTGAAGCAGATGATCAGAGAGAACGCAAATAGTGCCTG
CAGGAAAGCATTGGCAGGGCTGCCCAGAGAGGCCACATTAGGAGACATCCTGCAAAGATGCGAAGGGGTTGGAGGAGAAGAGTACAAGGC
GCAGATGCTTGCGGGGGCGATAATGAAAGGATTATCAGGAGTCGGGGAAAGGGGGCGGCAGTGCTTTCGGTGCGGACGGATGGGGCACCT
GATGGCTCAATGCCGAGCTCAGGACAAAAGTTGCCCCCCATCACAGCCGAGAAAGGGTGTGACTTGCTTTGAATGTGGGAAGCACGGGCA
TTATGCGAAACAATGCCGCTCGAGGCGGAGACCCCGACCGGCGTCGGGAAACGGGTGGAGGGGCCCCGCGCGGGCCCCGAATCACGCGTT
CCCCGTGTCAGCCGCGGAAGAGAGCCCGAAAAAGGTTTCTCTCATAGAGGAGTACCAGAGGCTTCGCTCCCGGGAGCCACTTGGGCTGAA
GTCCGATGCCTGGCCCCAGTGGGAATCCGACCCGGGGACACTGTAAACGTCCCAATACAGCCCCTCCCCTGGAGGGCCCTAGTGGTGGGG
CTTCGGGCAAGAGCGGCAGGGGTCGTTCATTCATCCGAACGGGGGGAAGGGCCCGGGAAGATTCCCCTCACCAACCATCATCCGTATGGT
ATATACCTGGGATCGGGAATGGTCATTGCTCGCGCTACGCCCCTTGACCCCCCGCCCCCTGATCCCCCGCCTCCAGCTATTGGAATAATA
CAGGAAATAACCCTTGATAAACCGTGGCGGACCCTCTTAGTTGAGGGAAAGCCCCTTAAAGGGCTCCTGGACACCGGGGCCGATCGCTCC
GTTATTCATGATTGCTGTTGGGCAGCGGAGTGGCCACTAGCAAACCACTCGATGGGGGTGCAAGGCGTGGGGGGACTGCAAGCCGCGAGA
GAGGCGGGGCGTTCGTTGATTTGGTCTTGCCGGGGAAGGCAAGGTGCGTTCGTCCCGCTGTGCGTTCAAGGGCTCCGTATGAATCTGTGG
GGTAGAGATGTTCTGCAGGGGCTGGGAGCCCACCTTATAGATGAAGCTGATCCTTTTTAGGTGGGGCCACTGGGTTCCGGGGCTTGTCGA
CACCCCCACTGGTATGGCTTCCACACCCACCGGTATGGGTGGACAAGTGGCCCCTGACCAAGGACAAGCTGCAGGCGCTGAGGGAACTAG
TTGCGCTCCAATTTGCACAGGGGCACCTAGAGGAATCGTTCAGCCCCTGGAACGCCCCCGTATTCGTTATAAAAAAGAAAGCCGCGGGGA
AGTGGCGCTTCCTCATGGATTTGAGAAAAATCAATGCGCTCATTGTGCCAATGGGACCCCTGCAACCCGGATTGCCCTCCCCTAACATGA
TTCCAAAAACTCACCAGATCCGTGTCATAGACATTAAGGACTGCTTTTACAGCATCCCGTTACACCCTGACGACAGGGTGAAGTTTGCTT
TTACTGTCCCGAGCCCGAATTTCGCAGAACCAGCGATCCGGTACCAGTGGAAGGTGCTGCCACAGGGCATGTCTTGTAGTCCCACCATAT
GCCAGTGGTTCGTGGGGCGGATACTCGCCCCTTTCCGCAAGGAATACCCGGGGGCGACGATCGTCCATTACATGGACGACATCCTTCTGG
GTATGCCCGACCAAGGGCGGGTACAGACGCTCACCCGGCGAGTGGTGGCAGCCCTCGCAGCCCAGGGTTTATTCGTAGCACCTGAGAAGG
TACAAGAATCAGCCCCGTACACGTACCTTGGGTTTGACGTCACCGAGACCCAGGTGACTCAAAGACCACCCCAAATTGACCCCCGAAAAT
ATGTCACGTTAAACGATATGCAGGGTTTAGTAGGGAGGATTCAATGGATGCGCGCGAGAACGCCCATCCCCTCCGCCCTCATGCAACCCC
```

TGTACGATCTCCTAAAAGGAGACCCCAATCTTAAATCGCACCGCGAATGGACGGAGTCCGCGAAAAGCTCACTCCGAGAAATCACACAGC
GGTTGGTGGGTAGCCATACTTGTCGAGCTGAACCCCACCTCCCCATGGAGGTCACGATATTTCGGGAGGGTAGCCTTTTCGTGGCTGTAC
ACCAAGGAGCCTTGATTCTCGAATGGTGTTATCCTCGAAACCCCTCCCGTGTTCTCCCAAAGGAGACTGAGCTTCTCGGTCGCTTTTGCC
AGAACGCCATACAGAGGGTAGTGGCTCTCTCAGCCACTTACCCTATAGTCCATGTCGGGATAGCCATGGGAGACCTTGAGGCGATAGCCA
GGGACAGCTTCCTGTGGGCCGTGCTCCTACAGCAGGCCACCTTTACGGAACGCTCCCCGTTGACTCTTAGCCAATTATACGGGGGATCGG
ACATTCTGATCCCTCGGGTGATCTCCGATACTCCGGTCGGGGGAGACGATGTCTTTACGGACGCCACCAAGGAACACCGGGCGGCAGTTT
TCAATCAGACCACCGGTGCCTTGTCGGTGCTCGACACACCATACGGCTCGACTCAGCGAAATGAACTTTTCGCTATCATATGGGCAATGA
CTACCTACCCCCAGGCGATCAACATTATCTCGGATAGCCTGTATGCCGTTCATCTAGCCCGGAGAATCGAAACCTCCATACTTTTCGACA
GGCACTCGGAGATTGGGAACATGATCTCTCAGCTTCAGGCTGCCGTAGCGGCTAGGGGATGTAAGGTATACCTCGTGCACGTCCGTTCCC
ACACGGACGGACAAGGGCCGATCTTTGAAGGCAACCGCACTGTGGATGCCAGCCTACACCCCACAGGGCCGTCGCTAATGGGACTGGATG
CTGCAGCTGCGGCGCATAGGGAGTTCCATCTCCCGGCCACCTCGCTCTGTCGGCTGTATGGGGTCACTAGAGAGAAAGCCCGGTCTATTG
TCCGGCGCTGTACTCGCTGTATTCCGTTCACCCCACGGCCGGCAGGGTCGACGGGTGTGAACCCCCGGGGGCTCACGCCAAACGAGCTGT
GGCAAATGGATGTCACCCATTGGGGGACCTCCACGATTCATATGACCATTGACACCTTCTCCGGATTCATGCTGGCGACACGCCAAGCAG
GAGAAGCCGCAAAACATGTGCAAAACCATTTGTACCATTGCTTTGCCACTATAGGCACACCCCAGGAGATAAAGACGGATAATGGCCCCT
GCTATGTCTCCAAAGCCATGTCCCTCTTTTTCTCCTCCTTTGGCATCTCCCATATTACCGGGATACCTTACAACCCCAATGGTCAAGGTA
TAGTGGAAAGGACAAACAGGACGCTGAAAACTCTCCTGAAGAAGCAGGGGGTGGGGAAGCGGGTCACGCAGTACGCCCTAGACAAGGCAA
CGTACACCCATAACTTTCTATCTGTTGATCGGGAGACGGGACTCTCCCCCGCCATGCGGCAACTCTGCCACAGGTCCGCGACCGTGAAAC
CCCGGCGCCACCACCCACACGCCCGCACTATGGGGCGAGCGATGTGGCGTACTGTGGAGGGCGATTGGCGCGGTCCTGATCCGGTACTAA
TTCGGGGTGGAGGATATGCTTGTATCTCCACAGGTGACGGCCCCGTTTGGATCTCCTCGCGACATCTCCGCCTCGTCGAGGAAGACGATG
CCCAGCCGCAGGGGAAGTCCCGGCGCCCTGATTCCCCTCCTCCTCCCCCTCCTTCCCCTTCCCTTTCTTCTTCCCGGGGGCGCGGTGGGG
AGGGAGGCCGGGACGCGGATTAACTGGACTGCTCCGGTAAATAGCCGCCTGAAAGGTGTTGAGGAGCAGAGCAGGGTTTGGAGGTGGGCC
CTCCTCGCCAATGCGCCTTTGTTGCGCATGGTCTCCTGGAGGGAGGACTCCCCCCAACTCTCATTTCGAGGGAACGTGACCAAGCTCACA
GGCCAGCACCGGGGTCAGTCCCCGGGCACTGACTAGCGTGACATCCGTATTCGGGGCGCTATCTTGTGCAGCACTGGTCCCCCCATATGC
TGGTACACCGGCAATGGGGCTTTCTCGAACCAGCCCGAGGTGACCCGATCTTGCTTGCGGTTGCTGAATTTTACTTCTACACTGGAAGTC
CACATACCCCAAGGTCAGGGGCATCCCACACACCGAAACCTTACCCTGCTGGGGATCCACGGCACACTTCAGTGCACCCCCGGCGGCAAA
AAATCCAGCGAGGGAAGCGGGGGGTATTTTGGCCCTATCCGCAATGCGAGCTCGCTCCCCGGTCCCGGGAGAATGTGCGCCGAGGGTCGG
GGTGGGGTTTGGATGCCTCTAACTGACTGTTCCACGGGCATATTCCGCTGTATTCCCCATAATGATGACGGTACCCTCCTGTTCTCCTGG
GATAACTACGTTGGGGACCGCACATATACAGAGGCGGCCGCACCACGGCTTTACGTGCTACCGTGGGGGATGCACCACCCCGGTCTCTGG
AAAGCCTTTACTTCCCTGCGCTCCCCCACCGCAACTGTTAGGGTCCGCTCGTGCGCCTGCCCTGACGGGGCGACGTGCCAGGATTACCCT
TTCTCGGTATATCGTATTATTTTCCGTGAGGGTCTGAACGTTTCCCTGCACTTGCAGGGCAATTGTGCCGACCATGATTATGCCTTTCAA
GGGAATGCCCTTATCCCTGCCCCTTACGCGCTTATAGCCTGCCCAAGTACTGACCCCAGCTGGAACCTTACTGCCTCCCGGTCAGCTTAT
GGCGACTTTTCCTTGTCCACCACGGGTTCCTCTCAATGCGTCCTTATGACTCGGGCGCGACCCATGACCACGGAGGGTGCGGCAGGTTAC
GCTGTTTGGCTGGTCCGGCGGCACCGCTACACCCTTATACCCGTTAAATCCGCTGACCCTTGGTTTTCCCAGGCCTCCGACCATCGATGG
CACACCCTAGCGACCCGCTTGCGCGGCTTGCGCACCCGCAGGGAATTATTTGAGGCCATGCTGGGTATCGCGAACCTCGCGTTTTCTGCC
TTCCAGGAATGGCAAATCCTGAATCTGTACGATCGGACTGGTTTCCTCGCGTCTCAGCTCTCGGCCTTCATGCACTCCACCCAACTGGCG
TGGGGGGTTCAGCGTCACCTGGACGCTTCCCTGCAGCAGGAGCTGATCGGCCTTAAGCACGTGATCGAGACGCTCGGCGACGAGGTTCAC
CTCCTGTCCCTACGGCAGGCGGTGAAATGCGACTACCGGTATCGGCACGTCTGCGTGCTTCCGCTCCGCTGCAACGCCACCAACTCTTCG
TTCCCTGCCTCCTGGGACGGAGTAAAGAAACACCTGGAGGGACTTTTCCTGGCCGCCAACGTGACTGGCGAGTTACAGGAACTGTGCCGC
CTTATTGACGAGCTCAACGCCCAGTCGCTGCGCTTCCCCAAGGCTTGGGACGACCTAGGCCGCGGGGAGGACGCGGTGGAGGAACTTTGG

GACTGGATTAAACGCGTGATGCCCGGGGGATTGGCCCCTTGGCTGGTCTCGGTCCTTGCCACCACTGGGGGGGGGGGGGGTTGCTTTTGC
TCATGTGCCTCCCGTGCCTCCTTCGTCTCTTTTGCTCCATTGTTGCCCGGGTGGTCCGCGAGCCTCAGGTCGAGATGCTCCCATTGCGCG
GACGCCCTTAGCGGTCCCCGCCATAACGGGAGGAGGA

>Tacu_ERV2#ERV
GTAACTTCTTTGGCGACCACGAAGGGACCCTCGTGCTGTGATCTGGATAGCAACAATCTCCGTCAGACTTTGTAAGTAAGGATCCGCTTT
TAGGGTCTGAGTCTCCCCTTCTCTCTTTCTGGGTTGTTAATTTGGAGCGCCCGGGTGAGGTTCCGGGTTAGCGGTTTCTTCCTCCTCCAT
TTACACGGTGGTGGATGGTTTCTCCCCCCTTTGGTTAGAAGGGGCGAGGGGGTTTCTCCTCCTGGGTTAGGAGGTGGTTTGTTTACCTCA
GAAGATGTTCTGAGGGGGCCTAGGAGACGTCCTAGGTGTTCTTTCTGGGTTGAGCGGAAAGGGGGTTCCGGTTGTTGGAACCATGGGCAC
GAGTCTCTCTAAGACTAGTTGGTGCTTGGCTAAAGTTCTCCAATCATGGGATGAGGGTCACTTGCCTGTAACTAAAGGAATGTCAAAAAA
CAACTTGTAGGGTGTCACAAGGCCTGGACTAAGTTCTGGGACCAACAGGGTTATTCTTTCTGGCCCGAGGGAGGGTCTTTTGATCCTCAA
ACCCTTCACAAACTCCGGGAGGTTTTATCTTGTTCCCGGTCTGATGAAATGTGCTATTGGTATTGTTGGGCTGAAGTTAGGAAGAATCAG
TCCTCTCAGACTAAAGTAGCAACACTGTCACCTGTGGCAGAAAAAGCTGCGTTGTCTTCAGTCCCGAGTCCGCAGTCCCACTCAGCGGAC
