Evolutionary histories of transposable elements in the genome of the largest living marsupial carnivore, the Tasmanian devil

Gallus, S; Hallström, BM; Kumar, V; Dodt, WG; Janke, A; Schumann, GG; Nilsson, MA.

Supplementary information

Supplementary Table 1. Genomic location of L1-1_SH copies longer than 6,000 nt.

Supplementary Table 2. Presence-absence table for the phylogenetic screen in Dasyuromophia.

Supplementary Table 3. ERV1 poly(A) element characteristics.

Supplementary Table 4. DNA transposons in the Tasmanian devil genome.

Supplementary Table 5. Scientific and common names of the investigated marsupial species.

Supplementary Table 6. Primers for the phylogenetic retrotransposon marker screening.

Supplementary Table 7. OC-1 and hAT-1_MEu transposon primers.

Supplementary Figure 1. L1 length graphs of RTE and L1 in opossum and Tasmanian devil.

Supplementary Figure 2. Conservation plots of ORF2 in opossum and Tasmanian devil.

Supplementary Figure 3. Screen for identical SINE screen in opossum and Tasmanian devil.

Supplementary Figure 4. RT-PCR of the OC1_Das DNA transposon in dunnart.

Supplementary Figure 5. Phylogenetic tree of OC1_transposase ORFs from different mammals.

Supplementary Figure 6. Genomic PCR of OC1_Das and hAT-1_MEu DNA transposons.

Supplementary Figure 7. Phylogenetic tree of the 17 individual full-length L1 copies.

Supplementary dataset 1. Fasta sequences of the phylogenetic retrotransposons used for the evolutionary analysis.

Supplementary dataset 2. Fasta sequences of the consensus sequences of the ERV1 element.

Supplementary Table 1. Characteristics of L1-1_SH copies longer than 6,000 nt with intact 5' and 3' UTRs and no nested integrations of transposable elements.

Scaffold	Chr	Location	TSD	Size (nt)
GL835730	1	chr1_GL835730_random:4900-11929	AAAAAT/AAAAAT	6,642
GL835482	1	chr1_GL835482_random:18539-25398	GATCAGT/ GATCAGT	6,860
GL841388	2	chr2_GL841388_random:525,411-532,547	TATCTA/TATCTA	6,739
GL842553	2	chr2_GL842553_random:c12580-5862	AAGATGTTATTACAG/ AAGATGTTATTACAG	6,719
GL841234	2	chr2_GL841234_random:c698412-691981	3'UTR N	6,432
GL841955	2	chr2_GL841955_random:19706-26707	ATAAAACAA/ATAAAATAA	6,606
GL841930	2	chr2_GL841930_random:72123-78892	CCCAGA/CCCAAA	6,770
GL841199	2	chr2_GL841199_random:c968408-962018	ACATTTAAGGA/ATATTTAAGGA	6,391
GL849815	3	chr3_GL849815_random:94666-101246	AAAGAATCCATT/AAAGAACCC ATT	6,581
GL857653	4	chr4_GL857653_random:27886-34843	AAAAGCAACAATA/AAAAGCAA CAATA	6,574
GL857349	4	chr4_GL857349_random:8002-14653	TCATTTA/ TCATTTA	6,652
GL859690	4	chr4_GL859690_random:c8386-1489	-/-	6,898
GL860397	4	chr4_GL860397_random:c9175-2429	AAGAAAAAA/AAGAAAAGA	6,747
GL861587	5	chr5_GL861587_random:c196766-190082	GATAATTT/GATAATTT	6,685
GL861677	5	chr5_GL861677_random:805749-812510	AGACATGAATATGG/AGACATG AATTTGG	6,762
GL861642	5	chr5_GL861642_random:c234073-227515	TCAGTCA/TCTGTCA	6,559
GL865003	6	chr6_GL865003_random:c139531-132898	AAGAACATAAAGAT/ AAGAACATAAAGAT	6,667

Chr: chromosome, TSD; target site duplications. -/-:not found.

Supplementary Table 2. Presence absence table of phylogenetically informative retrotransposon insertions for Dasyuromorphia.

	12 WS	20a WAL	20b WAL	100b L1	29 WS Δ	6 WS	5 WS	100c L1	7 WS	15 WS	16 WS	19 WS	8 WS	97 L1	44 WS	30a WS	81 L1	87 L1	30b ERV	83 L1	95 L1 Δ	100a L1
ShaDB	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Sha*	?	+	+	+	+	+	+	+	+	?	?	+	?	?	?	+	?	?	+	+	+	+
Dge	+	?	?	+	+	+	+	+	+	+	+	+	+	+	+	+	?	+	+	+	-	-
Pap	?	?	?	+	+	+	+	+	+	+	?	+	?	+	?	+	+	+	+	-	-	-
Pta	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	+	D	?	+	?	?	?
Afl	?	?	?	?	+	+	+	?	+	+	?	+	?	+	?	+	+	+	-	-	-	?
Pin	?	?	?	?	?	?	?	?	+	?	?	?	?	?	+	+	?	?	-	?	?	?
Scr	?	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-
Mfa	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pgu	?	-	-	-	-	?	?	-	?	-	-	?	?	-	-	-	?	?	-	?	-	-
Iob	-	?	?	?	-	-	-	?	-	-	-	-	-	-	-	?	?	-	?	?	?	?
Mla	?	?	?	?	?	?	?	?	?	+	?	?	+	?	?	?	?	?	?	?	?	?
Nty	?	-	-	-	-	-	?	-	?	?	?	?	?	?	?	?	?	?	?	?	?	-
Meu	?	-	-	-	?	-	-	-	-	-	-	-	-	?	?	-	-	-	-	-	-	-
Mdo	-	-	-	-	?	?	-	-	?	-	-	?	-	?	?	-	-	?	-	-	-	-

