Title

Monotreme-specific conserved proteins derived from retroviral reverse transcriptase

Authors

Koichi Kitao[1], Takayuki Miyazawa[1*], So Nakagawa[2*]

Affiliations

[1]Laboratory of Virus-Host Coevolution, Institute for Frontier Life and Medical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8507, Japan

[2]Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa 259-1193, Japan.

*Correspondence to: takavet@infront.kyoto-u.ac.jp (T.M.) and so@tokai.ac.jp (S.N.)

Abstract

Endogenous retroviruses (ERVs) have played an essential role in the evolution of mammals. Many ERV-derived genes are reported in the therians that are involved in placental development. However, the contribution of the ERV-derived genes in monotremes, which are oviparous mammals, remains to be uncovered. Here, we conducted a comprehensive search for possible ERV-derived genes in platypus and echidna genomes and identified three reverse transcriptase-like genes named *RTOM1*, *2*, and *3*. They were found to be clustered in the *GRIP2* intron. Phylogenetic analysis revealed that *RTOM1*, *2*, and *3* are strongly conserved between these species, and they were generated by tandem duplications before the divergence of platypus and echidna. The *RTOM* transcripts were specifically expressed in the testis, suggesting the physiological importance of *RTOM* genes. This is the first study reporting monotreme-specific *de novo* gene candidates derived from ERVs, which provides new insights into the unique evolution of monotremes.

Introduction

Endogenous retroviruses (ERVs) are remnants of retroviral genomes found in the host genomes. ERVs are retroviruses that infect the host germline cells and are integrated into the host genome (Johnson 2019). Young ERVs retain their viral open reading frames (ORFs), but gradually lose their intact ORFs due to the accumulation of mutations. However, proteins expressed from ERVs sometimes evolve as functional genes in the host (Ueda et al., 2020). A typical example is the syncytin genes, ERV-derived fusogenic genes, which are expressed in the human placenta (Mi et al., 2000; Blond et al., 2000; Blaise et al., 2003) and are required for mouse placenta formation (Dupressoir et al., 2009; Dupressoir et al., 2011). Syncytin genes have been independently acquired from different ERVs in different mammalian lineages, which is a representative example of the convergent evolution (Imakawa et al., 2015). In addition, other ERV-derived genes that does not show fusogenic activity have also been found to be expressed in the placenta. For example, *HEMO* encoding a secreted envelope protein (Heidmann et al., 2017) as well as *gagV1* and *pre-gagV1* genes (Boso et al., 2021) are highly expressed in the human placenta. In contrast, it is unknown whether ERV-derived genes are co-opted in monotremes that are egg-laying mammals.

Here, we attempted to determine whether there are ERV-derived genes specific to monotremes. Comparative studies for the detection of ERV-derived genes have been conducted in mammalian genomes, including the platypus (Nakagawa and Takahashi 2016; Wang and Han 2020). However, for monotremes, only the genome sequence of one species, the platypus, was available (OANA5), the quality of which was limited (Warren et al., 2008). Recently, high-quality monotreme genomes of platypus (mOrnAna1.p.v1) and echidna (mTacAcu1.pri) were sequenced using long-read sequencing technology

(Zhou et al., 2021). Taking advantage of these genome sequences, we conducted comparative analyses and detected three novel ERV-derived genes specific to the monotreme lineage.

Results and Discussion

To comprehensively search for ERV genes in monotremes, we extracted ORFs from the genomes of platypus and echidna. The amino acid sequences obtained by the virtual translation of these ORFs were used as queries for the sequence search. We used the hidden Markov model (HMM) of the retroviral genes in the Gypsy Database 2.0 (GyDB) (Llorens et al., 2011) as the subject of the sequence search (Supplementary file 1—Table S1). We identified ORFs similar to *gag*, *pro*, *pol*, and *env* genes (Figure 1A). These ORFs are presumed to be a mixture of (1) ORFs that are evolutionarily conserved and (2) ORFs of young transposons that retain their ORFs. To exclude young ERV ORFs, we performed the clustering analysis based on the amino acid sequence identity. Since young ERVs are thought to be included in large clusters due to their mutual similarity, we removed sequences that belonged to large clusters consisting of more than 10 sequences. This step could also exclude evolutionarily conserved but highly duplicated genes such as SCAN domain-containing genes (Emerson and Thomas 2011), which is out of the scope of this study. Next, using the platypus ORFs as queries, and the echidna ORFs as the subjects, we conducted a sequence similarity search using BLASTp. We obtained ORF pairs with high amino acid similarity (Figure 1—figure supplement 1A; Supplementary file 1—Table S2). One of these ORFs was *ASPRV1* that is a known ERV-derived protease gene acquired in the common ancestor of mammals and is responsible for skin maintenance

77 (Matsui et al., 2011). Next, we focused on the ORFs that are absent in human genome and
78 identified three ORFs. They were located tandemly in the intron of the *GRIP2* gene in the
79 opposite direction (Figure 1B). All three ORFs showed high similarity to the reverse
80 transcriptase (RT) of spumaretrovirus in GyDB (Supplementary file 1—Table S3).
81 Therefore, we designated these genes as *RTOM* [RT-like ORF in Monotreme], and three
82 genes were named as *RTOM1*, *RTOM2*, and *RTOM3* in order of their location from the 5′
83 direction (Figure 1B). To rule out the possibility that the *RTOM* genes were acquired
84 before the divergence of humans and monotremes and were lost in humans, the *RTOM*
85 coding sequences were searched using tBLASTx with in the genomes of six mammals,
86 two birds, eight reptilians, and two amphibians (Supplementary file 1—Table S4).
87 Significant hits excluding the RT regions were not obtained other than in platypus and
88 echidna. Therefore, the *RTOM* genes were thought to be acquired after the divergence of
89 therians and monotremes.
90
91 We found that the *RTOM* genes in the platypus genome were computationally annotated
92 in the RefSeq database (Figure 2A). *RTOM1*, *2*, and *3* genes of platypus contained two
93 introns in the 5′ UTR, and the entire *RTOM* ORFs are expressed as mRNA excluding a
94 second splicing variant of *RTOM3* that partially lost its ORF (Figure 2A). In echidna,
95 *RTOM2* and *RTOM3* gene structures were annotated in the RefSeq transcripts; however,
96 *RTOM1* was not annotated. By conducting transcriptome assemblies of RNA-seq data of
97 echidna tissues (Supplementary file 1—Table S5), we reconstructed the *RTOM1*
98 transcript (Figure 2B; Supplementary file 2). As a result, all echidna *RTOM* transcripts
99 have two introns in the 5′ UTR, which was similar to observations for platypus. According