CTGATAGGTACCCCAGTTCCTATCCCCCCCTGCTTATTTACCTCCGGTACCACTTCTAGCCCCTGTTCGTGAGAACTACCAGCAGATCCC
AGGGGCAGGCGGGAACATCGTAGTAGCCGGAATAAGGGCTCATGTCCCTTTCAAGCCCCAAACCCTTATAACCTGGCAGCAGAATACTCC
ACCCTTTCAAAAAAATCCTCAGGAAGTGTATCGAAGGGTCCTGGGAATTTTTAGAACCCATTTTCCTACTTGGGCGGATGTAGAAGAGCT
GCTTGAATTGTTTTTCACTTCTGATGAGCGCTCCTGTCTCAGAGCCTTGACTGTAAAGTTAGAGAACAATGGTCAGCCTCCTGTGGCTTG
GCCAACAATTGAGCCTAATTGGAATTTAAATGTGGAAGCAGATTGGAGTCGCCTTAATCTGGCGCGGCAGACTCCCCTGGCAGCCATCCA
AGAATTGGGCCGTAAACCCCCTAATTGGGACTCCTTCCATGGGATGATCCAGAGGGAAGGGGCAACCTCCGATGACTTTTTCACTCGACT
TAGCGAGGGAGCAGAGACGCTTGTACAGTTGAATCTGAATAATGAAAGAGACGCTAGAGCCCTAGCTACTGCCTTTGTGAAACAGGGGAG
GTTCGCCAATTATTTTGATAAGTATATTCCCGGGTGGGCAGCAAAACCTCTGTCAGAGCTGCGAGAGATTGCGCGTCATATACAAAATAT
GAAACCCACTAATGCTGCAAAGGTCCTGCTTATGGGCAACTCCTACTCCCTGGGGCGTTGCTTTCATTGTCAACAAGAGGGTCATTTTAA
GAGAGATTGTCCTAAATTGCGTGGAAATAAGCGGGGTAGATACCCACGGGTTTTTCCCCCAGGTTGCCCGAACAGTACTAGACCCAATCA
GGGTCTAGCCCCGAGGAATGGACCTCGTATGCCCCCAACCCAAAACACAGGTCCCAACTTTGGCCCCTTTGGACCTTGGGACCCAGAGGG
ACCCCAATGACTCTCAGTTCCCCTCTCATTTCCCCCAAATATGCGATCCCCTGGGCATGAGTATCCCGATGTGCATTCCCTACTGGCTGG
ATGCTCCCTGGGGATTGATTTCCCCGATCTGAGCCTCTATGATGAACCGCGGGTAAAGCCATATGGCATGGCTCCCGTGGTGACTGATGA
CCGGCTCGTCACCCGTTCCCCTGAGTCGTTCTGTGAATTGCCCATTAATACAGAGTCCTATAAATTGCTAGTTGATACGGGTGCAACCTA
TTCAGCGCTCCCTTCCTTTCCTCCTGGAGCCCGAGGAAAGGAAAGGATAAGTATAGTAGGGGTAGGTGGGAAAGTTCAAACTGTTTCTAG
AACACAGCCCCTACAGTGCTGGTTACAGGGGATTTACATTCACCACCCGTTCCTAATCATTCCTTCCACTCCATCTCCTTTAATGGGACG
AGATCTACTATGCCGTTTGAATGTGACTTTGGTTTGTGATCCGGAAGGAATTGCCTTGACTCTAGGAAGCCTCCTGGGTAAACCTGATCC
AGTGCGCACTCCACCTGAACAGCTGAGTTCCCTCATGGGACTCCTGGATGCATCTCCCTGCACAGATGATCCCACTCTAAGGACATTGCC
CCTGACACTGTGGAGTCGTGGTCCGTCAGATGTAGGTTTTCTTCATCGGGCAGAGCCAGTTACAATTACAGTCCGAGAGGACCGGGCACC
TCCCCAAGTCAGACAGTATCCAATGTCTGCAGAAGGTAAAGCGGGTTTAGAGCCCTTGATATCTGCCTTTTTGCATGAGGGAATTTTAGT
CGAATCTCGGTCTCCTTGTAATACTCCTGTGTTGCCTGTGCGAAAACCAGGCTCCCCTCTGTCCTATCGGTTAGTCCAGGACTTGCGTGC
AATTAACTCTTACGTTTTGCCTATGCATGCTGTGGTTCCTAGTCCTACGGCAGTAGTTAGCTCTGTGAGCCCCGAAGCAACCTGTTTTAC
TTGTATTGATCTAACTGGTGCTTTCTTTACTATCCCGGTGTCCTGTGAGAGCCAGTACCATTTTGCCTTTACCTGGGAAGGTCGTCAACT
AACTTGGACGAGGCTTCCCCAGGGTTTTGTTCACTCCCCAACTATCTTTTTGAGAGAATTGTGCAGAGCCTTATCTCCCCTCATATTACC
CGAAGGGATCAGTATGTTCCAGTATATGGATGATATCCTTATTGCTGGGGAAATGAAGGATCTGTGTCGGTCAGGTTCCCTCCAGGTCCT
AGCATGTTTGCATCAGCTTGGTCTTAAAGTTAGTCGTGAGAAACTGCAGTGGTGTCAGCCCCAGGTAAAATATCTTGGGTTCATGTTGAG

GGCGGGAGAAAGAGCAATCGCCCCCAAGCGAGCAAGCCTGATTCAATGCATGCCTCGCCCCACTACTAAGCGGGCCCTGAGAGGTTTTCT
AGGTGCGGCTGGGTTTTGTTGAGCTTGGATTCCAGAATTTGGGTTGTTAACCAGACCCCTTTTTGAGCTCCTAAAAATGACTTCCCCAGA
GCCCCTGGACTGGACTCCTAAGGCTGTAGAAGCCTTTCAGACACTAAAATCATGCCTGAGTTCAGCCCCTGCCCTGGGACTCCCTGACTA
CAGTAAGCCGTTCTATTTGTACATTCATGAACGCCGCGGAGTGGCCTCAGGTGTTTTGTGCCAAACTCTCGGTCCTCACTATCGACCAGT
AGCCTATTATTCAGGTGTCCTTGATCCTGTCATTCTAGGTCAGATACCCTGTATCCGCTGCATAGCTGCGGTGGTAATGCTGCTTGAAAA
GTCTCAGGATATCCTCTTGGGACATCCCGTTGTTGTTCGGACTCCGCATGAAGTTGTGCGCCTCCTTCGTGGAGCAGCCACCCAAGCCTG
ATCTGTGGCACACCTGACTAAATATGAGGTTACTTTATTGGAAAATCCGCAGGTACATGTGGAGCGCTGTGGCTCCTTGAATCCAGCTAC
CCTGCTTGAGTCCTCTCCCATGACTGACCAAATTCACGATTGTGACGAAGTAGTTCGGGAAACTGTCGCTCCTCGATCTGACCTTCGCGA
TGAGCCTTTGTCCTCTTCTGATCTAACTTTGTTTACAGATGGCTCCTCCTTTATGGACCAAGGGGTGCGCCTCACTGGCGCTGCTGTTGT
TACCCTAACCTCTGTGCTGTGGACCGGCTCACTCCCAGCTTCTTGCAGTGCACAGGCAGCTGAGCTTGAAGCCCTTACACAGGCATTATT
ACTCGGTAAAGGGAAGCGTGTGACCATCTATACAGACTCTATGTATGCCTTTGGGGTCTGTCATGCTACAGGTACCCTGTGGCAAAGTCG
GGGGTTCCTCACTTCAGCGGGTCGTCAGGTTGCTAATGGGGATCGTATAGAATGCCTATTACAGGCCCTCTTGCTCCCAGCCGAGGTGGC
TGTCGTGCATGTGCGTGCCCACACCCAAGGAAGTGATTCCCCGAGTTTGGGCAATCGCCTAGCCGATGAGGCCTCTCGTTGGGCAGCTCG
GTATGCCCCCTTTGTATGTGTAGCTCCTGTTATGACCCTGGCCCCCACGTCTATCCCTGGCACTCTGTTGCCCTCCCTGATTACTGATAA
GGAAAGAGCTGAGTGGCCTCGGCAATATGAGGCACTGGAACGGGACGGACTCTTTTGGGTCCCTGATGGATGCCCTATTTTGCCTGCTGC
TGCACTGCGACCAGTTTGTCTGGGTCTTCACCAGGAGACTCACTTTGGTGCAGATGCCCTGGCAGCCACAGTTCTTGGAGTTTGGTTTGC
CCCAGGTATCCATCCTGTTGCCAAACGTGTTACAGCCAGTTGTGCTACTTGCCAGAGCTTCAATGCTCACCCTGGTCGTCAGGTTCCCTT
GGGAGGACGCCCTTGGGTCTATTCCCCTTTTGAGGGTTTGCAGGTTGATTTTGTGACACTTCCAAGGGCTGGTCGTTTCCGATATTTACT
GGTCATTGTTGATCACCTGACTGGGTGGATTGAGGCTTTCCCTTCCTCCCGAGCAACCGCTCTTACAGTGGTGAAATGTCTCTTGCGGGA
AATCATTCCCAGGTTTGGACTACCTGCTGTTATTGATTCTGATCAGGGATCACATTTCACTGAGTCTGTTCTATCCGAAATCTATCGATC
ACTTGGAATCAAACGTTCTTTACATGCACCCAACCATCCCCAGAGTTCAGGGAAATTGGAACATGCTAATCGTGAGTTGAAAACCCTTTT
GGGAAAGCTTTGTATGGAAACTCTTCTCAAATGGCCCGAGGTTCTACCTCTAGCCCTTTTTCATCTACGATGTCGCCCCCAGGCTCCTTT
AAATATATCTCCTTATGAGATGCTATTTGGTCATTCCCCGTATTGGGGAAGCCTTTCAGATCCCCGACCTCCTCCAACCCTACAGGTGGG
CGATGATGCCCTGCGAGCCTATGTTTTGTCTCTCCAGGCAATTCTAGCTGAGCTGCAGGCTGCTGGAATCCTTCGTCAGCGCATCCCTCT
CACGGATCATCTGCATCCATTCATCACTGGAGATTGGGTTTGGGTTAAGCGCATGGTTGGCAACGCTTCACTTAAGCCATCCTGGGAGGG
ACCCTACCAGGTTCTCTTGTCTTCCTTTTCTGCAGTTCGTGTTGCTGAGCGCCCCTCATGGATCCACCACTCCCAGGTGAAGCCAGCCGT
TTTTGATGCTTCGCCTCTTGATTGTACTCCCCCATCGGTTTCTGCTGACTCTCAGGATTCTGAGCCACTTGCTGTCCCAGCTTTACGGCT
GCGTCGAACTTCTACCGGATGGGCCTCGGTCCTGTGAGATTTTGTCTGGTTGCTTGTTTGTGTGTCTTGCTGCTGGCTAGTTTTGGTGTG
TGCCTTTTGGGAAGGACAGACTGATTGATGTTTCTGCCCCAGTATCCCTCCTGAAGCTTCGAGTTTCCTTCAGCTACAGCATTTTGGCTT
TTGGGCTCTGTTCATTCAGCTTCCTGTGTTTCTCTCACTTCCCTTTTTCTAAGACCAGTTAACATGAGGCTCCTCTGCCTCCTTTGCCTC
TTCCTCGCCTTGTCTTTGGGGGTTAGTGCTTGGCAGGATAATCTAGCAGTGACACTCGCGGGTTCCCTGAATGTCTCTTTTTTGACTTCT
GGGAATACCACCTCTTGCTGGATCTGTTTACATGTTCCTAGTTCCCACCGCTCGGGCTTCCTTTTACTCGGGGTCCCCTTGGCCCCATGC
CGTACAGATGAGAGTTGCTTCTATGATTTTTTTAAGGGCTCCTGGTATCCTTGGTATCCAGGTGACACACGATATATATCTTTCCGTTCC
CGTAACAGAGAGCCTGCGATCCGGGTGTCCTCTCGGGTATATGCTGAAGTGTGCTTCGTCTTCTTTAATTCTTCTTCTTTAAGCCTTCTC
AGCTTTCTTCCTCCTATCCACTGTGCAACTCCACTCTTTACTTTTGCAGGGGTGTAACACACTATTCTGTTACTAATCCTCAAACCCCTT
TCTGGACATTTGCTTCTAACTGTTCTGCCCGGACCTTGATCTCTGGGAGAGTTTCTAAACAGTTTTGGAATATTACTACCTTGCCACGCA
TACTTAATTTCACTGGAAAGCTGGATGTGGGCATACGATTTGTTCCCGGTATAGGGCCCAATTCCGCTGTACAACTCCTCGGTGGCCTAC
TGAAACAAGCCTTTGGCCGCAACCAGAGCCTCTGCTTGCCAGAGGGCTGGTACTTTCTTTGTGGCCGTTGGGCCTTCCCTTGTCTTAATT
TTTCCTCCCCTTGGGAGGGTGCGTGTACTGTGGGTGCGGTTTCCCCTTTTCTTTCCATTTTTCAGAACCTCTCGCTACCCCATACCCGTA

ACATGGGCTTTCACTGCTCTCGGCGCACTGCCCCTCCGTGGGCGCTGGGGGACTCCTTTTCCTCAGGTCGTGCCTTCCTGCATGGCCTCC
TCCCTGTTGTAAGTATAGAGATTCTCGAGCACTCCTTATTCAATCTCTCGGCTCAATTTAGACTTTTTGCTAATGCCACTGTGGATGCCC
TTACAGCCCTGAAGGAAGAGATCCACTCGGTCTCTCAAGTCATTATCCAGAACCGATTGGCGCTGGACATCCTCTTGGCCAACCAAGGAA
GAGTATGTGCTCTGATCAATCAATCCTGTTGCTTTTACCAAGACCAGTCGGGCCATATTGAAACTGATCTCTCCATCCTATGAGGAGTCG
TGCAGTCCCTCCCGTCCCAGTCTGCCCCCAGTTTGGACTCTTGGTGGTCATGGCTGGATTTTCCATCCTGGTCTGGCTGGGTTGGGTACC