For each locus (indicated by the number in the top row) selected species (three letter code is explained below) were tested for presence (+) or absence (-) of a specific element (**WS** = *WSINE1*, **L1** = *LINE1-1-SH*, **WAL** = *WALSI1a*, **ERV** = *New ERV Element.*). The addition "a", "b" and "c" to the locus number indicate elements from a locus containing two or more informative insertions. The "?" indicates lack of sequence information for this locus, as in this case the species was tested for amplicon size but not sequenced. A "Δ**"** marks a large indel and "**D"** stands for deletion of the element and part of the genomic flank sequence. Sequences for the *Sarcophilus harrisii* genome individual (ShaDB) were directly taken from the database. For some loci an individual (Sha*) from Copenhagen zoo was sequenced to verify the genome sequence. The background colors highlight the phylogenetic grouping the marker supports.

Sha – *Sarcophilus harrisii* (DB=database; *Copenhagen zoo), **Dge** – *Dasyurus geoffroii*, **Pap** – *Parantechinus apicalis*, **Afl** – *Antechinus flavipes*, **Scr** – *Sminthopsis crassicaudata*, **Mfa** – *Myrmecobius fasciatus*, **Pgu** – *Perameles gunnii*, **Iob** – *Isoodon obesulus*, **Mla** – *Macrotis lagotis*, **Nty** – *Notoryctes typhlops*, **Meu** – *Macropus eugenii*, **Mdo** – *Monodelphis domestica.*

Supplementary Table 3. Descriptive information of the 5 copies of a poly(A)tail ERV1 element as well as its fasta sequence in the Tasmanian devil genome.

1	Chr1 GL834777	188 nt	forward	CNOT1 intron
	poly(A)tail: AAAAAAAAAAAAACAAAACAAAAA			24 nt
	TSD: ATGTAA/ACGTAA			6 nt
>1-chr1GL834777 ATGTAAGCAGAGGTTCCAAACATGCCAGCCTTGCTCCTGGCATGCCAGGACCCTCTGTCCACTGGAACCCTGTTT CCGGTGCCCTTCTTATCTCTATCTTCACCTATTTCCTTAACTATACTTTAAACTTACTTCCAAACCCCATAATAAACC TCTTTTATATCAATCTTAAAAAAAAAAAAACAAAACAAAAAACGTAA				
2	Chr2_GL844869	335 nt	reverse	intergenic
	poly(A)tail: AAAAAAAAAAAAAAAAAAAAAAC			23 nt
	TSD: AGGACCT/ AGGACTT			7 nt
>2-chr2GL844869 AGGACCTTTTGATTCACAAATCCTGGAACCTTTGAATTCCAATAGAAGATCAGGACCTGTCCCAGCCCCACCCG GATCTGAGCCAGCTTGGGACTCCACCCACAGGCCTCTCTAGTTAGATCTCCCATTTTAAAAGAGCCAAGCTGGTAC CCCCTCTTTGCAGAGGTCCCAAACATGCCAGCCTTATGCCTGGCATCCCAGGACCTTCTGTCCGCTGGAATTGTGTC TGGTGCCCTTCTTATCTCTACCTTTACCTATTTCCTTAACTATACTTTAACCTTACTTCCAAACCCCATAATAAACCT CTTTTATCAATCTTAAAAAAAAAAAAAAAAAAAAAACAGGACTT				
3	Chr4 GL856963	332 nt	reverse	intergenic
	poly(A)tail: AAAACAAACAAAAAAACCAAACAAAAAAAAAAC			33 nt
	TSD: GAGATGATTCA/ GAGATGATTCA			11 nt
>3-chr4GL856963 GAGATGATTCACCTAGTCCCTTTGAATTCCAATAGAAGATCCGGACCTGTCCCAGCCCCACCCAGATCTGAGC CAGCTTAGGGCTACACCCATAGGCCCTTCTAGCTACATCTCCCATTATAAAAGGGGCCAAGTTGGGACCTCCTCTT TGCAGAAGTCCCAAACATGCCAGCCTTATACCTGGCATGTCAGGACCCTCTGTCCACTGGAACTCTTGTCCAGTGC CCTTCTTATCTCTACCTTTACCTATTTTCTTAACTATACTTTAACCTTACTTCCAAACCCCATAATAAACCTCTTTTA TCAACCTAAAACAAACAAAAAAACCAAACAAAAAAAAAAACGAGATGATTCA				
4	Chr1_GL835600	169 nt	reverse	intergenic
	poly(A)tail: AAAAAAGAAAGAAAGAAAGAAA			22 nt
	TSD: AGAACTGCATGGA/ AGAACTGCATGGA			13 nt
>4-chr1GL835600 AGAACTGCATGGACAGAGGGTCGAAACATGCTACACCCCAAGGACCCTCTGTCCACTGGAATCCTGTTTCCA GTATCCTCTTATTTCTACCTTCACCTATTTCCTTAACTATACTTTAACCTTACTTCCAAACCCCATAATAAACCTCTT TTATCAATCTAAAAAAGAAAGAAAGAAAGAAA AGAACTGCATGGA				
5	Chr4_GL856888	197 nt	forward	intergenic
	poly(A)tail: AAAAAAAAAAAAAAGAAAAAAAAA			24 nt
	TSD: AAGAAATGACCAGCAGG/AAGAAATGACCAGCAGG			17 nt
>5-chr4GL856888 AAGAAATGACCAGCAGGGGACCCTCTCTTTGCAGAGGTTCCAAACATGCCAGCCTTATGCCTGGCATGCCA GGACCCTCTGTCCACTGGAACCCTGTTTCCAGTGCTCTTCTTATCTCTACCTTCACCTATTTCCTTAACTATACTTTA ACCTTACTTCCAAACCCCATAATAAACCTCTTTTTCAATCTAAAAAAAAAAAAAAGAAAAAAAAAAGAAAATG ACCAGCAGG				