100 to the alignment of the six amino acid sequences of platypus and echidna RTOM proteins,
101 RTOM2 lacks a region shared by RTOM1 and RTOM3, but the C-terminal region was
102 conserved among the RTOM proteins without insertion or deletion (Figure 2C). To
103 investigate the tissue-specific expression of *RTOM* genes, we analyzed the RNA-seq data
104 of platypus and echidna (Supplementary file 1—Table S5). In platypus, *RTOM1*, *2*, and
105 *3* were commonly highly expressed in the testis (Figure 2D). *GRIP2* was expressed not
106 only in the testis but also in the brain, and its expression level was lower than that of the
107 *RTOM* genes. This suggests that the *RTOM* expression was not a result of the *GRIP2*
108 expression. We further investigated the mapped reads using Interactive Genome Viewer
109 (Thorvaldsdóttir et al., 2013). It was found that *RTOM3* showed a splicing variant with
110 an intron in the coding region, as shown in the RefSeq transcript (Figure 2—figure
111 supplement 1). In echidna, we found that all *RTOM* transcripts were specifically
112 expressed in the testis, similar to platypus. Expression of *GRIP2* in echidna testis was
113 also relatively low, strengthening the idea that the *RTOM* expression is independent of
114 *GRIP2* expression (Figure 2E). Given the higher expression level of *RTOM2* in both
115 platypus and echidna, this gene may play a central role of the RTOM proteins. It is still
116 possible that the relative expression levels of three genes may change according to tissues
117 and developmental stages that were not examined in this study.
118
119 To obtain insights into the viral origin of the *RTOM* genes, we performed a BLASTp
120 search of the amino acid sequence of platypus RTOM1 against the NCBI virus database.
121 We found that retrovirus Pol proteins from various distinct lineages, namely
122 gammaretrovirus, deltaretrovirus, epsilonretrovirus, and spumaretrovirus, are similar to

123 the RTOM1 proteins (BLASTp: E-value < 1E-20). In all hits, the retroviral Pol proteins
124 showed high similarity to the latter half of RTOM1 (about 370-607aa) (Figure 3A).
125 Domain search against the Pfam database (Mistry et al., 2021) in the HMMER web
126 service (Finn et al., 2011) revealed that the latter half of RTOM1 and RTOM3 contain RT
127 domains (Figure 3—figure supplement 1). A phylogenetic tree was constructed from the
128 RT regions of the RTOM proteins and the retroviral Pol proteins (Figure 3B). The RTOM
129 proteins appear to be more related to class III retroviruses, including spumaviruses or
130 spumavirus-related MuERV-L (Llorens et al., 2009). The tree topology of the RTOMs
131 strengthens the validity of the inference that three *RTOM* genes were generated by tandem
132 gene duplications before the divergence of platypus and echidna (Figure 3C). In the non-
133 RT region of RTOM1 (approximately 1-369aa), no significant hits for retroviruses were
134 obtained (Figure 3A). We performed a BLASTp search for all non-redundant proteins in
135 the GenBank database for the non-RT region of RTOM1; however, no similar proteins
136 were found except for RTOM2 and 3 (E-value < 0.05). Therefore, the non-RT region of
137 the *RTOM* genes does not seem to be derived from ERV genes or conserved host genes.
138 Considering the structural divergence of the non-RT region, such as deletion of RTOM2
139 and splicing variant of platypus RTOM3 (Figure 2C), the RT region is a core domain of
140 the RTOM proteins, and the non-RT region may provide functional modifications specific
141 to each RTOM protein.
142
143 During the 187-million-years history after diverging from monotremes, therians have
144 acquired many ERV genes and evolved their unique features, especially the placenta
145 (Imakawa and Nakagawa 2017). Our work revealed that monotremes also domesticated
146 ERV genes which emerged and was conserved more than 55 million years ago, the

147 divergence time of platypus and echidna (Zhou et al., 2021). Calculation of
148 nonsynonymous and synonymous nucleotide substitution frequencies of *RTOM1*, *2*, and
149 *3* shows that they are under purifying selection, strongly suggesting their physiological
150 importance (Figure 3D). One possibility of their functions is that they are restriction
151 factors against ERVs and retrotransposons as well as exogenous retroviruses. For example,
152 gag-derived *Fv1* (Best et al., 1996) and env-derived *Fv4* (Ikeda and Sugimura, 1989)
153 inhibit retroviral infection in mice. It is possible that the RT domains in the *RTOM* genes
154 may compete with retroviral infection and/or retrotransposition as antagonists. Another
155 possibility is that RTOM proteins are involved in physiological functions unique to
156 monotremes. To investigate that, it would be important to clarify which cells (viz. germ
157 cells or somatic cells) in testis express the *RTOM* genes. Retroviral genes like *RTOM*, in
158 which only RT domain is co-opted, have not been reported in other vertebrates to the best
159 of our knowledge (Naville et al., 2016). Thus, the future functional elucidation of *RTOM1*,
160 *2*, and *3* will provide us the new aspects of ERVs functioning in mammals.
161