TGACCAAAGGATTTCTTGCACTCCTTCTAACGTTACTGTGCATCTTGTTTTTGGTGTGCTGTGGTTTTCAGCTCCTGAAGGCCTGCCTCC
GACGTGCTTCATCTACCCTGATGCTTCTACAAACCCTGACCACACCCCTGTCCAAAGCTGAAGTCACGGACATGCAGCGCCATCTGGACG
CTGTGTTTCGCTGAGGAGGGGGATTGTTGTGTGTGCTCAGTTGTGCCTGAATTTTTCTGTTGGATTTTTCTGTCTTTTCTTTTTGCCTGG
ATTTTTGTCTTTTCCTTGTCCTATGCTTTAGGCGCCGGCTTTGCAGAAGACCTTTGTTATCCCGAGGAGGGTCGTATGTGATCCTCCCGG
TGGCCATGGGCCTCGGAAGGATCAAAGGGGGGA

>Tacu_ERV3a#ERV
GTGGCGCCCGAACAGGGACAGGGACCTGAACCCTGGACCCTCAGACTCAGCACCACAACCCTCACGAAGCAACTCATCACCGCGCAGTGA
CTCCCCAAAGAACGGAAGAGGTGAGGGCATTTAAGTTTACTTTCGTTTTCAAACACCAGGCTTACAGCAGGGAAGAAAATGGGACAAATT
CAATTCCGTTCTCGGAATCATTATATCCAATTACTTAAGGATGTTGTAAAGGATGTTGTAAAGGTTGATTATCACCACTGCCCAGATTGA
ACAACTTTTGGAATGCGTGGAAAGATGCTGCCCTTGGTTTCCAAAGGAAGGAATAATGGATTTAGAACGATGGGAAATAGTGGGAAGGCA
GATGAGTTCCAGACTACGAGAAGAAGGACCCCAGGCTTTTCCTATCGCTGCGTTCTCCATATGGAACGTATTAATGATTTGCCTTACTCC
TGAACACAAAATTCCCTTGAGGGAAAAGTTGGCGGGGATAGATGTGGCCTCCGGCCCGGACAAAGCTATGATTTCTCCGGACCCCCCTTT
GCCTCCCCCGCCTGAATCTGCATTGATAGACTGTAATCCCGAGCAGTGGCCGAAAGGAAAGTGGAATCCTTTCACTGAGCCAGCTCCTTC
CGCTCCACCCCCTCCACCACCCCCTCCATGGGTCCTGAAACCAGCACAAACACCTCTGCAGCGAGCGATGGGGGCTGCTATGGAAGGGGG
AGAGGAGATACCTATTGAACCGTTGATGGGTTTCCCTGTGATGGCGCAACCTGACCCTGCTAATCCTGGTCAGTTAATACGTACCCACGA
ACTCTTGAATTTCAAATTATTAAAAGAAGTGAAACAGGGTGCTGTGACATATGGACTCACGGCTCCTTATGTGCTGGCACTCATAGAAAG
TATTGCTTCCTCAATCCTTTGCCCGTGTGATTGGCAAACCTTAGCCGAGACTGTCCTGGATGGCGGAGACTATTTGCTTTGGAAGGCAGA
ATTTTTTGACTGGTGCGCGGAGCAAGCGCGTCGCAATGATCGCGCCCAACCTCCAGTTCAAATCACTTTCGAGATGCTTACAGGGACTGG
GCCTTTTGCGGACACTGCCAAGCAATTAACTTACACTCTGCAAGCATATCAATAAATAATGATAGCGGGGAAAGCTGCATGGCAAAAGCT
CCCGGCGAAAGGAGAACAAAACCCCAGTTTTACAAAAATTTTACAAGGACCGTCAGAACCATTTGAACAATTTGTCAGCCGCCTGACGAT
GGCTATCCAGCGCAGCGTCAGTGATAATGAGGCGGGAGTTATTTTTTGAGACAGCTGGCTTATGACAATGCTAACCAAGACTGCAAGCG
TGCCTTAGCTGGGGGCAAACGGACCTTGTCCATCACTGACATGGTACGCCGATGTCAGGATGTAGGGACAGCCTCTTTCAATGCTAATCG
AATGGCCGCAGCAATGCAAAGACAAACTAAGGCAGACATGGTGTGATTTCACTGTGGAAAGAGCGGACATTTGGCGAAGGATTGTCGTAG
CCCACGACAGAACTTTTCCAGACCACTACCAGCCTCAGTATGCCCAAAATGCAAAAAAGGACGCCACTGGGCTTCTGAATGCCACTCGAT
GCTCTCAAATAATGGCAGCCCTGTGCCGGGAAACGGACGCAGGGGCCGGCCCCAGGCCCCAACAACAATAGGGGCAGCGATCTCTCTGCA
ACACGCAGGTGTTCCAAAGGAACAGGCTCTGACTTCTCCCGACAACTGTCCACCATCGGCGAATTGCAGCGAGCCACTAGCGGCTCCGCT
GGCCTCGACCTCCCCGTGGTGGGATACTTGATTATAACACCTGAGGTGGGGGTTCAACGAATCCCCACTGGGGTCGCCGGCCCTCTCCCG
AAGGGTACTGTGGGACCCATATTGGGCCGTAGCGGACTAACTCAAAAGGGGCTGATTGTCCTTCCGGGAGTGGTGGATGAAGACTATCTG
GGGGAATTAGCTGTTTTGGCATACAGTCTTCGGGTTGTGCAACTCCCTGAATGTTCACGCATTGCTCAATTGGTGATCTTTCCTTTTTAC
GCTTTAGGCTCCGTTCGGACCGAGGATCTGCGGGGGGCAAAGGGTTTTGGCAGCTCCGGCATAGGAGCTTATTGGTCCGCTGTCATTGGG
CAAGACCAGCCGAGATTGACAGTCCGTTTAAATGGGCGAGAGTTTATTGGTTTGGTAGACACAGGGGCTGACCGTTCGGTCCTAACAAAA
AGGGGCTGGCCCCCGGCATGGCCACTGCAAACCGGGGCTGCCCCGTTGGAAGGCATCGGAGAGGTTACCTTCCCCTCTCAAAGTGCCGCT
TACCTACACTGGGAGCTTGATAGCCGATCAGGATACATTCAACCATATGTGTTAGATCGCCTCCCCGTTAACTTATGGGGACGTGATCTT

TTAGGACAAATGGGGGCCGTAATTACGACCGACAATTGTCACCCGTGTAATTCACCGCGCCCTTTTCCTTCGCGACCACTGCAAAGGTCG
GGGTCAACCGCCCAACAGCCCAACGAATAACGTGGGCTTCTGATGCACCAGTACGGGTGGAGCAGTGGCCGCTACCAACCCCGAAATTGG
CTGCTTTACAAGTGATAGTCACGGAACAATTAGCGGCGGGGCATATTGAACCCTCCGATAGCCCTTGGAATTCCCCTGTCTTTGTGATAA
AAAAGTGCTCGGGGGCCCGGAGGATGTTGACTGATCTCCGAGAAATCAATAAGACGATGCAGCCGATGGGAGCGCTGCAACCTGGGCTCC
CGAACCCGGCTATGATCCCTAGAAATTGGCCGATATTAGTAGTAGATATTCAGGATTGCTTTTTCCCTATTCCGCTCCACCCGGACGATC
GTATTCGATTCGCGTTTTCAGTCCCTGTGGCCAACAGGACTCAACCTACTGCGCGATGTCATTGGACGGTCCTCCCCCAGGGAATGAAAA
ATAGCCCCACAATGTGCCAGACTTACGTTGCGCTTCTGATAGCGCCACTCCGCGTTAAGTATCCTGAGGCATATATTATTCACTATATGG
ATGACATACTCTGTGCTTGCGCCACCGCAGCCCAACTACACGCCTTACGGCCTGCTCTTCTGGTGGCTCTAGAATCTGGAGGGCTGCGGG
TTGCACCCGGGAAAGTCCAGACAAAACCCCCGATCACCTATCTGGGACATATTTTGACAGATCGGACAGTGTCCCCTGTGGCCCCTACCT
TGGATGTTTCAAAATTAAAAACCCTTAATGATTTCCAAAAATTACTGGGGGAATTAAACTGGGTGCGCCCTTATCTTGGAATCCCCACCG
ATCAATTATCTCACCTTTTTGCTACCTTGCGAGGAGCCCCGGAGCTCTCTTCCCCCCCCCGCTCCCTTTCACCACAAGCCGCCCGGGAGT
TAGAGCAGGTAGTCAAAAAATTGGCTCAGGCTACGGTGGACCGTGCCACCCCAGGGATACCCTTATCCGCCGTTGTCGTTCCCACCTCTG
TCATGCCTACAGGGGTTATATGTCAAGAACAACAATTGGTTGAGTGGATGCATTTGCCCACTCAGCCCCACCGCTCTGTTTCCCCATATC
TGACTTTGGTAGCCCAGTTATTGGGTCGTTTAATTCGGCGCATTACCGTCCTTTCTGGGACCGGACCCTCCACATTATACCAACCATATA
CTTCAGATCAACTTGCCACACTTCTTGGCAACGATCCCGATTGGCAGATCCTTTTGGCAATGCGTACCGGGGCTTGGGTTTCTGGTCTCC
CAGATATTCCACGATTAAAGGTTTTCTCACCTTTTGCATGGGTATTCCCCTGCATCACGGTTCCGGACCCGATACCGGGGGCTACCACAG
TTATTACAGATGGAACGACAGGCCGAGCCGCCTATTACTGTCCTCCGGACACTTCCCGCGTTATCCCCGTTCTGTCCCATTCCGCTCAAC
ACGTGGAACTTGGAGCTGTGATGTTAGTCCTTCGCGATTTCCCTCAATCACTTAATATTACCTCAGATAGCCGGTATGCATGCTCTGTTA
CTCTTCACATTGAGACTGCTGCGCTCCGGTTTTCCGAGGCAGCCATTTTTCCTTCGTTCCAGGAACTTCAAGCAACGATTCGCGCTCGTT
CTCATCGTTTTTACATTTCTCATATTAGGGCGCATTCGTCCCTCCCCGGACCATTGGTTGAAGGTAATGCTTGTGCGGACCCACTTACTA
GGGAAATTGCTAGCGCACCAAGGGCGATTATTTCTGTTGCACCACTGTTGGCTCCGGCTCTTCTCCCTTTAGATGCTACAGCCGCAGCCT
CGCAAGCCCACCGGCGCTTCCACCAAGGGGCGGGAATGTTGCGCCGTCAATTTGGCATTTCTCGGGAAAGCGCGCGGGCTATTGTGAAAG
CTTGTACCTCCTGTGTTACTACATTACCTGCGGCCTCTGAAGCCACGAATCCCCATGGTCTTGCGCCAAGAAAGCTGTGGCAAATGGACG
TCACGCATTACCCTCCTTTTGGCAGATTGGCTTTTATTCATGTTACTGTTGATACCTGTACATTATTTACTTTTGCGTCGGCTCACACTG
GTGAATCAGCCAAACATCTTATGGGTCATTTGTTTCTCGCTTTTGTCTTTATGGGAATTCCACAGGTATTGAAAACCGACAATGGCCCCG
CATATACTTCTAAGGCGTTCCGGTCTTTCTGTAATATTTTTGCTATCCGCCATGTTACGGGTATTCCATACAATCCCACGGGTCAAGCCA
TTGTCGAAAACCGTCATCGTTGGCTCAAAGCTTTTTTGGAAAAACAGAAAGGGGGAGATGAACTGACTTCTCCCCATAGACAATTACAAT
CTGCTATATATACCCTTAATTTTTTAACTATGGATGACAAAGGTCTCACCCCTGCTGAAAAGTGGGGTGGCAGGGACATTAAAGAACACC
CTGAAAGAGTCAAATGGAAGGACCCTCTCACGCGCCAGTGGCGAGGACCAGGTCCCTTATTAACATGGGGCCGAGGGTATGCTTGTGTCT
TTCCAGAAACTGAGGACCGCCCGATCTGGGTAACCACAAAGAACATACGTCGCGCCCTTTCTACAGCACGCCCTCCCTCACCCTGCATAC
TTCAGGAGCTGACAGAGAACCAGGATACACACTCCGCGACCCTTCCTCTGGAGGATCAGCAGCAAAGTTCCTAAAAATGATGATTGCTTT
CCTGATTCTCGCCTCCTCCACTGTTGTGATTACTGAGGACTATTGGATTTTTGTTCCGTACCCTCCGTGGGTGTGACCAGTTACGTGGCA
AGATCCAGTGGGAGATATTCTCACTAATGCTTCTGACCTTGTAGGGGGGGATGCGATTCCACGGGAGGGACCAGGAGAAGAGGCACTTGT
CAATGTCTCCCTCCTTGCTACTCGCACTCAACTTTCTTTTCAACGTTGTGATTATCGTTAGCCTCATATCTGTGTGACCCCATATTATTA
TAATGCCACATTTCCCGGGTTTTCGTGGTCCGCGATCCGAAATCGGCTCCAAGGAGCCTGGCATTCTTCAAATGTCTCCCTTGATTTGTC
ACACCTTGCCTCACTCTCCGATGCTATGCAAAATACCCCTCCCTTGACCATTCCTGCCGATATTGCTTCACCGATCATTTCTGCCCTTGA
GGCGGTTAATCCTTTTAATTGGCTCTCAATGCTGCTTCATAGTCTCCCATATATTATTTTGTTGATCTTGGCCCTGTTTTGCCTGCCCTG
CGTTGCCCGATGGCTATCTGTTTTCCTTTGCCCGCTTCTTACTCAACTCCATGCTTTGCAATTGTTTCAAAAACAAAAAGGGGGACTTGT
GGCAGACCTAACATCAGGGCCTAGGTAGT

>Tacu_ERV3b#ERV
ATGGCGCCCGAACAGGGATAGGGACATGAACCCTGGACCCTCAGACTCAGCACCGCAACCCTCACGAAGCAACTCATCACTGCGCAGTGA