Target site duplications are indicated in red.

Supplementary Table 4. DNA transposons longer than 1,500 nt in the Tasmanian devil genome.

Name	>1,500 nt	Total copy number/genome
OC1_Das/hAT-2	132	5,100
hAT-1_MEu	34	31,283
Mariner3_MD/MarsTigger8	10	1,540
Mariner1_MD/ MarsTigger6	1	5,921
Charlie1	1	1,882
Charlie1b_Mars	15	2,372
Charlie24	5	159

When the element name in RepBase and Repeatmasker is different, both names are given.

Supplementary Table 5. Scientific and common names of the investigated species.

Common name	Scientific name	Order	Abbreviation
Tasmanian devil	*Sarcophilus harrisii*	Dasyuromorphia	Sha
Western quoll	*Dasyurus geoffroii*	Dasyuromorphia	Dge
Dibbler	*Parantechinus apicalis*	Dasyuromorphia	Pap
Mardo	*Antechinus flavipes*	Dasyuromorphia	Afl
Brush-tailed phascogale	*Phascogale tapoatafa*	Dasyuromorphia	Pta
Fat-tailed dunnart	*Sminthopsis crassicaudata*	Dasyuromorphia	Scr
Planigale	*Planigale sp.*	Dasyuromorphia	Pin
Numbat	*Myrmecobius fasciatus*	Dasyuromorphia	Mfa
Eastern barred bandicoot	*Perameles gunnii*	Peramelemorphia	Pgu
Southern brown bandicoot	*Isoodon obesulus*	Peramelemorphia	Iso
Greater bilby	*Macrotis lagotis*	Peramelemorphia	Mla
Marsupial mole	*Notoryctes typhlops*	Notoryctemorphia	Nty
Tammar wallaby	*Macropus eugenii*	Diprotodontia	Meu
Wallaroo	*Macropus robustus*	Diprotodontia	Mro
Honey possum	*Tarsipes rostratus*	Diprotodontia	Tro
Common ringtail possum	*Pseudocheirus peregrinus*	Diprotodontia	Ppe
Common brushtail possum	*Trichosurus vulpecula*	Diprotodontia	Tvu
Wombat	*Vombatus ursinus*	Diprotodontia	Vur
North American opossum	*Didelphis virginiana*	Didelphimorphia	Dvi
South American opossum	*Monodelphis domestica*	Didelphimorphia	Mdo

Supplementary Table 6. Primer sequences for the amplification of the phylogenetic retrotransposon markers.