162 **Materials and Methods**

163 *Identification of conserved ERV genes*

164 The platypus genome (mOrnAna1.p.v1, GCF_004115215.1) and the echidna genome
165 (mTacAcu1.pri, GCF_015852505.1) were used for the ERV gene screening (please see
166 fig. 1). The 240-nt ORF flanked by stop codons were retrieved using the getorf program
167 in the European Molecular Biology Open Software Suite (Rice et al., 2000). For HMM-
168 based sequence search, hmmscan was used (Expected threshold: 1E-5) in HMMER3
169 v3.2.2 (Eddy 2011). ORFs were clustered using CD-HIT v4.8.1 (Li and Godzik 2006)
170 with 50% amino acid identity. The sequence search for platypus ORFs against echidna

171 ORFs was conducted using BLASTp v2.10.0+ with an e-value < 1E-50 (Camacho et al.,
172 2009). To detect the presence of homologous sequences beyond the monotreme lineage,
173 deep homology searches were performed using tBLASTx v2.10.0+ with e-values <1E-5
174 on the genomes of six mammals, two birds, eight reptilians, and two amphibians
175 (Supplementary file 1—Table S4)

176 To further examine the distribution of the bitscore of the BLASTp search, we
177 extracted ORF pairs that showed high similarity (Figure 1—figure supplement 1A). Then,
178 we examined the overlap with the RefSeq annotation and identified the genes to which
179 the ORFs belonged. Finally, by extracting genes that had no homologs in the human
180 genome, *RTOM1*, *2*, and *3* were detected. To validate this approach, we performed similar
181 analyses on the genomes of human (GRCh38.p13) and marmoset
182 (Callithrix_jacchus_cj1700_1.1) (Figure 1—figure supplement 1B and C). They diverged
183 43 million years ago (Perelman et al., 2011). We succeeded in identifying known ERV-
184 derived genes such as *PEG10* (Ono et al., 2001), *RTL1*/*PEG11* (Charlier et al., 2001),
185 *ASPRV1* (Matsui et al., 2011), *NYNLIN*/*CGIN1* (Marco and Marín 2009), *ERVV-1* and *2*
186 (Kjeldbjerg et al., 2008), and *ERVMER34-1*/*HEMO* (Heidmann et al., 2017)
187 (Supplementary file 1—Table S6). This suggests that our method is sensitive enough to
188 identify ERV-ORFs conserved in platypus and echidna.

189

190 *Expression analysis*

191 RNA-seq data of platypus (20 samples from 6 tissues) (Marin et al., 2017) and echidna
192 (11 samples from 7 tissues) (Zhou et al., 2021) were used (Supplementary file 1—Table
193 S5). Low-quality reads were trimmed and filtered using fastp v0.19.5 with default options

(Chen et al., 2018). The filtered reads were mapped to each reference genome using HISAT2 v2.1.0 (Pertea et al., 2016). Based on the 11 RNA-seq sequencing data mapped on the echidna genome, we obtained the echidna *RTOM1* transcript by conducting transcriptome assembly using Stringtie2 v2.1.6 with "--merge" option (Kovaka et al., 2019). We added the coordinates of the echidna *RTOM1* transcript (Supplementary file 2) to the RefSeq gene coordinates. We then calculated the expression levels for 20 platypus and 11 echidna RNA-seq samples using the Stringtie2 program with default options (Kovaka et al., 2019).

Phylogenetic analysis

Representative retroviral Pol amino acid sequences were retrieved from the GyDB collection (https://gydb.org/index.php/Alignment?alignment=POL_retroviridae_Biology_Direct_41_2009&format=txt) (Llorens et al., 2009). A multiple alignment was generated using MAFFT v7.487 (Katoh and Standley 2013), and poorly aligned regions were removed using trimAl v1.4.rev15 (Capella-Gutiérrez et al., 2009). A phylogenetic tree was constructed using IQ-TREE2 v2.0.8 (Minh et al., 2020) with 1000 replicates of ultrafast-bootstrap (Hoang et al., 2018). The tree was visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Acknowledgments

We would like to thank Editage for English language editing. This work was supported by Grant-in-Aid for JSPS fellows 20J22607 to K.K. and JSPS KAKENHI 20K06775 and 20H03150 to T.M. and S.N. The super-computing resource was partially supported by the NIG supercomputer at ROIS National Institute of Genetics.

References

Best, S., Tissier, P. Le, Towers, G., & Stoye, J. P. (1996). Positional cloning of the mouse retrovirus restriction gene Fvl. *Nature*, *382*(6594), 826–829. https://doi.org/10.1038/382826a0

Blaise, S., de Parseval, N., Benit, L., & Heidmann, T. (2003). Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(22), 13013–13018. https://doi.org/10.1073/pnas.2132646100

Blond, J.-L., Lavillette, D., Cheynet, V., Bouton, O., Oriol, G., Chapel-Fernandes, S., Mandrand, B., Mallet, F., & Cosset, F.-L. (2000). An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *Journal of Virology*, *74*(7), 3321–3329. https://doi.org/10.1128/JVI.74.7.3321-3329.2000

Boso, G., Fleck, K., Carley, S., Liu, Q., Buckler-White, A., & Kozak, C. A. (2021). The oldest co-opted gag gene of a human endogenous retrovirus shows placenta-specific expression and is upregulated in diffuse large B-cell lymphomas. *Molecular Biology and Evolution*, *38*(12), 5453–5471. https://doi.org/10.1093/molbev/msab245