CTCACTAAGGAACTGAAGAGGTGAGGGCATTTAAGTTTACTTTCATTTTCAAACACCAGGCTTACAGCAGGGAAGAAAATGGGACAAATT
CAATTCCGTTCTCGGGATCATTATATCCAATTACTTAAGGATGTTGTAAAGGAGCATGGGTTGATTATCACCACTGCCCAGATTGAACAA
TTTTTGGAATACGTGGAAAAATGCTGCCCTTGGTTTCCAAAGGAAGGAAAGTGGAATCCTTTCACTGAGCCAGCTCCTTCCGCACCACCA
CCTCCACCACCCCCTCCATGGGTCCCGAAACCAGCACAAACACCTCTGCAGCGAGCGATGGGGGTTGCTATGGAAGGGGGAGAGGGTTTC
CCTGTGATGGAGCAACCTGACCCTGCTAATCCTGGTCAGTTAATCCGTACCCACGAACTCCTGAATTTCAAACTATTAAAAGAAGTGAAA
CAGGGTGCTGCGACATATGGGCCCACCACTCCTTATGTGCTAGCACTCATAGAAAGTATTGCTTCCCCAATTCTTTGCCCGTATGATTGG
CAAACGTTAGCCAAGACTGTCCTGGATGGCGGAGACTATTTGCTTTAGAAGGCAGAATTTTTTGATTTGTGCGTGGAGCAAACACGTCGC
AATGATCGCGCCCGACCTCCAGTTCAAATCACTTTCGAGATGCTTACAGGGACTTACAGGGATTTACAGACTCGTGAAAGCTTGTACCAG
TCCCTACCCAAAGGGGTACTGATTACATGGAAAACTGTTTTTACGGTGCTGCGGGCGCTGCATCCCACTGAACGGGTGTTACAGGAGGCA
CCGGCGATGCGCGACTTGGGGGCAGGGAGTGATCAGGAGGACCCCCTGTCCTTTGAAGTTTTGTCACAAACAGAAGGATTGGAGCCGATG
TACGCTACTGTTCCGGATCCCGAGGACTTGTATAGGGACCAGGATCATAATGACTCTGGCGATACTTCTGACTCTGGGACTGCAGACCAG
AAGGAGGGATCTGATCTATACCTGCCTTTGGCCCCTCTTACAATGACGGCTGCTGCCGCACCCACAGCAAAGGACCTTTTGCAGATCCAG
AGCCTGCAGGCGGTATCCCGTTCGTCAATTGTTCCCCCTCCCCCCCTTGCTCCCCCTTTGTCTGCTGCGCCGCCGTATGTATATGCTGTG
CCTGCTATGGCGCAGCCTGCCTAGACAGGGGCCCTTTTGGGCTGGCGGCAGGAGGCATTGCGCCCAGGGGACATACAGCTACTGCAAGCG
ATGCCGGTGGCTTATCTTCCCAGCCGGCCGGCTCAGTATGACTCTCTCCCTTATGAGTTGATTAAGGAGTTGCGGAAAAGTGTTCGTGAT
TATGGGTTACAGTCGTCCTATACAATGAATCTAATAGTGGCCTCTACCAACCCCGAAATTGGCTGCTTTACAAGTGATAGTTGCGGAACA
ATTAGCAGCGGGTCATATTGAACCCTCCGATAGCCCTTGGAATTCCCCTGTCTTTGTGATAAAAAAGCGCTCGGGGGCCTGGAGGATGTT
TACTGATCTCAGAGAAGTCAATAAGACGATGCAGCCGATGGGAGCGCTGCAACCTGGGCTCCCGAACCCGGCTATGATCCCTAGGAATTG
GCCGATATTAGTAGTAGATATTCAGGATTGTTTTTTCTCTATTCCGCTCCACCCGGACGATCGTATTCGATTTGCGTTTTCAGTCCCTGT
GGCTAACAGGACTCAACCTACTACGCGATATCATTGGACGGTCCTCCCCCAGGGAATGAAAAATAGCCCTACAATGTGCCAGACTTACGT
TGCGCTTCTGATAATGCTGCTCTGTGTCAAGCATCCTGAGGCATATATTATTCACTATATGGATGACATATTATGTGCTTGTGCCACCAC
AGCCCAACTACACGCCTTACGGCCTGCTCTTCTGGCGGCTCTAGAATCTGGAGGGCTGCGGGTTGCACACGAGAAAGTCCAGACAAAGCC
CCAAATCACCTATCTGGGACATATTTTGACAGATCGGACAGTGTCCCCTGTGACCCCTACCTTGGATGTTTCAAAATTAAAAACGCTTAA
TGATTTCCAAAATTTATTGGGGGTATTAAACTGGGTGCGCCCTTATCTTGGAATCCCCACCGATCAATTATCTCACCTTTTTGCTACCTT
GCGAGGAGCCCCGGAACTCTCTTCCCCCCGCTCCCTTTCACCACAAGCCACCCGGGAGTTAGAGCAGGTAGTCAAAAAATTGGCTCAGGC
TACGGTGGACCGTGTCACCCCAGGGATACCCTTATCCACCGTTGTCGTTCCCACCTCTGTCATGCCTACAGGGGTTATATGTCAAGAACA
ACAATTGGTTGAGTGGATACATTTGCCCACTCAGCCCCACCGCTCTGGTTCCCCATATCCAACTTTGGTAGCCCAGTTATTGGGTCATTT
AATTCGGCGCATTACTGTCCTTTCTGGGACCAGACCCTCTACATTATACCAACCATATACTTCAGATCAACTTGCCACACTTCTTGGCAA
CGATCCCGATTGGCAGATCCTTTTGGCAATGCATACCGGGGCTTGGGTTACTGGTCTCCCAGATATTCCACGATTAAAGGTTTTCTCGCC
TTTTGCATGGGTATTTCCCTATATCACGGTTCCGGACCCAATACCGGGGGCTACCACAGTTTTTACAGATGGAACAAAAGGCCGAGCTGC
CTATTATTGTCCTCCAGACACTTCTCGCATTATCCCCATTCTGTCCCATTCCGCTCAACACATGGAACTTGTGGCTGTGATGTCAGTCCT
TCGCGATTTCCCTCAATCACTTAATATTATCTCAGCCGGTATGCATGCTTTGTTACTCTTCACATTGAGACTGCTGTGCTCCGGTTTTCC
GAGGCAGCCGTTTTTTCCTTTGTTCCAGGAACTTCAAGCAACGATTCGTGCTTGTTCTCATCGATTTTATATTTCTCATATTAGGGTGCA
TTCGTCCCTCCCCGGACCATTGGTTGAAGGTAATGCCTGTGTGGACCTACTTACTAGGGAACTTGCTGGTGCACCAAAGGCGATTATTTT
TGTTGCACCACTGTTGGCTCCGGCTCTTCTCCCTTTAGACGGTACAGCCGCAGCCTCGCAAGCCCACCGGCGCTTCCACCAAGGGGCGGG
AATGTTACGCCATCAATTTGGCATTTCTTGGGAAAGCGCGCAGGCTATTGTGAAAGCTTGTACCTCCTGTGTTACTACATTACCTGAGGC
CTCTGAAGCCACGAATCCCTGTGGTCTTGCGCCAAGAAAGCTGTGTCAAATGGATGTCACGCATTACCCTCCCTTTGGCAGATTGGCTTT

TATTCATGTTACTGTTGATACCTGTACATTATTTACTTTTGCGTCGGCTCACACTGGTGAATCAGCCAAACATGTTATGGATCATTTGTT
TCTTGCTTTTGCCTTAATGGAAATTCCACAGGTATTGAAAACCGACAATGGCCCCGCATACACTTCTAAGGCGTTCCAGTCTTTCTGTAA
TACTTTTGCTATCCGCCTTGTTACGGGTATTCCATACAATCCCACGGGTCAAGCCACTGTCGAAAATCGTCATCGTTGGCTCAAAGCCTT
TTTGGAAAAACAAAAAGGGGGAGATGAACTGACCTCTCCCCATAGACAATTACAATCTGCTATATATACCCTTAATTTTTTAACTATAGA
TGACAAAGGTCTCACCCCCACTGAAAAATGGGGTGGTAGGGAAACTAAGGAACACCCAGAAAGAGTAAAATGGAAGGACCCCCTCACGCG
ACAGTGGTGAGGTCCAGATCCCTTATTAATGTGGGGCCGAGGTTATGCTTGTGTCTTTCCAGAAACTGAGGACCACCCGATCTGGATACC
TACAAAGAACATATGCCGTGTCCTTTCTACAGTGCGTTCTCCCTCACCCTGTGTACTTCAGGAGCTGACCGAGAACCAGGAGACAGACTC
TGCGACATTTCCTCCGGAGGATCAGCGGCCCAGTTCCCCAAAATGATGATCGCTTTCCTGGTCCTCGCTTTTTCCACCGTTGTGATTACT
GAGGACTATTGGACTTTTGTTCCGTACCCTCCGTGGCTGCGGCCCATTACGTGGCAGGACCCGGTGGGAGATATTCTCACTAATGCTTCT
GATCTTGTAGGGGGGGATGTGATTCCATGGGAAGGACCAGGGGAAGAGGCACTTGTCAATGTCTCCCTCCTTGCTACTCGCACTCCTATA
TGTTTCTCTGCAGCTCCGACCCCAAGTCCGTGTCTGTCTCTGCTCCCCCTCACTACTTGGGCTCAAGAAAGAGAATTGCGTGAACCCCAG
CCCCGTTTGGTGTACTGTGGGATTACCCGTTTGACTACTGTATCCCGTCGTTTGGGAACCTCACCTTCGGTACCCAAACTTCCTTCTTGT
GTGTTTGCCAGCTTGTCTTGGCCACCATCAGAGTCTGTTTTGCCTAACTGGCAAACGTGCCGACAACCCTTCCCCACGGTTATTCATTTG
TCCTCTCACGTATTCCCAGCCACTCTACTGAATTGGGGCAGACATGACAACCTTATTTATTTATCCCCTTTGCATGGTTATCTTGGGTTT
CTGCCCACGGATGCTTTCCAGGCTTTAGGTTCCTCTCATGCCGGAACTTCCTCCCCCTTGTACCCTAGTCTTTGGCGGGCTTTTGCTGCT
TTGGGTCCAATCACGCTTGCTTTTGCTCGATTACAGGTCCCTTTCGATTCTATCCCAATGCATATGGAAGAACGGCAAATAAGAGCATGT
ATATTGCCACCGCAAGCCTTCCTTGTTGGCTACCCATCCCTTATCCCCTTCCCTTCCCTCTGGAATATTACATGTGATAATTGTAGACTT
ACCCAATGTATATCCGCCCATGATGCCACTGCTACTATTCTTACAGTAGAACAACCTCGTCATGCTATACTACCCACCTCATTATCCCAC
CCTTGGGCTGAGTCCCCTGCGGATTCTGCCCTGCAGATGCTAGCTGATCATTTTCGTATTCCAAAATGGTGACATCGTCGGTTTGTTGGA
TTGTTAATCACTGGGTTTATCACTTTGGCATCCCTTATTACTGCTACCACCTTATCCTCTATTACCCTGCATACTGAAATAACACAGGCC
TCCTTTGTTAATTCCCTTGCAAGAAATGTTTCTGATGCTCTGCAAACACAGTTATGATTACATAAAGCCTTTACCGCTCGCCTCTCTCTT
CTACAGCATTCTCTTATGCAATTAGGAAATGAGGTTGATGCTCTGCGCACTCAACTTTCTTTTCAACGTTGTGATTATCGTTACCCTTAT
ATCTGTGTGACCCCATATTATTATAACGCCACATTTCCCGGATTTTCGTGGTCCGTGATCCGAAATCGGCTCCAAGGAGCCTGGAATTCT
TCGAATGTCTCCCTTGATTTGTTACACCTTGCCTTACTCTCCGATGCTATGCAAAATACCCCTCCCTTGACCATTCCTGCCGATATTGCT
TCACCGATCATTTCTGCCCTTGAGGCGGTTAATCCTTTTAATTGGCTCTCAATGCTGCTTCATAGTCTCCCATATATTATTTTGTTGATC
TTGGCCCTGTTTTGCCTGCCCTGCCTTGCCCGTTGGCTATCTGTCTTCCTTCACACGCTTCTTACTCAACTC

>Tacu_ERV4#ERV
TTTGGGGGCTCGTCCGGGATCCCCACCATTTTAGGCGGACCCCTTCGCTCCGCCATTCGGGATCGGCGCAACAGACCCGGTACCGGTGAG
TCACTTTGTCAGGCCTCGCGAGGGTTTGGGAGTATAGGAACGGCAGGACGCTGCCTTTGTTCCGACTCCACTCGGATCAGGGGACGCTCT
GATCTCGAGTTTGGGTCTCATGGTCGAGTACAGTTGTCCGTGAGACGTGTATGTTTTGGGCTGTTTGTTTGGTTGTTGGTCCGTCGTTTT
TGTTCTTTGTGCGTCTTGCGAATTGCCGGTGGACCCCGGCTAGAGCCGAGGGAGAGGGGGCGCATGCGTAATGTTTTAATCTTCGGGCGA
GTACGGGGTACAGGAAAGGGCCCTTCAAGCCCGTTGGAACATCTTCTTCGGTATTTTTGTCTCCCAGTTTTTCTGCTCGCACTTGCCGTC
CTCTCCCTCTCGGCCAAGATGGGACAGGGCAGGTCAAAGGGCCCCTTGAGCCCGTTAGGGTGTTTGCTCAAACACTTCTCTGATTTCCAG
CGGCGAGCTGATAACTATGGCGTGTCTGTTAATAACTTTGACTTACGCAGGTTTTGCGAGTTGGAATGGCCCACCTTTAAGGTTGGCTGG
CCTGACACTGGGACCCTAGACATAGGGGTGGCGGCCGCCGTCCGCCGAGTTGTGGACGGGAACCCAGGCCACCCGGACCAAATCCCATAC