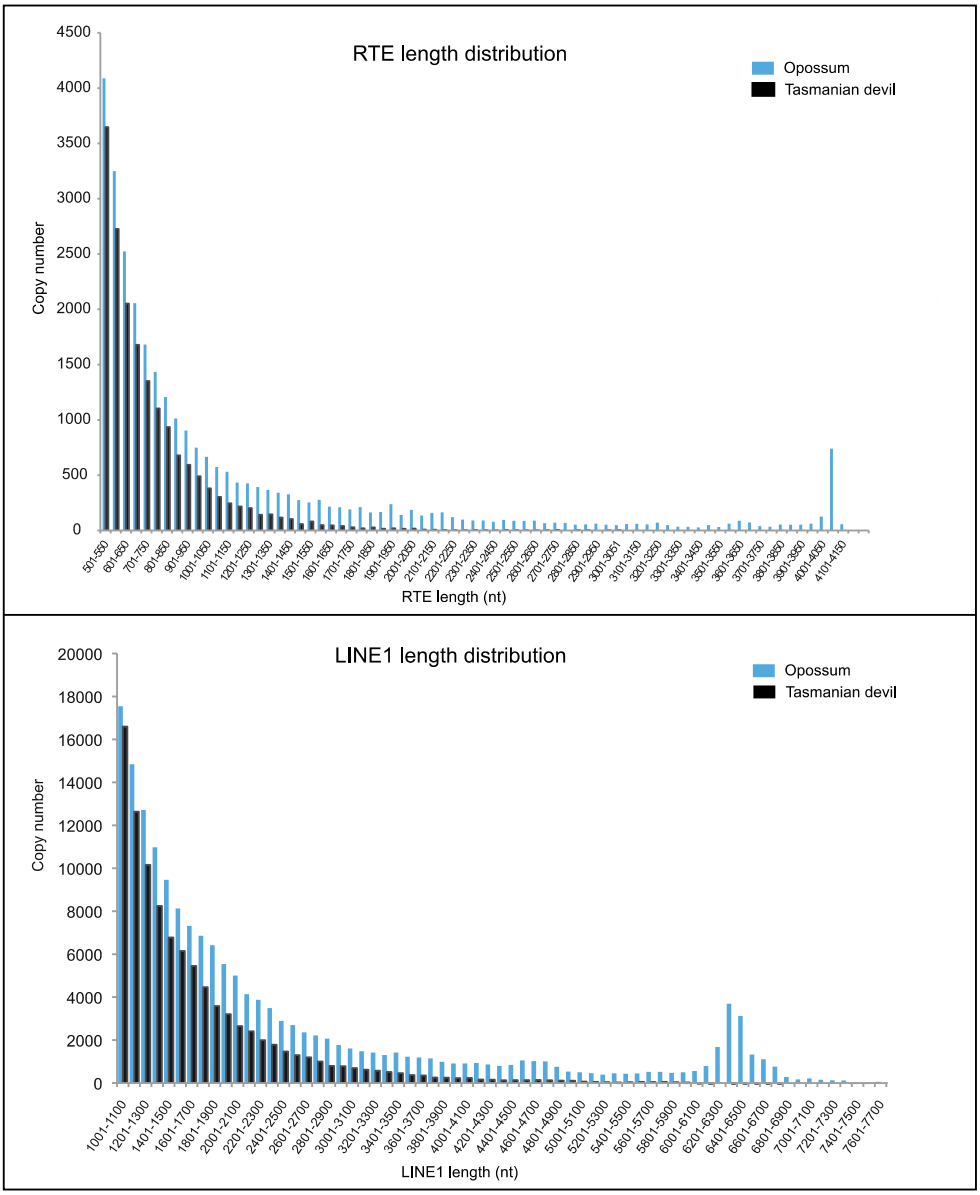
Marker	Forward	Reverse	Length	Gene/Chr.
05	CTCACCCTGTCCTACAACCG	GGARTGGGTAGTCATAAAGGCCC	812	chr1_GL834472 FURIN
06	GGCCAAGTTTCTGCAGGAAGC	CACCAGCTCCATGACGATGTAG	860	chr1_GL834472 FES
07	GGGGGCATGAAACACAACAC	CCCTCAGCAAATGGTTCCAC	920	chr1_GL834506 SLC12A6
08	CAACTGCAGACTGGTGCCC	GGTCAAAGTCATCTTTGCAGGC	666	chr1_GL834508 THBS1
12	CTTGCAATCCTGGACCATCCC	CCCAAAATTCGAGGACATGATGG	862	chr1_GL834593 DICER1
15	GAAGCACAAGGTCAGCAGG	GCCTCAACCGCTGTTGTTC	735	chr1_GL834651 GLE1
16	GGCACAGAGAGGGCATCC	CCAAGGCCTGGAGCTGAAG	713	chr1_GL834652 Novel
19	CTGCTGCTGGCTTACCCC	CACCAAACCAGCCTGACG	573	chr1_GL834713 ATP6V1B2
20	CTGACTAATGAGCAGGTATGGCAG	CCACCTACTCGAACAATGCTGG	927	chr1_GL834670 ZNFX1
29	CAGCCTCCTGGAGAACATGC	GAGGACTTGCAAAATGATGCAC	733	chr1_GL834769 CDK10
30	GTGGAACTCAGCAAACCTATGC	CCAGCAGCACTCGTAAAGTACCC	700	chr1_GL834777 CNOT1
44	GATTCCCAGAAATCACTGC	TTCCGTTTTCCTTCCTATG	642	chr2_GL841447 SRRD
81	GACTAAAGTTTTTGCTCACCAAGG	GTCCAGAGCAGCAAGACAGG	1,042	chr1_GL834508 CASC5
83	GGACGTGGCATGTGCATG	GTATCCAGGGTGGGTCACC	1,092	chr1_GL834625 TBC1D9B
85	CGACAGTGCATTAGCTGGACC	GAAACCAACAATCAGGTGCAAA	1,508	chr2_GL841614 ADCY1
87	GCACCATCCACTGCCTCC	GTCCTGCGACGTCGTCTCC	1,014	chr3_GL849766 RTN2
95	GATCACATGCCTGAAGG	CAGATTCTTCATGATCAGC	1,105	chr1_GL834643 RHPN2
97	GGCAGTGCAAAGATCTGAGTCA	GAGCCAGAGAGATTCATCTGAGG	1,031	chr1_GL834650 NLRC5
100	GGCCACAGGAGCACTCC	GGGGCATACAGTTTCAAAAT	780	chr1_GL834659 Novel

The length of the amplification product in Tasmanian devil together with the gene and chromosome location of the marker are indicated.