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Charlier, C., Segers, K., Wagenaar, D., Karim, L., Berghmans, S., Jaillon, O., Shay, T., Weissenbach, J., Cockett, N., Gyapay, G., & Georges, M. (2001). Human–ovine comparative sequencing of a 250-kb imprinted domain encompassing the Callipyge (clpg) locus and identification of six imprinted transcripts: DLK1 , DAT, GTL2, PEG11, antiPEG11, and MEG8. *Genome Research*, *11*(5), 850–862. https://doi.org/10.1101/gr.172701

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Dupressoir, A., Vernochet, C., Bawa, O., Harper, F., Pierron, G., Opolon, P., & Heidmann, T. (2009). Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(29), 12127–12132. https://doi.org/10.1073/pnas.0902925106

Dupressoir, A., Vernochet, C., Harper, F., Guegan, J., Dessen, P., Pierron, G., & Heidmann, T. (2011). A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proceedings of the National Academy of Sciences*, *108*(46), E1164–E1173. https://doi.org/10.1073/pnas.1112304108

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Emerson, R. O., & Thomas, J. H. (2011). Gypsy and the birth of the SCAN domain. *Journal of Virology*, *85*(22), 12043–12052. https://doi.org/10.1128/JVI.00867-11

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl), W29–W37. https://doi.org/10.1093/nar/gkr367

Heidmann, O., Béguin, A., Paternina, J., Berthier, R., Deloger, M., Bawa, O., & Heidmann, T. (2017). HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors. *Proceedings of the National Academy of Sciences*, *114*(32), E6642–E6651. https://doi.org/10.1073/pnas.1702204114

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Ikeda, H., & Sugimura, H. (1989). Fv-4 resistance gene: a truncated endogenous murine leukemia virus with ecotropic interference properties. *Journal of Virology*, *63*(12), 5405–5412. https://doi.org/10.1128/jvi.63.12.5405-5412.1989

Imakawa, K., & Nakagawa, S. (2017). The phylogeny of placental evolution through dynamic integrations of retrotransposons. *Progress in Molecular Biology and Translational Science 145*,89–109. https://doi.org/10.1016/bs.pmbts.2016.12.004

Imakawa, K., Nakagawa, S., & Miyazawa, T. (2015). Baton pass hypothesis: successive incorporation of unconserved endogenous retroviral genes for placentation during mammalian evolution. *Genes to Cells*, *20*(10), 771–788. https://doi.org/10.1111/gtc.12278

Johnson, W. E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. *Nature Reviews Microbiology*, *17*(6), 355–370. https://doi.org/10.1038/s41579-019-0189-2

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kjeldbjerg, A. L., Villesen, P., Aagaard, L., & Pedersen, F. S. (2008). Gene conversion and purifying selection of a placenta-specific ERV-V envelope gene during simian evolution. *BMC Evolutionary Biology*, *8*(1), 266. https://doi.org/10.1186/1471-2148-8-266

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with

302 StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-
303 1910-1

304 Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular
305 evolutionary genetics analysis across computing platforms. *Molecular Biology and*
306 *Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

307 Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large
308 sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659.
309 https://doi.org/10.1093/bioinformatics/btl158

310 Llorens, C., Futami, R., Covelli, L., Dominguez-Escriba, L., Viu, J. M., Tamarit, D.,
311 Aguilar-Rodriguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G. P., Maumus, F.,
312 Munoz-Pomer, A., Sempere, J. M., Latorre, A., & Moya, A. (2011). The Gypsy
313 Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Research*,
314 *39*(Database), D70–D74. https://doi.org/10.1093/nar/gkq1061

315 Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H., & Moya, A. (2009). Network
316 dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biology*
317 *Direct*, *4*(1), 41. https://doi.org/10.1186/1745-6150-4-41

318 Marco, A., & Marin, I. (2009). CGIN1: a retroviral contribution to mammalian
319 genomes. *Molecular Biology and Evolution*, *26*(10), 2167–2170.
320 https://doi.org/10.1093/molbev/msp127

321 Marin, R., Cortez, D., Lamanna, F., Pradeepa, M. M., Leushkin, E., Julien, P., Liechti,
322 A., Halbert, J., Brüning, T., Mössinger, K., Trefzer, T., Conrad, C., Kerver, H. N.,
323 Wade, J., Tschopp, P., & Kaessmann, H. (2017). Convergent origination of a
324 Drosophila-like dosage compensation mechanism in a reptile lineage. *Genome*
325 *Research*, *27*(12), 1974–1987. https://doi.org/10.1101/gr.223727.117

326 Matsui, T., Miyamoto, K., Kubo, A., Kawasaki, H., Ebihara, T., Hata, K., Tanahashi, S.,
327 Ichinose, S., Imoto, I., Inazawa, J., Kudoh, J., & Amagai, M. (2011). SASPase
328 regulates stratum corneum hydration through profilaggrin-to-filaggrin processing.
329 *EMBO Molecular Medicine*, *3*(6), 320–333.
330 https://doi.org/10.1002/emmm.201100140

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.-Y., Edouard, P., Howes, S., Keith, J. C., & McCoy, J. M. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, *403*(6771), 785–789. https://doi.org/10.1038/35001608

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Nakagawa, S., & Takahashi, M. U. (2016). gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. *Database*, *2016*, 1–8. https://doi.org/10.1093/database/baw087

Naville, M., Warren, I. A., Haftek-Terreau, Z., Chalopin, D., Brunet, F., Levin, P., Galiana, D., & Volff, J. N. (2016). Not so bad after all: Retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clinical Microbiology and Infection*, *22*(4), 312–323. https://doi.org/10.1016/j.cmi.2016.02.001