ATTACCATTTGGATAGATATTATAGTAGACAACCCTAAGTACTTAAAGGACTGCGGGTGCCGGCCCCTCAGCGCTTCTAAGGTCCTCGTA
TCCAGTACTCTAGGGTCCAAAGCTGCTCGCAAGCCTCCCGTGCTCCCCACGCAACCGGAGAGCCCGAGGAGGGGGCGAGCGCCTCCTCCG
CGAGCGCCCCCTCCCCCCTATAGGGAACCCTCAGCCCCGCCCGAGGAAGAGGCTTTCCCCAACCAGATTCCACCGGACCTCCAAGCCCC

CCCCACACCCGAAGTGGGACTGAATTTGGGCCGACAGGAAGGGCACCGGGGGTCTCGGGAATATATCCCCTGAGGGAAACAGGAGAAAGG
GATGAAACGGGGCGGCCTGTGCGGACATATGTTCCTTTTACCACGTCAGATCTGTACAATTGGAAGAACCAGAATCCTTCCTTTTCCCAA
GCCCCGGAGGAGGTAATCAACTTACTAGAGTCAGTCTTCTATACCCATCAACCTACTTGGGACGACTGCCAGCAACTCCTCCGCGTCTTG
TTTACGACGGAGGAAAGGGAAAGAGTAAAGGCAGAGAGCAAAAAGGAGGTCCGAAATATTCGCGGTGAACCAAGCACTGACGCAGGGGAA
GTGGAGGCCCAGTTCCCCTCTGGCAGGCCTGATTGGGACCCCAACACCCCGGGAGGGGAGGCCAATCTGAATCAATACCGCCAGATCCTC
TTACGGGGGCTACGGGCGGCGGCCAGAAAGCCGACTAATCTCTCTAAGATAACCGAGGTCCGGCAGGGGCCAACGGAAAGTCCTACGGCC
TACCTGGAACGACTATATCAGGCCTACCGGACCTGGACCCCCATAGACCCTGGGAGTCCTGATAATCAGGCAGCTATAGTAATTCAATTC
GTGTCGCAGTCGGCCCCAGATATCCGAAAAAAAGATTCAAAAAATGGATGGGTTTCAGGGAAAGCCTCTCTCTGAGCTGGTAGCCATAGC
CCAGAAGGTTTTTGACCAACGAGAGGACCCCACCAGAACAACTTATGAATTAACCCAAAAAATGGCGAGGGTCCTCCTAGCTCGAGAGGA
ACATTCAGAGAATAGGCGACGGGGAGGCAGGTCAGGTCAGAAGAGGCTGCCCCTGGGAAAGGACCAATGTGCCTACTGCAGGGAGAATGG
GCACTGGAAACGGGACTGTCCCAAGTTAAAAGGGGGCGCAGCTCCGGTCCTGGTAGAGGAGGAGACTCAATAGGGCCGTCGGGGTCCCTC
AGCCCTCCAGGAACCCAGGCTAAAGTTAAAAGTCGGGGGGCAATTGATTGATTTTCTGGTTGACACAGGGGCAACCCATTCAGTAGTGCA
GAAACCCGTTGGTCCAATGACAAGGGATACGGTGACTATTGTAGGGGCCACCGGGGCCACGTGCAGGTACCCTAAATCGGAAGGTCGAAT
TGTTGATCTAGGGAAGGGATTGGTAACACACTCCTTCCTAGTTATTCCCGAATGCCCTGACCCCCTGTTGGGACGGGACCTCCTGCACAA
GTTAAGGGCCACCATTATATTCCCCGAAGCGGGGACCCCTGAAATTAGAACTGAAGGCAAGTTACTGCTGTCCTCACCCTTGGTGGAGGA
GTATCGTCTGTTCACTGAACAACCTGCACAAAACCTCGCCCTCTTAGATTTATGGAGGGAGGAGATCCCCGGAGTATGGGCAGAATCGAA
CCCTCCGGGACTCGCTACTACCCAGGTCCCCGTGCATGTCCAGCTTACCAGCACGGCCCTGCCGATCAGAATAAGGCAATACCCTATAAG
TCTGGAGGCTAGAAGGAGCCTCAGGGGGAGTATTCGGAAATTTAAGGAGGCAGGAATATTGAAACCCGTCCACTCCCCTTGGAATACCCC
CCTCCTACCCGTCCGGAAAACTGGGACCTCGGAATACCGCATGGTACAGGACCTGAGGGAGGTGAATAAGCGAGTGGAAACCATACACCC
CACTGTTCCCAACCCTTATACCCTCCTCAGCCTTCTGCCACCTGACCGAACCTGGTATTCGGTCCTAGATCTTAAGGACGCATTCTTCTG
TATACCTTTGACTTGTCAATCACAGCTCCTGTTTGCATTCGAATGGATAGACATGGAGGAGGGGGAGTCGGGCCAATTGACCTGGACCAG
ACTGCCCCAGGGATTTAAGAATTCCCCCACCTTGTTTGACGAAGCTTTGAGTAGAGATTTGCAGGGATATCGATTTGACCACCCAACAGT
AACGCTCCTCCAGTACGTAGACGACCTTTTGATTGCCGCCGGGAGTCGAGATGAATGCCTCCAAGCTACCAGGGACCTGCTGGTCACTCT
AGGATCAATGGGGTACCGCGTGTCAAGCAGCAAGGCCCAGCTGTGCCAGGAGGAGGTCACTTACTTGGGATTCAGGATCAAGGACGGGAC
CAGGACGTTGGCCCAGAGCCGGGTCCAGGCCATCCTGCAGGTCCCAGCCCCGAAGACCAAGAAGCAGGTACGAGAGTTCCTGGGCACGGT
CGGCTACTGCAGGCTCTGGATCCCCAGCTTCGCGGAGTTGGCACAACCCCTATACGCCGCCATCCGAGGGGCCGATGCCCCCCTACGATG
GACCAGTACCGAAGAGGAAGCCTTCCAGCGGTTGAAAACGGCCCTGCTGCAGCCACCTGCTCTGGCCCTACCCGACCTGGACAAGCCCTT
CCAGCTTTTTGTAGACGAGGCAGAGGGTGTTGCCAAGGGGGTGCTCATGCAGACTCTCGGCCCCTGGAAGAGACCAGTGGCGTATCTCTC
CAGGAAACTGGACCCCGTGGCCGCCGGATGGCCCCGCTGTCTGCGGGCCATTGCAGCCGCCGCCCTCCTGTCCAAGGAAGCGTCGAAGTT
AACCTTCGGGCAGAGTTTGGAGATCACCTCGTCTCACAACTTGGAGGGTCTCCTGCGCACGCCCCCGGACAAATGGCTGACCAATGCTCG
AGTAACCCAATATCAGGTCCTGCTCCTGGACCCACCCCGGGTGATCTTCAAGCAAACTGCGGCACTTAATCCCGCAACCCTGCTGCCAGC
AACTGACGACTCCTTGCCCCTGCATCACTGCGCGGACACCCTGGATGCCCTAACCACCACCCGCCCGGATCTGACCGACCAACCCCTTGC
CGACGCTGAGGCCACGCTCTTCACTGATGGGAGCAGTTACGTGAAGGAAGGCCTGAGGTATGCGGGGGCGGCCGTGGTGACAACGGACTC
CATCGTCTGGGCTGAGGCACTCCCGAAAGGGACGTCGGCCCAGCGGGCTGAACTTATAGCCTTAACCAAGGCGCTGGAATGGAGCAGGGG
TAAGACTGTGAACATCTACACCGACAGCCGTTATGCGTTTGCTACCCTGCACGTACATGCAATGATCTACAAGGAAAGGGGACTGCTGAC
TGCCGGGGGCAAGGCCATCAAAAACGCCTCTGAAATTTTAGCTCTTCTAACGGCCATCTGGCTGCCAAAGCGTGTCGCCGTCATCCACTG
CAGAGGACACCAACAAGGTGAATCGTTGGAAGCATTGGGAAACCGGCTGGCTGACAAGACAGCCCGGGAGGTCGCTAAGAAGTCACCGGC
AATTCAGGCCTCCCTGTGCGACTCGCCCCGTACCCCAGTTGACTGGGTCCCAGTGGACACCCCACAATATACAAAAGAGGAAGAGGCTCT
CGGCCAACGGCTTGGCGGAACCACTGACTCGACCGGCTGGTGGCGACTCCCTGACGGGCGGATCCTACTCCCAAAAGCAGTAGGGAGGCG

GGTAGTCGAGCAGACCCACCGTGCTTCCCATCTTGGGGAATCCAAACTGGCCGCGGTCATACGAAAGCACTACCTCATCTGTGGCATCTA
CGGGGCAGTAAAAGACGTGGTGCGCAGGTGCGAGGCCTGCGCTCGGGTAAATGCACAATCCGTCCCTACCAGCTCGGCCGAGAACGTCCG
CGACCGAGGACTGGCCCCCGGGGAACATTGGGAAATTGACTTTACCGAGATGACTCCGGCCCGGGGCGGCTACAAGTATTTGTTGGTCCT
GGTGGATACCTTCTCCGGATGGGTGGAGGCTTACCCGGCGAAGGGGGAAACGGCTCAGATTGTCGTCAAGCACCTGGCAAATGATCTAGT
CCCGCGATTTGGACTGCCACTTCGTATTGGGTCTGACAATGGTCCGGCTTTTGTCGCAAAGATAACTCAGCAGCTGGCCTCCGCGCTCCG
GATCACCTGGAAACTACACTGTGCGTACCGGCCCCAGAGCTCTGGGCAGGTGGAAAGGATGAATCGGACTTTGAAAGAAACTATCACCAA
ATTAAAGATGGAAACTGGGGGTGATTGGGTCGCGCTTCTCCCCCAGGCCCTCCTCCGGGCCCGGTGCACACCAGGGAGGGAAGGCCTGTC
CCCCTTTGAGATTGTCTATGGTCTGAGGCCCCCCCTGGTGCCCCGAGTCGGCCTTGACCAGCTTGCCGAAGTCACCCATCGGTCTTTGCT
TAAGTCCTTACAGGCGCTGCAGGCTACGCGGTCCCTCGCCCGGACGACCTTGGCAGATCAGCAACCTGGGGCAGAGGTCCATCGAGGAAG
GGAGCCCCTCTTCCATCCTGGGGACCTTGTCTACGTAAAGAGACTCGACCCCCGGCAGCTCGCTCCCCGCTGGGACGGGCCCTTCACTGT
CGTTTTGAGCACTCCCACTGCCGTGAAGGTAGCTGGTAAGACCCCGTGGATCCACCACACCAGGTTGAAGGGTGCCCCAGAGTGCAGTGG
AACATGGGGGATCGATCCTGCCTCCACCCCTCTCAAGTTAAAACTCTCCAGACGTTTGTCATAACCTTGTTCGGAGTCTGCTGTCTCTCT
GGGGTTCTGGGGGGCCTCGGGCCCCCCCCTAACAAAGAGGCACTGATAGCGGCCCTGTGGGGACCCCCCTGTAACTGTCGAGGGGGCGTC
CAAGAAACCGTACCCACCACCTATACGCGATCAGTTACTTGTGGCGGCTCGACGGCTTACCTAGTGTACAACAGGGGAGTGGGTGGAGGA
TATCATCAACATTGGGTATGCACTCGCCGACCTAAGGTACTACTGTCCCCGGACGGTCGACCTGGGCCCTGCCCAACGGCATGCCAGGTC
ACCTCCCAGATGCACTCCACTTGCTATAGCGCGGCCCAGCAGTGTAACCACACGGACGGGAGGGTCTATTTGACTGCCGTCCTACAGAGA
ACTTACAGTGGCTCCTTTGGGGGAGAAATGGACCACTCCAAATACGCCCAGGCCTCCTGCACGGGGACCGTCGGTCAGCCCGTCTGTTGG
CCACTCCGGGCCCCCCTCCACATCTCTGATGGTGGGGGGCCGTCCGATCGCGTCCGGGAAGCGCAGGTGTCGGAACGCATGAAGGAGGTA
ATCCAGATCTTACACCCCTCGATCCGGTATCACCCCCTGGCGCTACCTAGGCCCCGGAGCCCCGGTCTGGACCCTCAGACTGCGGATATC
CTCGCCGCCACCCATCAAACTTTGAATGCCACTAACCCTGCGCTGGCGGCAGATTGCTGGCTCTGCATGCCCTTCGGCCATCCTATACCC
ATTGCAGTGCCCGGGAGCGCGTCCAAAAACGCTTCCGTGTCCGGGCCCGATTCTCCTTACCCCGGAGGAAACTGTACTCACAACCTTCCC
TTTCGTGTGCAGCCGCTTTGGGCCCCCCCTGTTCCCTGTTACCTTAGGTCAGGCCCCAACACCAGTATCGCCTTGGACGTAGGCTTCGCC
TCTTGTAACTGCTCTCAAACCGTCAATGTCTCCGCCCCACTGTGCCCGGCCCCCGGTCGAGTCTTTGTGTGCGGTGGGAACTTGGCTTTC
ACGGCCCTTCCCGCCAACTGGACGGGTCTTTGTGTCCAAGCCTCCGTACTTCCTGACATCGACCTTATTCCAGGTGACGAGCCTATTCCA
CTCCCCAGCCTGGATTATATCGCCGGTAGACATAAGAGGGCCATTCAGTTTCTCCCCCTACTCCTGGGCCTTGGGTTGGCCGGTGCGGGC
ATGGGAGCAACGGGTCTAGGGGTGTCGGTCCACTCCTACCATAAATTGTCCACCCAGCTCATTGAGGATGTCCAGGCTCTTTCAGGCACC
ATCCGCGATCTACAGGACCAGATTGACTCCCTTGCTACAGTGGTCCTACAAAACCGGAGGGGCCTGGACCTGCTGACAGCTGAACAGGGC
GGGATCTGCCTAGCCCTAAAGGAACATTGCTGCTTCTACGCTAACAAATCCGGGATCGTTCGGGACAAGATCCGCAAGCTCCAGGAGGAC