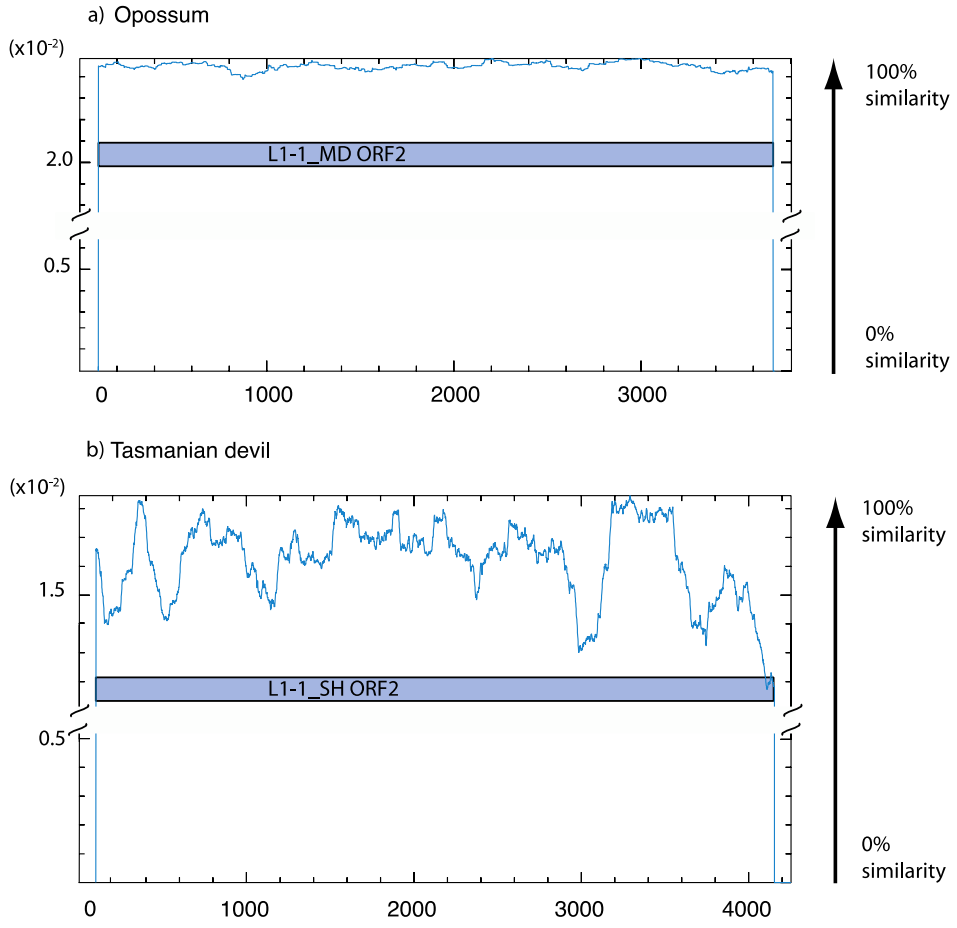
Supplementary Table 7. Primers for amplification of the DNA transposons OC1_Das and hAT-1_MEu.

Primer name	Sequence	Position
OC1ShaF	ATGATGTCAAGAAAAAGAAAAATTGACTC	601
OC1ShaF2	GTGGAGGGAGAATTTGTTAAAGAATG	944
OC1ShaR	GCCTGCTTTTGGCTAATGAGATG	2,381
OC1ShaR2	CTCAGACAGAAATTCCTAAAATTGTCTG	1,569
hATMEuF1	GCCAAAACTTCATTACAAATT	1,871
hATMEuF2	GGCCATTTCAGAAATTGTTGGAAATTC	390
hATMEuR1	GTATCAAATTAGTATATCTTTGGA	2,359
hATMEuR2	CAGAGGTAAGGTGGCTCTCAA	1,001

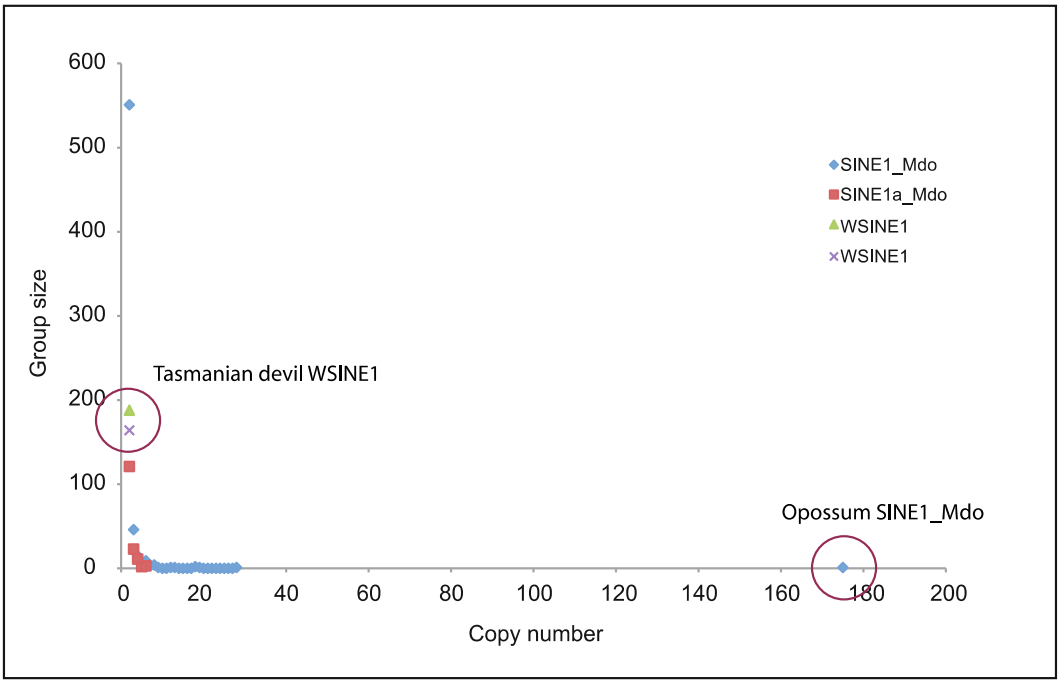
The position refers to the consensus-sequences of OC1 and hAT-1_MEu.