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Ono, R., Kobayashi, S., Wagatsuma, H., Aisaka, K., Kohda, T., Kaneko-Ishino, T., & Ishino, F. (2001). A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21. *Genomics*, *73*(2), 232–237. https://doi.org/10.1006/geno.2001.6494

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P. C., Silva,

A., O'Brien, S. J., & Pecon-Slattery, J. (2011). A molecular phylogeny of living primates. *PLoS Genetics*, *7*(3), 1–17. https://doi.org/10.1371/journal.pgen.1001342

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

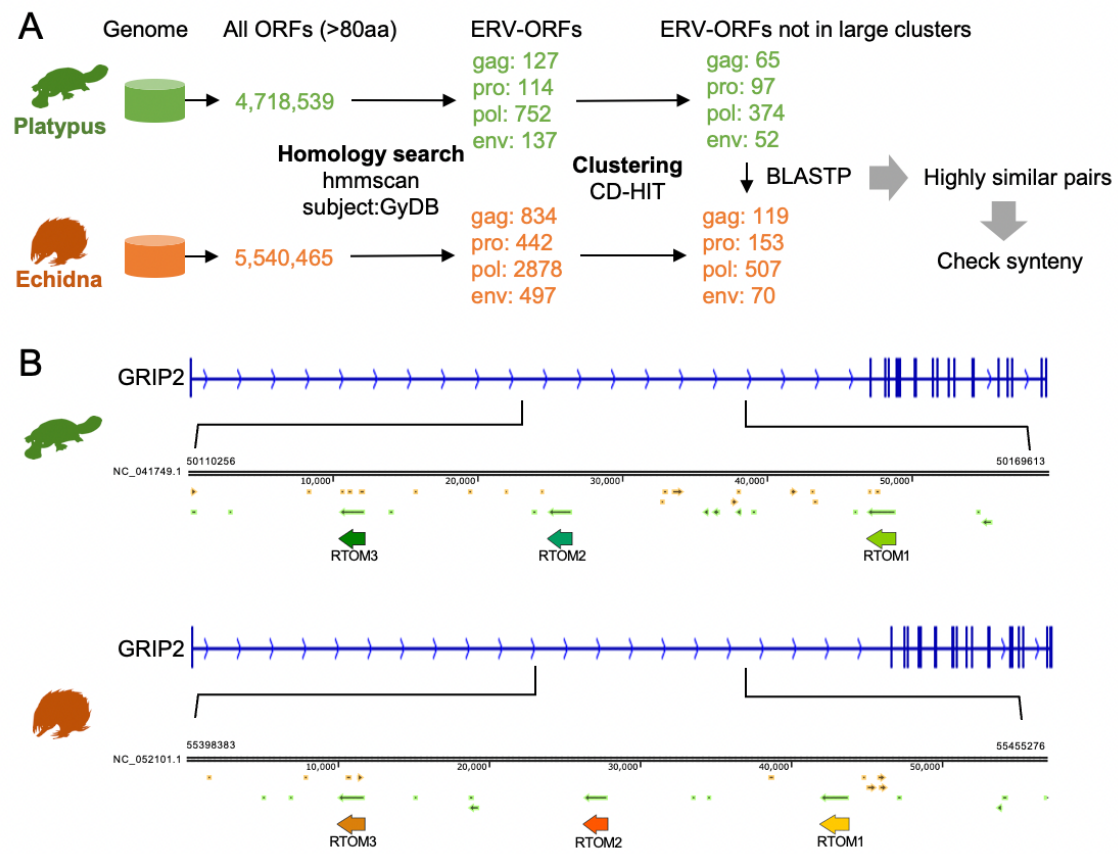
Ueda, M. T., Kryukov, K., Mitsuhashi, S., Mitsuhashi, H., Imanishi, T., & Nakagawa, S. (2020). Comprehensive genomic analysis reveals dynamic evolution of endogenous retroviruses that code for retroviral-like protein domains. *Mobile DNA*, *11*(1), 29. https://doi.org/10.1186/s13100-020-00224-w

Wang, J., & Han, G. Z. (2020). Frequent retroviral gene co-option during the evolution of vertebrates. *Molecular Biology and Evolution*, *37*(11), 3232–3242. https://doi.org/10.1093/molbev/msaa180

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., … Wilson, R. K. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, *453*(7192), 175–183. https://doi.org/10.1038/nature06936

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., … Zhang, G. (2021). Platypus and echidna genomes reveal mammalian biology and evolution. *Nature*, *592*(7856), 756–762. https://doi.org/10.1038/s41586-020-03039-0

391 **Figure legends**



392

393 Figure 1. Identification of *RTOM1*, *2*, and *3*. (A) Schematic representation of the in silico

394 screening for conserved ERV-derived genes in platypus and echidna. (B) Genomic

395 context of *RTOM1*, *2*, and *3*.