TTGGCCGTGCGGCGGCAGGAGCTGGCCAACAACCCCCTCTGGAGCGGCTTCAATGGACTCCTCCCTTATCTGCTGCCACTTCTGGGCCCC
TTGTTTGCATTGATCCTTGTGTTGTCTATCGGCCCCTGCCTGTTCAGAACTGGAGCACGTATGCTCCAGGATAGGCTGCAAGCTATTAAA
GTCCTGGCCCTGATGTCCCCGTATCAACCAGTGCCCCCTGAGGACCCCTCCCCCGCGTAACCTTGCGTTTGGCTTCTGTACCCACGCTTT
GCTGAGCGGTCAAAGATTTGCCCTTCACTGACAAAAAGCAGTGGGGA

>Tacu_ERV5#ERV
TTTGGAGGCCCTAGTGAGATGCACGGTGTCCTGTGTCACCGTGGAGACCCAACCTGGGAGGACCTCGGCCTCGGGGGAAGGAGACCATTC
TGCTCTGACTTCTAGGGGGGCCGGCCCCTGAGACGTTCCAGGGCCCCGGAACTCAAGGCTGAGAGGTCTAAACTGTCCCCTGACGCCGAA
TGAACCTCATTTCGGCCTGGTGGATCCTGGGATCGGCAACGATCAACAACTTCCAGAGGTAACCTGGTTTCTGTTTTCCGGAGGGAACGA
GTGCCGGACGCGGCAGTTAGCGCTTTGCGCCTCGATTCCCCGGGCACCCCAAGACGTTGGGTCTCTGCCCGATTCTGGTTCTGGTTCTGG
TTCCGCATTTGTCTGTGAATCTGTTCTATATGTGGAAGCGTTGCCTTTAGGAAATTCTGTTTTACGATCCTCTGTCTGAAACCTCCTCGC

TCACAGACTGCCAGGCAGGTTTCCACCTGGATCGGGGACGAATATCTTCAGGGCAACTTCTCCAATGTCTACTGGAAGGATGTGTTCGCC
TCCAGGTTGTTGTTTGTTTGTGTCTCTGTGTTGTATGAATGGGAGGTTCAGCCAGCACGCCAGAAACGCCTTTGAACTGTATGCTCAGTC
ACTTTAAAAAAGGATACCGGGATGGGTATGATTATGGGATTACCCTAAAGAAACAGAAGCTCATTCTATTTTGCATGAATGAGTGGCCCA
CCTTGGGGGTGGGGTGGCCCCCTCGGGGTAGTTTTGATAAGCAAATTGTAAACAAGGTTTGGAGAATTGTGACTGGGACCCCTGGCCACC
CCGATCAGTTCCCCTATATTGACGTTTGGCTGGACCTCATTACCCACCCCCCTCCCTGGTTGAAAAATTGCTCCCTGAAAAGGGGCATCC
GAGTTCTTTTTGCCCAACCTAAAAAAGGGCCCCCTTCCAGGGACAGCCCCCCCCAAAAAAAAGTCCTCCAGGAGTCCTAGGAGGATGATC
TTCCCCCCTCCTTGTAACCCGGGCGCCCTTCGGGACGTCCCGCAGTTTGAGAGCCCGCCCTGGACTCTCAGAGAGAGTCGAGTGGCATAG
GCCCCATTAAAGGTACCCAGATTCGAGTATTGGGATGGACCTCCAAAATATGGACGGGGGAGGTTACTTATCATTGTACGTGTTTCCATA
AGCGCTTAGTACCGTGGTCAGTGCACGGTTAAAGTTCTCTAGATACTAGAAAGTGACAGGAGTGTCTGTTTGACTTGAGGAAACGAAGGT
GGGAAAAGGAAAAATCTCAGTCAAAACCGAGTGGTGGGAAGTCCCTCCCGGAAATTCCGGAAGGCAGTCCGATGAACTTTCAAAGTATAA
GAGAAAAGAGAAGAAATTAATTATAAAAATTATTATTAAATAATTACATATTACTGCCTAACTTGGGCTGGCGCCCACCTCGGAGGGAAC
GTCTTCTGGCCAAAGTACGGATCCTCCGAGGACTGGGTTTGCCAAGAATTGAACATCTATGTAATTATGAGGGAACCCCGGAATGATGGG
GAATTTGTGTATGCCGCAGTCTGGCGGCCGTCGGTGGGTAAATTTGTGCTGAAGGAAATAGATGAGAAGAAAGATAGAGGGGGAAGAAAG
AGAAAAACCGGCACACCAAAGTGATGTTGCAGACTCCAAAACAACAGGAAAAGCAAAGAAAGATTATGATATAAATCGTGCAGATGGTGG
CAAAGTTGATCATGATTTATTACGCCTGCGGCCGGCTGGGGCATAAGAAAAGAGGGTGTCCTCAGCGGGAAAAAGAAAAGAGTTCTAGGT
ACATGAAGGAAAAGGGAAGGTGAAAAAATTGCAGTTGAAGGAAATTGTCTGCTCAGCCATCTACACCCTGTGAAGCAGGTGTCCCATAAG
GGGGGAATGAGGGGTGGAGAGATTGCCTCTGAAGGAAATTGTATGTGTATGAAGGTTAGAGTAAGTGTGTGTGAAGGAGAACGAGAAGGG
ACAGGAGATAATTTTAGTAATTCCTTCGTGGAGGGGGAAGACTAGGGGGATCAGGAGCCCATAGAACAAGCCCACCCTGAGCCCTTGACA
AATTTGAGGGTGGGCAGAGAAAAACAAGATTAACAGTATTATATCCTTTTGATAGGCACGGGCGCTACCCGATCGTCTCTAACCAGACAG
CCGATGAGGGCCAGAATCGAGAGGGAAATCATTAGGATCTCGGGGGTGCAGGGGGAGAACTTCCCGGTCCCTGTTACAGAGACCTTAGGA
CTAGAATACCTAGGGTCAAACTTCTCGGGAAAATTCCTGATCATACCCGAAGCCGGGGTAAACTTGTTGGGAGGGCTCTTAACCTCGGCG
GGAAAAACCATTAAAAATAAGGAGATTTTGTCTTTACTAGATGCGGTCTGGATGCCGGCCCAGGTGGCCAGCATCCATCGCCCCGGGCAC
CAGCACGGGGACTCACCCGAGGCCATTGGCAATCAGGCAGCGGATGAGGCTGCCCGAGTGGCGGCAAAGTCCCCTCCCAGCGTTGCCCCC
TTATGCCCTGTTTTTGACCCTAATGATACACCGGCCCCTCACTACTCCCCCTCGGATGATTCCTTCGCAAAACAAAAAGGGGGGACTAGA
GACGGGTCAGGATGGTGGGTCCTTCCAGAGGGAAGAATTTTTGTCCCTGAGGAAGTAGGGAGAGAATGGATCACGCGGCTACACCAGATC
ACCCATCTGGGGGTATGAAAGATGGGCTTGTTGTTGAGAGATAGGTACTTTATCCCACACTTGGACAGCCCCTTAGCGAGCATAACCACC
CGGTGTGGAACATATGCACAGGTAAATGCGAAGCAGGGGAAGGCTGCTCCCTCAGGAGTCAGATTGCCGGGATTGCAGCCAGGAGAAAAC
TGGGAGGTAGACTTTACAGAGGTGAAACCCCCCGCGGCAGGCTATCGATACCGCCTGGTATTTGTTGACACTTTTTCAGGGTGGGTAGAA
GCTTTCCCAGTTAAACACGAAACAGCCATGGTAGTGGTGAAAAAGATTCTAAATGAACTTCTCCCGCGATTTGGTCTCCCACTGGGACTC
GGGTCTGAAAATGGTCCAGCATTCATAGCCAAAGTGTCCCAAGGCATAGCCAAACCTTTAGGAATAGAATGGAAATTACACTGTGCCTAT
CAACCACAGAGTTCAGGTCAGGTAGAAAGGACCAACCGAACTCTTAAGGAATTCCTCACCAAACTGGTCCTTGAAACTCAGGAAAATTGG
GTAATGCTCCTCCCACTGGCCCTACTCCGAAGCCGATGTACCCCCAATAAGTCGGGTCTCGCACCTTTTGAGATTCTGTTTGGTAGACCC
CCGCCAATCCTCCCTCTAATCCGGGAGGAACTCAAGGTGGACGCTACTAATTCTTCCTTGATTAAGTTCCGGCAGGGTCTCCAGAAAACA
CAGGGAACACTCCTGAAATCTGTCCGAAACGCCCTGCCAGTTCCCACCTCTGCACCCGCACACGCCTTCCAACCCGGGGACTCGGTCCTG
GTCAAGAAATTCACCGCCTCCGGCTTGGAGCCTAAGTGGAAGGGCCCTTACACCGTCATCCTGACCACGCCAACAGCCGTCAAGGTTGAC
TCCGTTCCTGTCTGGCTCCATCACAGTCGAGTGAAACCTGCTGCGGCCCCGACATGGAAGGCGGAGGCACAGGCCGACCCCCTAAAGCTA
AGACTCTCCCGCATTTCCTCCTCCTCGCCTCCCTAATGACCCCCTCTCCCTCGATGTCATTCAACCCTCATGCCCCAAAACCCAAAAACA
CCTGGACGATGAAAAAGGGGGATCAGGTACTCTGGACATTCTATTCAGAAGAAGGTACATGGACCCATAACCAAGGTTTGAATGATGGCC
GGTATTTCCGGTTGGATTTATGCTCCCTCTTCCCCACCTCGATCGGAGGCGCTTCCCCTTGTAGTAATCCGTTTTCTGAAGTAAGAATGT

GCCCGGGGTTCCTTACAGATGGCTGGGATGCCAGGTGCCTGGATAGGTCGTCCCATTTCTGCCCGAAATGCGCCTGTGTGACCGCCATTG
TAGCATCCTCGTTCCGCGGGGCGGCTTGCGGGGCCAAAGAGGGCTGCGTGGGGACGGATCCGCACCTGACTATCCAAAAGGACCCAACCC
CTGGATCCACAACCTTGTTTCTCACCCTTTGCAACCCTGAGAGTAGGTTCTGGAATACGCCCCACGAATGGGGCATGAGATTAGATGGCA
GAAATAGGGCCTGGGCGGACCCTGGGATTATCTTTACCATTGCTAAGCAATCCCCCGTAACATCCTACTTGCCAATCGGCCCCCTTGGGG
AATTTGCTCTGCCCCCAAAAATCCTCCCTAGACCTCGGGCCCCGTCCTCCACTGTCCAGTCCTTACCACGGGCCCCGTCCTCCACTGTCC
AATCCCATCAGACACCCACCGGGGCAATTGGCTCTCAAGAGCCAGGTGAGGGGGAGTCCTCCAGCCAGAAACCCCCTCATCCTTCAGTCA
TGGGCCTCTTACAGGCCGTTTATGGGGTGGTCAACTCCACTCGACCAGACCTGGGCCTAAGCTGTTGGCTATGCATGGATGCCCAGCCTC
CATACTATGTTGGAGTGGCTATCAATAACTCTGTGTCTCCCACCTCCGATTCTGACAACTGTGAATGGGACCAGCCAGGGTTGACTCTTG
GGGATGTCCAGGGCTCTGGGGTTTGCTTAATCTCGGATAACACGAACCTCCATACCTCCCCATACTCGCCTGTCTGCTCCCTCAGTGTGA
TGGTCCGGTCCTCCTCCGGTTCTGCTTACTTCCCCGCCCCACCGGGCACCTGGTGGGCGTGTTTGGACGGAATCACTCGATGTGTTTCAG
CCAGAGTTTTCCTTGCTCACCCCGGTGGCCCTCTCTGTGTGCTAGTCTCCATCGTTCCCAGAGTGTCCTTGTTGCCTGGCACTGATGGGT
GGGACCACTTTTCCCTGCGGGAGGATTGGTCCCTCCATCATAAGCGGGCTGCCCCGCTGTTCATCCCCATTCTAGTGGGGTTGGGTTTAG
CGGGTTCTGCCGCCCTGGGCACTACCGCACTGGTGCGGGGGGAGGCTAGCTACAGAGAACTCAGCACCCAGGTGGATATTGACCTCACCC
ACCTTGAGCACTCCATTTCCACTCTGGAGCGACAGGTTGACTCCCTGGCGGAGATGGTCCTCCAGAACCGGAGGGGTTTGGACTTATTGT
TTCTGAGACAGGGTGGCCTCTGTGCCGCCCTGGGAGAGGCCTGCTGCTTTTATGCGAATAACTCTGGAGTTGTTCAGGAGAGCCTCTATC
TGGTGAGGAAAAATTTAGCAGGCAGGCAAAGGGAGCGTGAACGGGCCGAAACCTGGTACCAGAGTCTTTTCCGGACATCCCCGTGGTTAA
CCACGCTTGTGTCTGCCCTAGCTGGCCCCTTGTTTCTCCTCGTAGTTGCCCTGCTCGTCGGACCCTGCTTAGTGAATCGCCTCCTAGAAT
TTGTTAAGTCCCGCATCAACTCTGTTAAGCTGCTTCTCATTAGGGATCTCCACTATCAATCCCTACAAACTGAGCCCGTTGGCCGGTATG
ACGATGTCGCCACAAACGTGTCAAGGGTTTGACACTCTGTCCATAAGAAGTGGGGA

>Tacu_ERV6#ERV
tccccccctcaaagacatacgacccttagtcctaaggggttgctgagggtcaatggtccatcttctggaacggcttcatgctgacagtggg
cgatgaactcataactagatgtgttcagggtcgtatgggtgtaccgcaaaccggccgtgagctcaacagcgagatggaggtggtgtcacg
gtcggggtgatttctcgtagagtctgtaagatgggggcagtgtccccctcaggttcgtcagttgacttctgtcccgtgggttggtgctct
tccctgtggagccggaggcggagttcgccgattcggtcacaggtgtactgtgagggagtgtcctcctggggagaatcagaaaaatctggg
acaggaggtgtatagggtttaatacgtgagttatgagtccacgaggactggccctgcagtttcacagcagtgggggtagcaagtatcacc