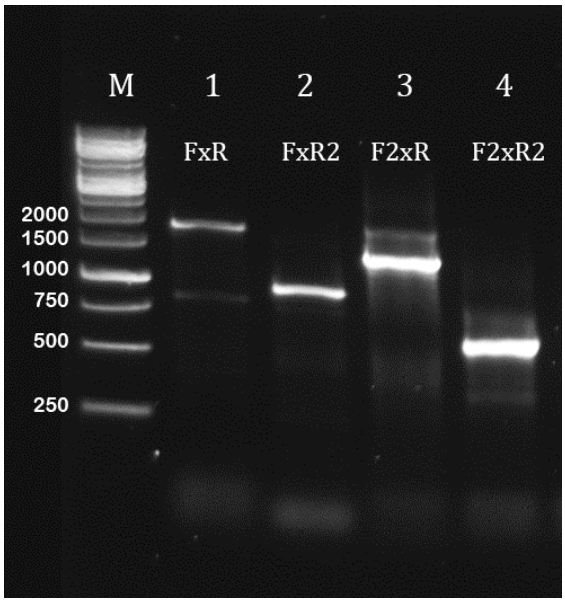
Supplementary Figure 1. RTE and LINE1 length distributions plotted against copy number in the opossum (blue bars) and Tasmanian devil (black bars) genomes. The majority of the RTEs and LINE1s are short, 5' truncated copies. The peaks for opossum around 4,000 nt and 6,400 nt indicate the full length RTE and LINE1 elements. The 17 LINE1 copies in the Tasmanian devil genome that exceed 6,000 nt are not visible due to the scale. In the opossum genome, 735 RTE copies were 4,050 nt long, corresponding to RTE_Mdo (4,088 nt) (Repbase ID RTE-1_MD), while only three were longer than 3,000 nt in the Tasmanian devil genome. The retrotranspositionally active L1-1_MD (opossum), which has a consensus sequence length of 6,356 nt (Gentles and Jurka 2005), represents the main peak at 6,300-6,400 nt, with a total of 6,815 copies.



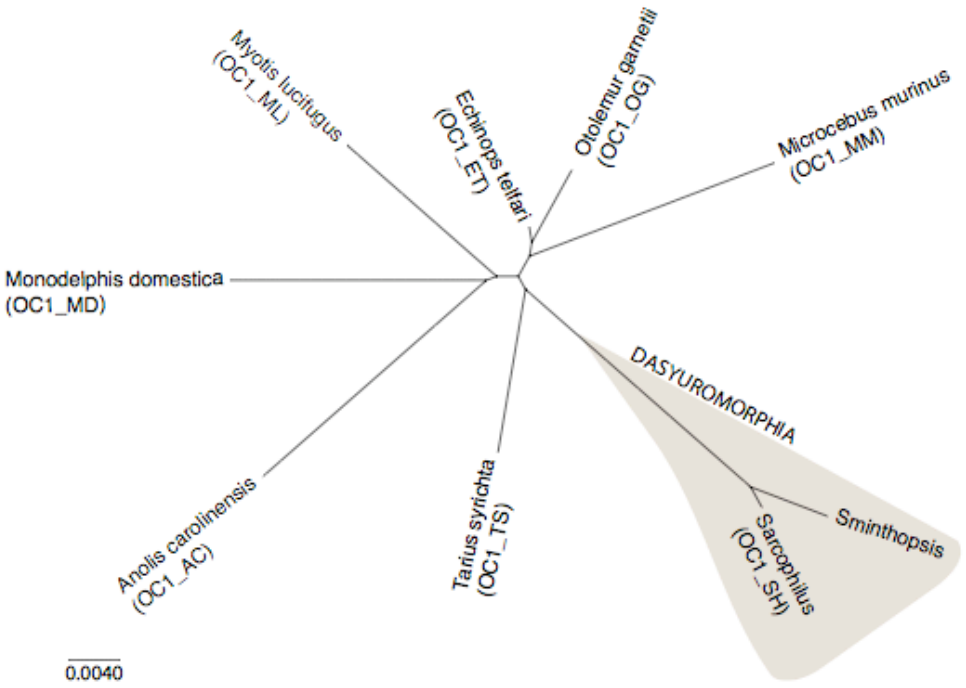
Supplementary Figure 2. Conservation plots for a) opossum L1-1_MD ORF2 and b) Tasmanian devil L1-1_SH ORF2 from the EMBOSS program PlotCon. At the X-axis the position in the alignment is shown while the Y-axis shows the similarity at each position in the alignment using an average similarity score. The average similarity score is calculated for each position as the average of all the possible pairwise scores of the nucleotide at that position. The program calculates a specified similarity matrix to use for gathering the pairwise scores. The blue lines indicate the amounts of variation found in the respective alignments. The greater deviation from the top line the higher the variation.