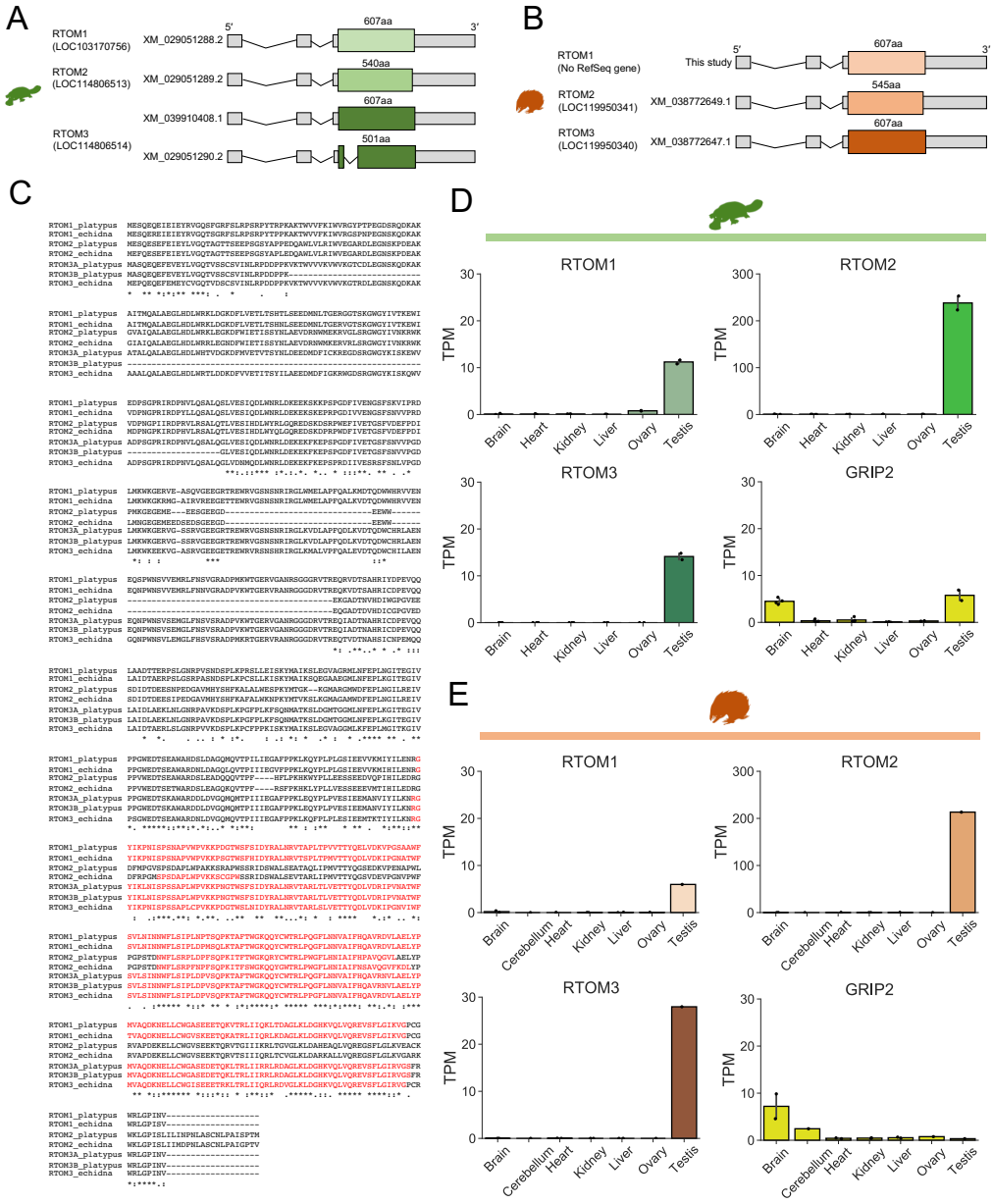
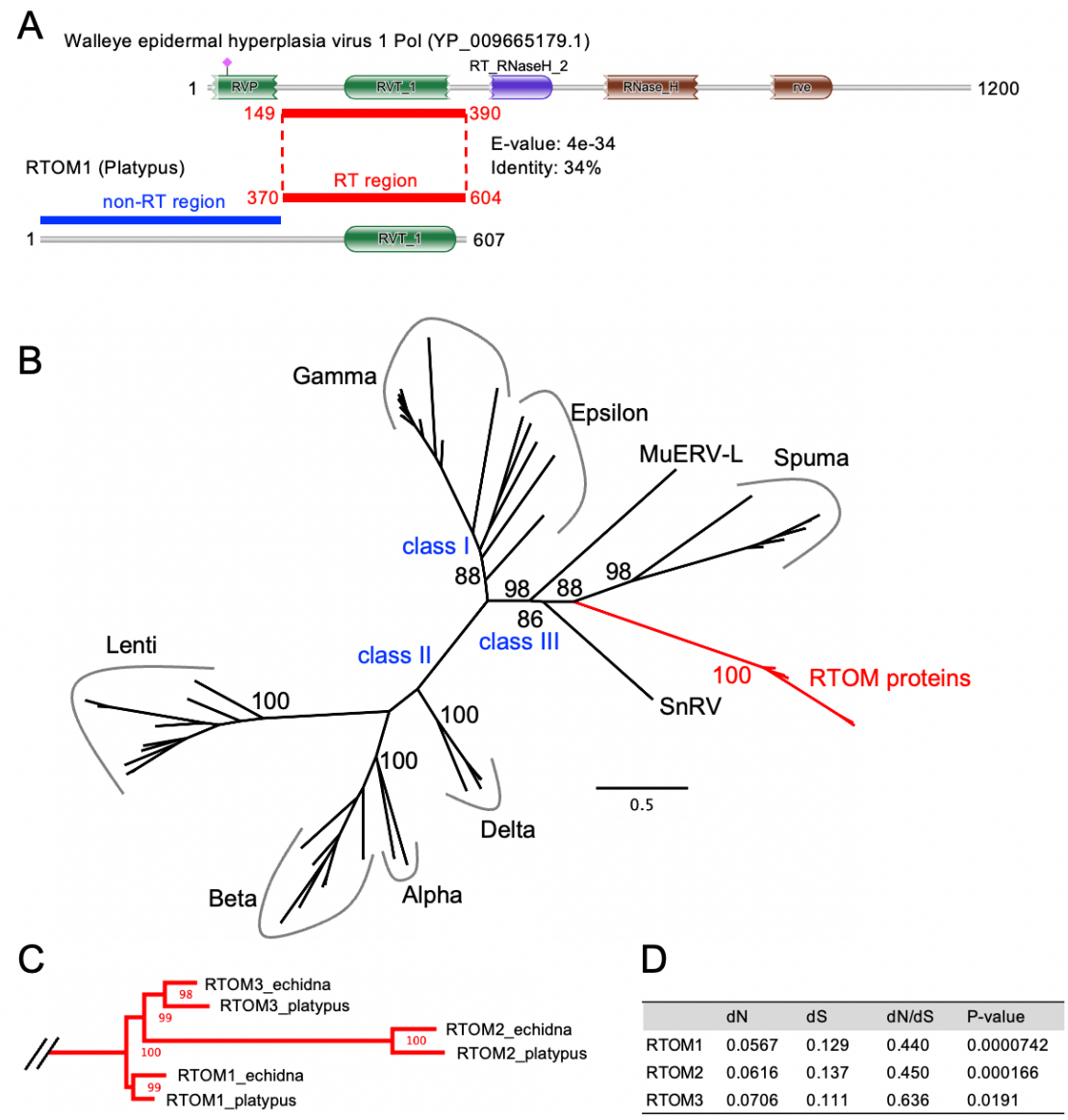