cggaaaggacctgtccattttggctgtaaatctgccttccccaccgtctgccaagttttcaagagaacaagtgagcctggtgttatggtg
gtgcccactggagccccgtctacatcgggggctggaaaatgataatttgcatacctgcgcagggccgcgtgtaccgcttctagctgggaa
gcaaagtgccctaaagtctgtgcttctgggtccaccagaagatcagtagtgagaaaggagcaaccgtacagcagctcaaaggggctgagg
cctaggcctttgcggggtgccatacggagtcggaggagacccgtaggcaaacactgaacccaggattgtcgggtctcgaggcagagctta
gtcataatacgttttaaagtgtgattagccttctctaccttccctgaggactgtgggtgccaggcagcatgcaggtagtaggtaatccct
agggtgcgggcaatggactgggtcacttgagcggtgaaggcaggtccgttatcgctttggagagagcgggggagtccaaagcgagggatt
atttccctaaggagccacgtgcccacctctattgcgcgctcagtgcggcatgggaaggcctcaatccatcccgtgaaggtatcaatcagg
actagcaggtaccggtatccagagcaagaaggcatgtgggtaaaatccatttgccagtcctctccaggcagctggccacgccgttgtact
ggtgtcagtaagggaggtttcctcaaggccccctcagggttagctctggcacagattgggcaagcgcggcaaaccctcttaatggtttcc
cgcatacctcgtcccacaaaaatgggttctaagagggtatgtaaggcatcccgacccaggtgggtagcctgatgcattccattcaacaac
ttccaagcggaggcttggggaatgaggagctttccctctggcgtttctagccatccatccgtccggagaatgcagccttcctcgtccgct
ttgcccaactccaactcggaatataccggggcacctgcaagctctgagagggttggtacaagagccaatgcactggctgcctcggtcctg
gtgcgggctgcctgtcgagcagctgtgtccgcaaaatgatttcccttgattatttctgtattccccttttggtgtccccgacagtgaatt

acagccacctcccgggactcccaaacagctgtgagtagctggaggatttcctctgcatgtttgacaggggtgcctcaggtagtgaggaga
cctcgctctttccaaatagtcccatgagcatgtaagacatggaagccatatttggagtcagtaaatacattaagactcatccccctttccc
aattccatagctcgagatagggctactagttctgctttctgtgcagaggtccctggagagaggggctgtgcctcgataacctcatggagg
ctcactaccgcgtactctgcatatcggactccctggtccatgtaactggacccatctgtgaaccagttagcctccgtgttgtcaaggggt
aagtccctaaggtctggtctgctagagtaactctcgtcaatttcacttagtcatgaactggggctggccagtccccttccccatcttctg
gaagagggagaagggttgctgggttcaaggagttacaggtctttaaggtaaccttaggggtgtcaagaaggagggcttgatattgcagca
agcggcctcccgtgagccaacgctcacccctatttcctagcactgctaggacccgatgtggggtatagacggtaaggggttggcccaaag
tcagctttctgctctcctgtaccgccacggcagtggccgcaactgcacagagacatactggccaaccagaggtgacactatccaaccgct
tggaaaaatacgctaccgcccgaaagtctggtcccaaagattgacctaaaaccgcggcacagtggtgtcgccgctcatctacatataagg
caaagggtttggctaggtcaggcaaggctaaggctggggctgaggaaagtttccgtttgagaattcgaaatgcgtccatctgttcgccac
cccattccagcggctcgacctctggcccagcaagggcatcatacagaggcttggccaggatcccgtagtttggaatccagatgcggcaga
aactggtcattcctaaataggcacgtaactggcgcttggtctgggggctgggatttgcaggatagcgtcttttctctctgggcttaggg
tcttccctaaagggcttaatacatgacccagatatctaacctcttgtttaaaggcttgtaattttgaaggggagacttggtacccagaac
aggccaggtggttaagtgtccggacaatgtcctccttcgcgccctcctcagtcgggctgcaaatgaggagatcatccacgtattgaatga
gggtgctctcacgtaattcaaggtcacggagatctcagcccaagcattccccaaaaaggtgagggctatcccggaatccttgcggaagga
ctgtccaggtgtactgatgtcccaacttttcccttggtccagtccactcaaaagtgaaaatatactgggacccttttactcaaggggatac
aaaagaaggcatccttcaaatctgttaccgagaaccagttagtatcccctggtaccagtgcaaggagagtgtatgggttcgcgactattg
ggtgtatagggattgtagctgcgcgcaccctccgcaggtcctggacacagcgataggtcccatctggtttctttaccaggaggacagggg
tgttgcaaggggaatggcatggcacaagtaacccgtgttcaagaaatcgctcgataattggttggatccctgcctgtgcctctgttggga
acgcatactggcgttggtgtggaaaatcggagggatcctttaggcgcacctcgactgggattgcattaatggctcggcccgggagacctt
tatcccaaacgaggggtcaacttttccacaaggtactcaggtaagtccaccctattggctgtggcagtgaggatgccaactggacgga
ccaatgaaaacaacgatcccaacttggtcctgatatccctccccagtaagggtgcggggttggagggaataataaggaatctgtgttcaa
acaattcctgctggtatagacatggtaaggggtcagtttggcgacaagattttgagatgccgtcgatccccaccacccttacgtcagatg
gatgggtaggaccgttaaagctgggaaggacagaataggtggcccccgtatcgataagaaaatttataatcttaccggacacttccacac
tcaccctggtttcggccatggagatggttcctgttggggcctggccgtttcctgggccccttgagtcccgactcttagacatgccaacag
ccaagaactgaccagcctccggattggcactcgcttccagtagtgggcagtcccgcttccagtgcccgtgtctccggcagaggggcacg
gtccaggtggctggcagctgtgaccaggcctccgtgcctgggggcactcccgagcccaatggccagacctgtggcacacaaaacagggct
tcccggaccctgggccacggacaaaagtcgtggccacgttgggcgaccttccagggtgagactgacccttcatcagggccgctacaagta
actgggcctgctctctgcccttcactttggccctgcggtccctctcatcccggactgcctggttacggttgttaaaaactctaaatgcag
cctctatcaactgatccgtgggggtttgtggccccaactgctccttctgtaattttcggcggatatcttcagctgattgattaataaaat
gcatattcagaatgacacttccctctaccgagtcaggatcaagctgggtgtacgttcgaatcgcctccgccagtctgagtctaaacatag
cgggggttctcggtagagtcctgggtcacagctcggacccgatcatagttaacttgcttcctaattccttttctccatgcccgccagtagg
catgtaatcatatgatcccggagctgtctatcagtactgtctccctgataattccaatctggtttggaccgagggacagctatcgctccc
agtgggtgcctcccatgactgttcacagcagcttcatcccctgccctttcagcataatcccaaactattctcttttcctctacggtcagg
caggaggaaagcaacagattaagatcttcccatgacaggtcatactgcaaagtcaggctatggaattcgtcccgaaagcacgcggggtct
gcagaatactgtcctaatttccgttttgccatttccaatttgctcaaagagaagggagcctgtacctgcactatctttcccctctcctct
gcaacttccctcaaggggagtatcctttcccctccttctggaggggcatagtgtgtgcctccccgggtgctggaagggcttagggcccag
gaggaggtggactggtatggaggaggacataaagggttggtgtttcccattccctcagggggcagaggcggggccgaaggtacctgagat
tcccagccagagggaggggaaagccatcctcccagggccccctgcagaggagaggatcatccagtatatctgaggaatcagaccttagg

gctggacccccctggaggggcccaggaaaaatacatgcggcagtttttttccttcagggctctgggataaagccatgaaagcttggacataa
gggatttctgaccatttctctaactgcctgcaatataagtccagctggaaaatagtattataacttacggacccatattcgggccagacc
tcctgatcctctaatttatactggggccatttcgtggtgcacaagaaaatcagttgctggcattttaaatcagccaaaccaaatttgtcc
ttattcttggtgttcagcaggcaacccaatggggagttgcctggaattgaagaaacttgtcccataacgggtgaaaaacattcctccttt
ggggtgctggaaaaaagttgtccaataatgggtggaaaacgttccttttttttgaggttctgtgtctgtttcgggttaagaccgacccgg
tcggcctttcttcccccatagtgagcctggggactagagggtgcctccgggagagcagacaagaacaatgccgcttctgtgcctgcctcg
ggctatgtccgagcgccgtgtggcgtcccactggtgctggaatacccggaccgcctcctgatgagaggtccttttccctgaccgacccgg
ttggccttccttcccccctagtgtacccaggaactagagggtgcctccgcgagaacagaccagaacaatcccggttctgtgcctgtctcg
ggttacgcccgtgcaccgagtggcgtcccactagtgctgaaatgcccgaaccgtctcctgatgagaggtcctttcccagaccaacccggt
cggccttccttcccccctagtgtacccgggcactagagagtgcctccgcgagagcagacaagaacgatgccggtcctgtgcctgtctcgg
gttatgcccgagcgccgcgtggcgtcccacttgcccaagctgtggtcccgccaactcaaggcgtcccgatgagaggttccttccccggac
cacctcctgatgagaggtcctcttcccggaccgcctcctgacgagaggtccttttatcccggaccgcctcctgacgagaggtcctttcc
ccggaccgacccggtcggtccttccttccccactagtgtacccgggaactagagggtgcctccgcgataacagacaagaacaattcaaaa
aaaaagaactcaccactccgatggcttcatgctgtgctgtggtcagccaggacggcggcagtcacatctactggatctccccgtacgggc
caccaaaat

>Tacu_ERV1_LTRa#LTR
CACGTGGGGAGCGTGGGAAAGATAGGGAGTATGTGCACACAAACCGTACCCAACGGCTCAAAGCCAAACAGACTCAACAACAGGAGATAA
AGGGACCAGGGCAAGAAAAACAAGCTTGCACCTGCAAAGCTCGGAGGCAACCTCAAGGCCATATCCACAGCCAGCAATGGTTGGCAACGG
GTGGCTGGGGGCCAGCCAGAGCAAACGCTTCGTGATGGACAGAGCTGTTGCTAGGCTAGTGGCTGGAGGGTGGCGAGTGAGTGTGCTACA
TGGCCTGAGAAAATCCTGCCCCCGGGCAACGGGGGGTATAAGAGACGAACAAGACAAGGGGGGGCACGCcCACACCACACACAGGTGCTC