Supplementary Figure 3. Plot of the group size distributions of identical SINEs in the opossum (SINE1_Mdo) and Tasmanian devil (WSINE1) genomes. In the Tasmanian devil genome, maximally two identical SINEs were found in each group (red circle). In opossum SINE1_Mdo is the only SINE that has one large group (175 copies) of potentially recently propagated copies (red circle).

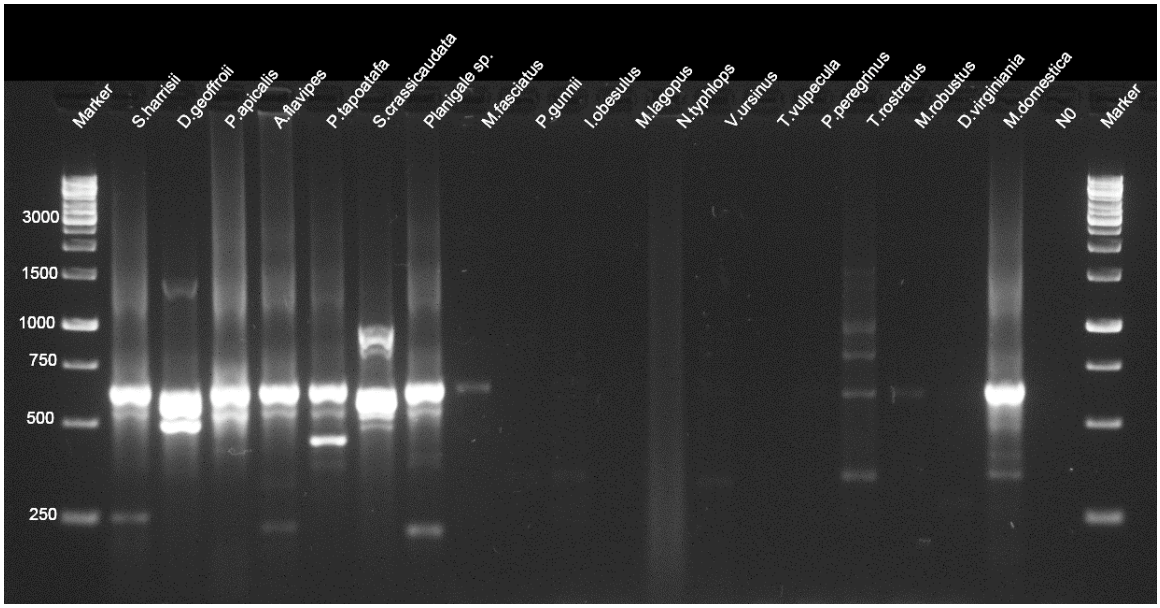


Supplementary Figure 4. RT-PCR of the OC1_SH DNA transposon in the fat-tailed dunnart. The amplified region (1) covers 1,909 nt of the transposase ORF from the ATG start codon to 10 nt from the stop codon. Four PCR (1,2,3,4) reactions were run for each species using different primer combinations. For primers see Supplementary Table 7. M: marker GeneRuler 1Kb.

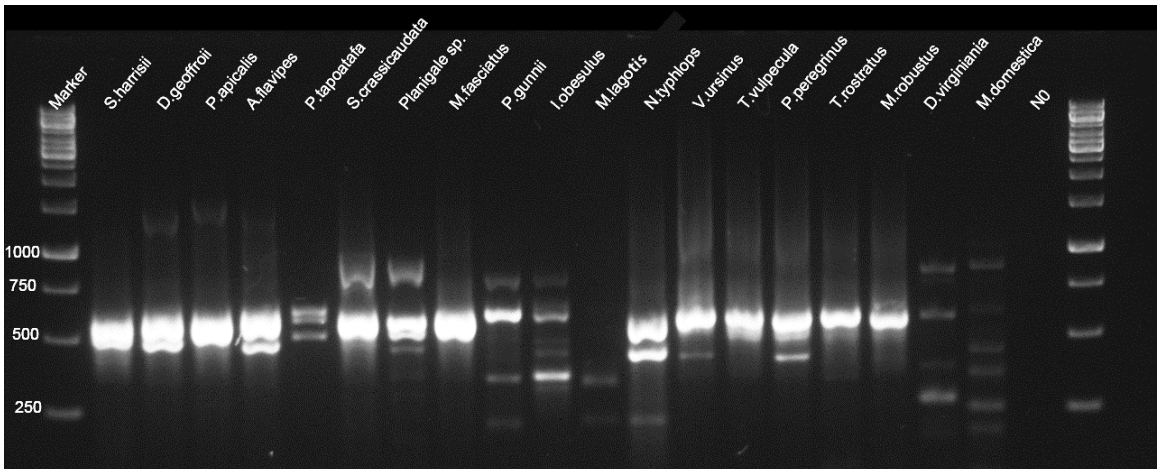


Supplementary Figure 5. Phylogenetic analysis of the DNA transposon OC1 transposase ORFs in the Tasmanian devil and fat-tailed dunnart genomes compared with OC1 from other vertebrates. The OC1 from Tasmanian devil and fat-tailed dunnart cluster together to the exclusion of the other species. The scale bar indicates amount of substitutions per site.