Figure 2. Expression of *RTOM1*, *2*, and *3*. (A) Schematic representation of the RefSeq transcripts of the *RTOM* genes in platypus. (B) Schematic representation of the reconstructed *RTOM1* transcript and RefSeq transcripts of the *RTOM2* and *3* genes in echidna. (C) Multiple alignment of the amino acid sequences of RTOM proteins. The amino acid sequence of echidna RTOM1 was obtained from the genomic ORF. "RTOM3A_plasypus" and "RTOM3B_platypus" are protein isoforms derived from

403 “XM_039910408.1” and “XM_029051290.2,” respectively. The regions showing
404 similarity to the HMM of spumaretrovirus RT domain in GyDB are indicated in red. (D
405 and E) Tissue-specific expression of *RTOM* genes and *GRIP2* in (D) platypus and (E)
406 echidna. Normalized expression levels are presented as transcript per million (TPM).
407



	dN	dS	dN/dS	P-value
RTOM1	0.0567	0.129	0.440	0.0000742
RTOM2	0.0616	0.137	0.450	0.000166
RTOM3	0.0706	0.111	0.636	0.0191

Figure 3. Evolution of *RTOM1*, *2*, and *3*. (A) Comparison between platypus RTOM1 and retroviral Pol protein. Walleye epidermal hyperplasia virus 1 is represented as an example. A region showing similarity to the Pol protein by BLASTp was designated as "RT region." A region that did not show similarity to any retroviral genes was designated as "non-RT region." (B) A phylogenetic tree constructed from the amino acid sequences of RT regions of the six RTOM proteins and the retroviral Pol proteins in GyDB. The multiple alignment is available in Supplementary file 3. Ultrafast-bootstrap values obtained from 1000 times

416 replication are shown in major branches. (C) Detailed representation of the clade of the
417 RTOM proteins. (D) The numbers of non-synonymous substitutions (dN) and
418 synonymous substitutions (dS) per site estimated by Nei-Gojobori method (Nei and
419 Gojobori 1986). Statistical significance of selection was estimated by the codon-based Z
420 test of neutrality using MEGA-X (Kumar et al., 2018).
421