TCTCTCCCTCCCTTTCTCTCTCTCTCTCGTTCTCTCTCTCTTTATTAACCATGTAATCGCTGATAGCAAATAAACGGCAACCAAGCTTTG
GACCTCTGGCTGACTCTTCCTGGTGTGAACGCGCGGCCTGCGTCCGGAGACGTCCGAAGACCCGGAGAGGGTAAGAACCCGAGAGTTGCC
CCCGAAACCCCGGGGAGCAACG

>Tacu_ERV1_LTRb#LTR
CGTAGGGAGCGGGGGAAGGCGCTGAGCAAGGAACAAACAGGGAACAGGCAGTATGTGCACACAAACCGTACCCAACGGCTCAAGGCCAAA
CAGACTCAACAACAGGAGATAAAGAGACCAGGGCAAGAAAAACAAGCTTGCACCTGCAAGGCTCGGAGGCAACCTCAAGGCCATATCCGC
AGCCAGCGATGGTTGGCAACGGGTGGCTGGGGGCGAGCCAGAGCAAACGCTTCGTGACTGGACAGAGCTGTTGCTAGGCTAGTGGCTGGA
GGGCGGCGTGTGCTACATGGCAAAGCCTCGATACGCCATTCCCCGAGAAAACCCTGCCCCCGGGCAACGGAGGGTATAAGAGACGAACAA
GACAAGGGGGGGCGCCTCTCTCTCCTCCTCTCTCTCTCTCTTCTCTCTCTTTTTTAACCATGTAATCGCTGATAGCAAATAAACGGCAAC
CAAGCTTTGGACCTCTGGCTGACTCTTCCTGGTGTGAACGCGCGGCCCGCGTCCGGAGACGTCCGAAGACCCGGAGAGGGTAAGAACCCG
AGAGTTGCCCCCGAAACCGCGGGGAGCAACG

>Tacu_ERV2_LTRa#LTR
ATTGTTGTCTACCAGCAAACTTGGCATGACTGACTCCATCTTGTGCATACCAACAGTAACCCTTTATTTTGTGCAGACCGAAAGCACTCC
AGTGTTTGTAAAGTAACATTAACCAAAGTGGCCCCTCTCTGTCTGGAATGTCCCCATCCTGTGTTGACCATTATCTCCTGATAACATCTT
GTAAACAGGGATTTTAACACTAACCAAACCATGACAAACAAACGGTTGTACTTTAACTTGGGATCTAGGAATGCCAAGGCCTCACGTAnG
GCACTACTCTCAGGAATCCTTTGTGTAGCTCGGATTTGAAACCCAATAAAAGAGGAAGAAGCGGCTGGGATCGGGGCTGCTTGTCACTCG

```
TACCTCTCCCGGGAGGTGAACGGGCAGGTAGCCACTTTCCTTCTTTACTGCTCTATCTGAACTCATGTCTCCGAGTCATTTTTCCTATCT
GCGTCACCACCCTGGGTACAAGAACCCTGCGGCCGATTTACACCGnGTTAACCTATTTTGCACCCTACTTGTCGATAACAn

>Tacu_ERV2_LTRb#LTR
tTTGTTGTCCACTAGCGGACTTGACATGACTGACTCCATTTTGTGtATGCTGACAGGAACCCTCCATTTTGTGCAGGCTGAGAGAACAAC
TCCTTTTTGTATTGTCTGCAACAGATGCCTTCAAGGCCATTAACAAGTAACACcTATCTATGTAAGGTCgTAGGGCAGATGACATCCTGC
ACAAGACAGTGACAACAGAAGATAACgAGAATTTACACACCTATTTATgCAAGGCCATAGGGCAGATAACAACAACCTTCACAAGGCAGT
gAAAACAGAGAATTTATAGCATCCTGTCGGTCCTAATTGGATTCCTGAAATTCCTAAATGCTCcGCTCCCATGTTGCTGTAACcCAATAA
AAGACACAGTGAGAATGGGATCGGGGCTGCTTGTCACTCGTACCTCTCCCGGGAGGTGAACGGGCAGGTAGCCACTTTCCTTCTTTACTG
CTCTATCTGAACTCATGTCTCCGAGTCATTTTTCCTATCTGCGTCACCAcCCTGGGTACAAGAACCCTGCGGCCGATTTACACCGATTAA
CCTATTTTGCACCCTACTTGTCGATAACAn

>Tacu_ERV2_LTRc#LTR
nTTGTTGTCCACTAGCaGACTTGACATGACTGACTCCATTTTGTGTATGCTGACAGTAACCCTCCATTTTGtGCAGGCCAaGAGAACAAC
TCCtTTTTTTTGTATTTTATGCAAGGCTCAGCAACAGATGTCTTCAAGGCCAGTAACACCTATCTATGCAAGGCCATAGGGCAGATAACA
ACAACCTGCACAAGGCAGTGAaAACAGAGAATAATGAGAATTTaTAACATCcTGTCGGTCCTAATTGGATTCCTGAAATTCCCAgTTGCT
CCGCTCCCATGTTGCTGTAACTCAATAAAAGACACAGTGAGAATGGGATCGGGGCTGCTTGTCACTCGTACCTCTCCCGGGAGGTGAACG
GGCAGGTAGCCACTTTCCTTCTTTACTGCTCTATCTGAACTCATGTCTCcGAGTCATTTTTCCTATCTGCGTCACCACCCTGGGTACAAG
AACCCTGCGGCCGATTTACACCGATTAACCTATTTTGTACCCTACTTGTCGATAACAn

>Tacu_ERV5_LTRa#LTR
nTGTTGTGATCGGGTTTGTCTGATCTCCATTTTGTTTTCCCATCCCTTCCTCCCATCCTTTTGTACACCCTTCTTCACGGAAGAAGAATG
CCCGCCAAAACTTCCCGCCACAAAACCTGCCTAACAAAAGCCCTCCCCCACCCTGACCTGCCTGACAAAAGCCCTCCCCCACCCTGTTCA
GGGCCGTTTATGACCCCAAGCGAAAATTAACCAATTGCAAGCCTCTGTTTCTGTAAACTTTCCACTGATTCGCAACTCCTGACCGTTAGC
CTTGCGTTTTGTGCCTTTATAAACTCCGGCCCTACCCCTGATCGGGGCTGCTTGATCTGGGTTCCTGGCCCCTTGCAGCCGCGCnCTAAT
AAAATCCACTTCTAAATTACTACCTGGGTCTCACTCGCTGATTCTCGGCACAACA

>Tacu_ERV5_LTRb#LTR
aTGTTGTGATCGGGTTTGTCTGATATCCAnTTTTGTTTTCCCATCCCTTCCTCCCATCCCTTTGTACACCCTTCTTCACGGAAGAAGAAT
GCCCGCAAAAACTTCCCGCCACnAAAACCTGCCTAACCAAGGCCTnCCCCCACCCTGACCTGCCTGACAAAAGCCATCCCTGTTGTCCAC
TAGCGGACTTGACATGACTGACTCCATTTTGTGCAGGCTGAGAGAACAACTCCTTTTTGTATTGTCTGCAACAGATGCCTTCAAGGCCAT
TAAACAAGTAACACTTATCTATGTAAGGTCCTAGGTCAGATGACATCCTGACAAGACAGTGACAACAGAAGATAACAGAATTTACACACC
TATCTATACAAGGCCATATGGCAGATAACAACAACCTTTACAAGGCAGTAAAAACAGAGAATTTACAGCATCCTGTTGGTCCTAATTGGA
TTCCTGAAATTCCTAAATGCTCCCTCCCATGTTGCTGTAACTGAATAAAAGAnnCGCAGTGAGAATGGGATCgGGGCTGCTTGATCTGGG
TCCCTGGCCCCTTGCAGCnnCGCGCGCTAATAAAATCCACTTCTAAATTTCTACCTGGGTCTCACTCGCTGATTCTCGGCACnAACA

>Tacu_ERV3a_LTR#LTR
TGTGGCAGACCTAACATCTGGGCCTAGGTAGTAAGCCAGGCCACCGGCCTTCCGTGCTCACAGGAAGACCTACAAGGAGCAATGAAGGAA
GTCAGGCTGCTAGATAAATAACTTGGCGCTGTTCTGAGAATTGCCTTGGGAACGTGCGATCTAGGCATGAGCTCAGCCTTGCTGATAACT
```

AGCCTTGTGATGAGCTCAGTCTTGCTGATGAGCTCATCTTGCTGATAAGCTCAGCCTTGCTGATAAGCTTGCAGTCATCCGCCTGTCCCC
TGACTCCCTTGAGGCTCCTCACTGTATAAAATCCTCCACAGTGTGCTGAATAAACAGTCTCCTTGTCCACCTTGAGACTCGCCTGGCCTG
GTTCTTCCATTGCGGTGCCCGTCTCCCCGCTGCGCCGGACGCGTGGAGGCAGACGGCAACA

>Tacu_ERV3b_LTRa#LTR
CATGTTTGAATTGTTTCAAAAACAAAGGGGGACTTGTGGCAGACCTAACGTCTGGGCCTAGGTAGTAAGCCAGGCCCCTGGCCTTCCGTG
CTCACAGGAAGACCTACAAGGAGCAATGAAGGAAGTCAGGCTGCTAGATAAATAACTTGGCGCTGTTCTGAGAATTGCCTTGGGAACGTG
CAGATCTAGGCATGAGCTCAGCCTTGCTGATAAGCTCAGTCTTGCTGATGAGCTCAATCTTGCTGATAAGCTCAGCCTTGCTGATAAGCT
CAGCCTTGCTGATAAGCTTGCAGTCATCCGCCTGTCCCCTGACTCCCTTGAGGCTCCTCACTGTATAAAATCCTCCACAGTGTGCTGAAT
AAACAGTCTCCTTGTCCACCTTGAGACTCGCCTGGCCTGTGTTCTTCCATTGCGCGTGCCCGTCTCCCCGCTGCGCCGGACGCGTGGAGG
CAGACGGCAACA

>Tacu_ERV3b_LTRb#LTR
TGTGGCAGACCAAACATCTGGCCCATCAATTACTGGGCCTAAGTAGTAAGCCAGGCCTTCCGTGCTCACAGGAAGACCTACAAGGAACTG
AGAGGAGAGCTGTGGACAGGCAAGAGTTACTTGTTATCGGCCCTACGGCCTTGCATAGACGGGTGTTGCTTGTTACTGGTTTTGAAGGTT
ACAGAATGGCTGTGTTAGTTATAAACACAGCAACATGGAAAAATACAGAATGTGACTGAAAGATATAAAAGTCTTGCTGCTTAACCCAAT
AAACAGTCTCCTTGTCCACCTTGAGACTCGCCTGGCCTGCGTTCTTCCATTGCGAGTGCCCGTCTCCCCACTGCGCTGGACGCGTGGAGG
CAGACGGCAACA

>Tacu_ERV4_LTRa#LTR
aTGTGAGGCTTGGAGAGGCTTGAGTCTCTACAACTAGACAGACCCCTATCTTGGGCAACCGGGCCAGTAACACAAGGAAGGAACATTTGC
ATAACCGCTCCTCAGACCCCTATCTTGGGCAACCGGGCCAGTAACACAAGGAAGGAACATTTGCATAACCGCTCCTCAGACCCCTATCTT
GGGCAACCAGGCCAGTAACACAAGGAAGGAACATTTGCATAACCGCTCCTCAGACCCCTATCTTGGGCAACCAGGCCAGTAACACAAGGA
AGGAACATTTGCATAACCGCTCCTCAGACCCCTATAAAGGGTGTAAAGTTGTGCTCGTGGGTGCTGCCGCTTCCCAAGCCTTACTATAAG
GCGGTAGCCCAGATTTAAATCTGCAATAAAGCTGCGACCTGCCCTCTTTGGCGCTGGGCAcGTCATCTAAATTTTGTCCGGTTGTCGTGC
GATCCAATTTGTTGGTGGGTGGGATCCCGGGCCCTGGAGACCTTGATCAGGTTTCCTTACAACA

>Tacu_ERV4_LTRb#LTR
TGTGAGGCTTGAAGAGGCTTGAGTCTCTACAACTAGACAGACTCCATCTTGTTAGACAGACTCCATTTTGTGGAAAATGCCTTGTCACGC
GGTTGGGCAGGGAGCCAGAACTTCTGGCCTGTAGCCCGGCCCCAACGCCCTTATCGGGAAGCACCTGGCAGGGCGGTTACCGTCAGCCTA
ACGGCCGGGAAAGCCCCTTGGGCAACCAGGCCAGTAAAACAAGGAAGGAACATTTGCATAACCGCTCCTCAGACCCCTATAAAGGGTGTA
AAGTTGTGCTCGTGGGCGGCTGCCGCTTCCCAAGCCTTACTATAAGGCGGTAGCCCAGATTTAAATCTGCAATAAAGCTGCGACCTGCCC
TCTTTGGCGCTGGGCAGGTCACCTnAAATTCTGTCCGGTTGTCGTGCGATCCAATTTGTTGGTGGGTGGGATCCCGGGCCCTGGAGACCT
TGATCAGGTTTCCTTATAACA

>Tacu_ERV6_LTR#LTR
TGATGGGGGCCCCTCATGCACCATGGAAAGGTTGCAGATATTGTTTGAATGTACGCGGTTCTGTTGCCATATCTGGTATGGGCCTTAGAC
CCTGCAGCCCACTGGCTGGGTCCAGGAACCAAACCAAATAAGGAGGTAGACTCCGCCCAGCTCTGTACCAAATCTGCACCAATCCAGTCC
CCACTGTTTTCAAATCTACCAATCACTGTGATCAGCAACAGCAGTATATAAGCCCATTGTATCGGGCACTCGGGGCCTTTTCCCTTAAGG

```
AAATGAGCCCGCCGGGTGCGTTACCCCCTATTCGGTCTTCGGCCTTACCAATAAAACTTTATTAAAACTTTCGGCAGTCACATGCTGTCT
GACTAATTCTTTAGTCGCCCGGCGACTTGGTGGCCATCCGGAACAACCGCCGCCATC
```