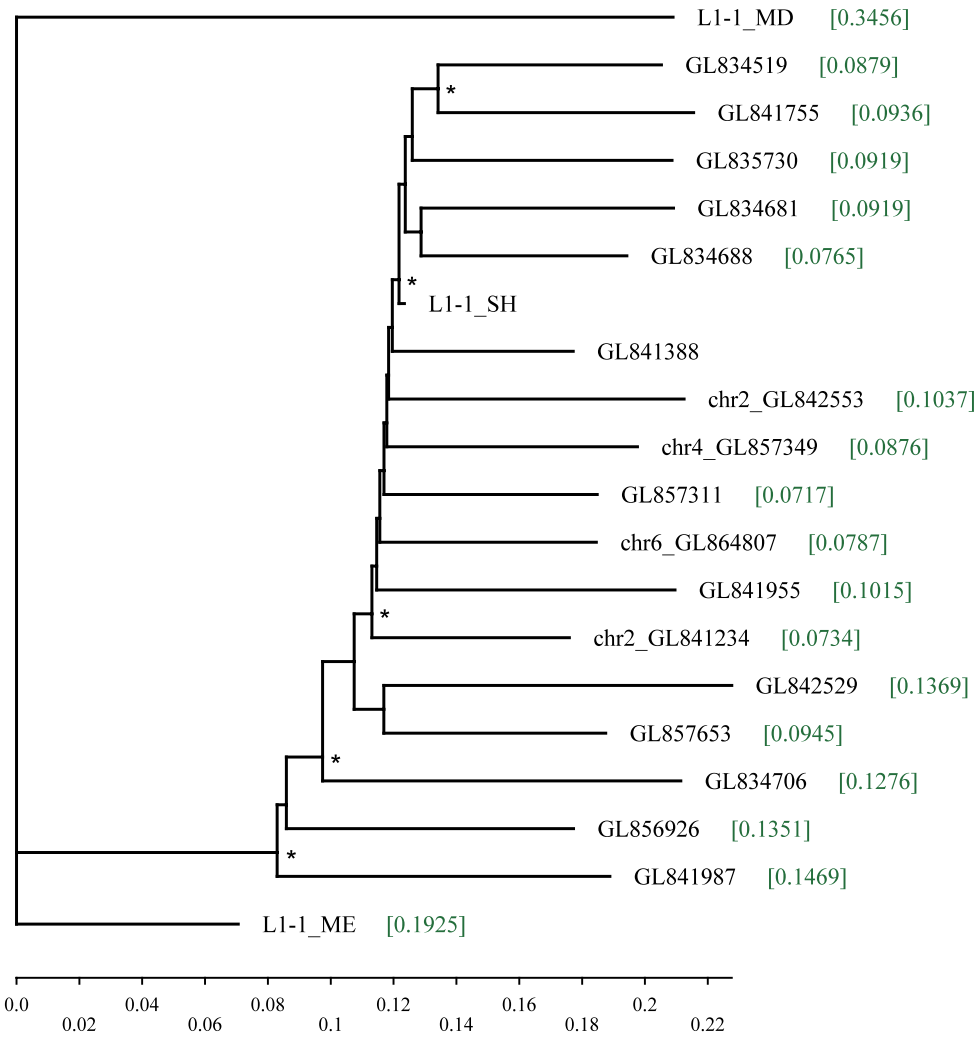
A. OC1_Das transposon



B. hAT-1_MEu transposon



Supplementary Figure 6. DNA Transposons. **(*A*)** Genomic PCR (533 nt) of the OC1_Das DNA transposon from all four Australian orders and one South American order. **(*B*)** Genomic PCR (611 nt) of the hAT-1_MEu DNA transposon from all four Australian orders and one South American order. The absence of the amplification in *M. lagotis* suggests a possible independent transfer into the four different orders. N0: negative contol. Marker: GeneRuler 1Kb



Supplementary Figure 7. Phylogenetic tree of the 17 L1-1_SH copies. The tree was calculated with maximum likelihood using the GTR+G(5) model on an alignment (4,234 nt) of ORF2 using the opossum L1-1_MD and tammar wallaby L1-1_ME ORF2 sequences as outgroup. The green value in square brackets following the scaffold name (supplementary table 1), is the pairwise distance to the L1-1_SH consensus sequence for each L1 copy. The scale bar indicates the evolutionary distance in substitutions per site. * indicate nodes that have over 90% support values as calculated by TreeFinder.