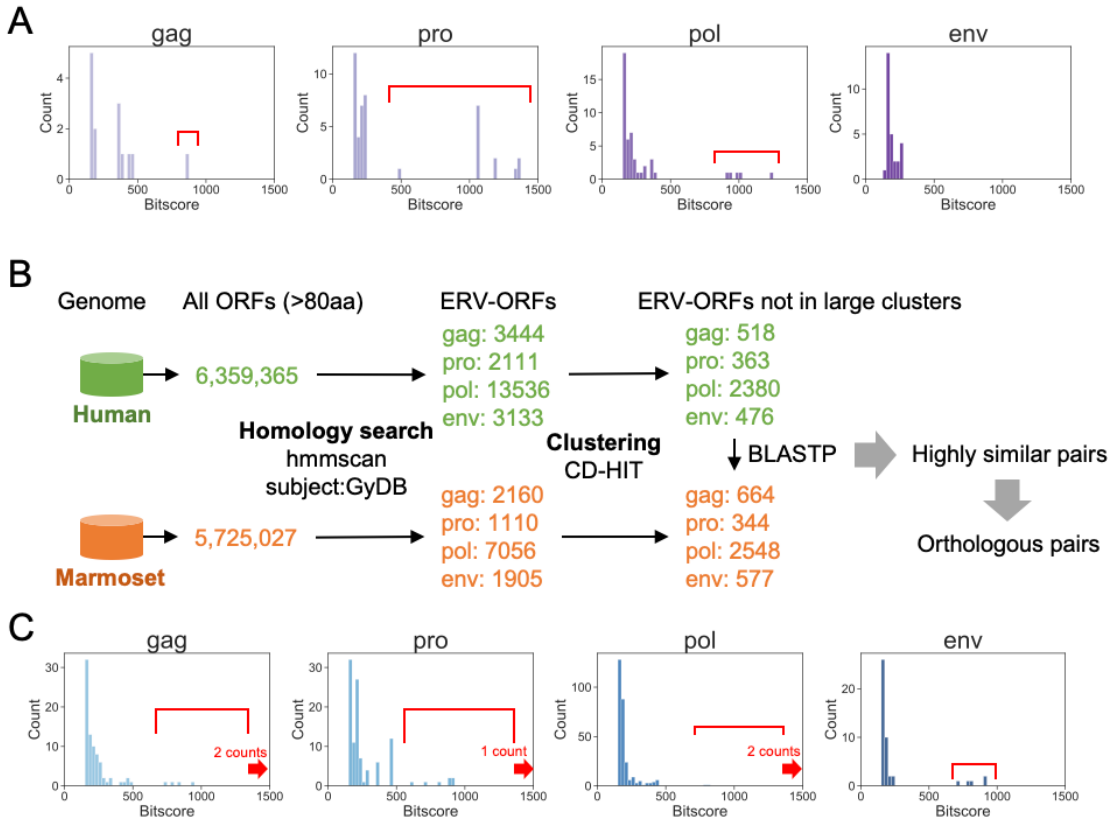


Figure 1—figure supplement 1. (A) Distribution of bitscore obtained by the BLASTp search for platypus ERV-ORFs against echidna ERV-ORFs. The ORF pairs with high bitscore indicated by red lines were used for further analysis (Supplementary file 1—Table S2). (B) Schematic representation of the control screening for conserved ERV-derived genes in human and marmoset. (C) Distribution of bitscore obtained by BLASTp for human ERV-ORFs against marmoset ERV-ORFs. The numbers of hits with bitscore higher than 1500 are indicated by red arrows. The ORF pairs with high bitscore indicated by red lines were used for further analysis (Supplementary file 1—Table S6).

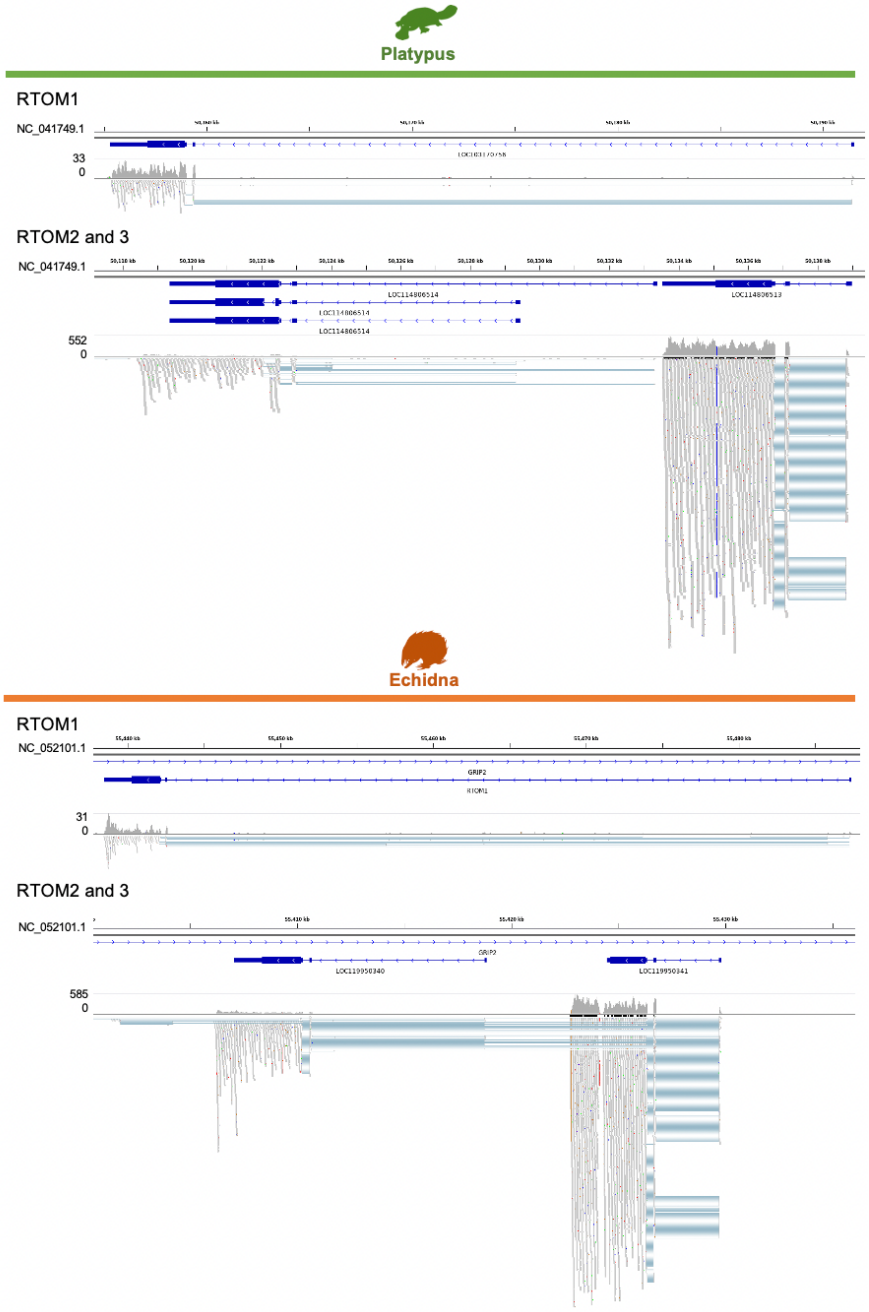


Figure 2—figure supplement 1. Screenshots of Interactive Genomic Viewer of RNA-seq reads on the *RTOM* genes. The transcript tracks in blue lines display the coordinates from the RefSeq GTF files. Thick blue lines indicate the coding sequences. Since there is no corresponding RefSeq transcript for echidna *RTOM1*, its gene coordinate was manually added from assembled transcripts in this study (Materials and Methods).

Figure 3—figure supplement 1. Protein domains in RTOM1, 2, and 3. The domain search was conducted using hmmscan in HMMER web server with default options (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan).