1 ***Letter***

2 **Discoveries**

3 **Title**

4 Monotreme-specific conserved proteins derived from retroviral reverse transcriptase

5

6 **Authors**

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

8

9 **Affiliations**

10 [1]Laboratory of Virus-Host Coevolution, Institute for Frontier Life and Medical Sciences,

11 Kyoto University, Sakyo-ku, Kyoto 606-8507, Japan

12 [2]Department of Molecular Life Science, Tokai University School of Medicine, Isehara,

13 Kanagawa 259-1193, Japan.

14

15 *Correspondence to: takavet@infront.kyoto-u.ac.jp (TM) and so@tokai.ac.jp (SN)

16

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 the 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.

32 Endogenous retroviruses (ERVs) are remnants of retroviral genomes found in the host
33 genomes. ERVs are retroviruses that infect the host germline cells and are integrated into
34 the host genome (Johnson 2019). Young ERVs retain their viral open reading frames
35 (ORFs), but gradually lose their intact ORFs due to the accumulation of mutations.
36 However, proteins expressed from ERVs sometimes evolve as functional genes in the host
37 (Ueda et al. 2020). A typical example is the syncytin genes, ERV-derived fusogenic genes,
38 which are expressed in the human placenta (Mi et al. 2000; Blond et al. 2000; Blaise et
39 al. 2003) and are required for mouse placenta formation (Dupressoir et al. 2009;
40 Dupressoir et al. 2011). Syncytin genes have been independently acquired from different
41 ERVs in different mammalian lineages, which is a representative example of the
42 convergent evolution (Imakawa et al. 2015). In addition, other ERV-derived genes have
43 also been found to be expressed in the placenta. For example, *HEMO* encoding a secreted
44 envelope protein (Heidmann et al. 2017) as well as *gagV1* and *pre-gagV1* genes (Boso et
45 al. 2021) are highly expressed in the human placenta. However, it is unknown whether
46 ERV-derived genes are co-opted in monotremes that are egg-laying mammals.
47 Comparative studies for the detection of ERV-derived genes have been conducted in
48 mammalian genomes, including the platypus (Nakagawa and Takahashi 2016; Wang and
49 Han 2020). For monotremes, however, only the genome sequence of one species, the
50 platypus, was available (OANA5). and the quality was limited (Warren et al. 2008).
51 Recently, high-quality monotreme genomes of platypus (mOrnAna1.p.v1) and echidna
52 (mTacAcu1.pri) were sequenced using long-read sequencing technology (Zhou et al.
53 2021), making it possible to search for conserved ERV genes in monotremes. Here, we
54 performed a comparative analysis of the genomes of these two monotremes and found
55 three ERV-derived genes that are specific to the monotreme lineage.

56

57 To comprehensively search for ERV genes in monotremes, we extracted ORFs from the
58 genomes of platypus and echidna. The amino acid sequences obtained by the virtual
59 translation of these ORFs were used as queries for the sequence search. We used the
60 hidden Markov model (HMM) of the retroviral genes in the Gypsy Database 2.0 (GyDB)
61 (Llorens et al. 2011) as the subject of the sequence search (supplementary table S1). We
62 identified ORFs similar to *gag*, *pro*, *pol*, and *env* genes (fig. 1A). These ORFs are
63 presumed to be a mixture of (1) ORFs that are evolutionarily conserved and (2) ORFs of
64 young transposons that retain their ORFs. To exclude young ERV ORFs, we performed
65 the clustering analysis based on the amino acid sequence identity. Since young ERVs are
66 thought to be included in large clusters due to their mutual similarity to each other, we
67 removed sequences that belonged to large clusters consisting of more than 10 sequences.
68 Next, using the platypus ORFs as queries, and the echidna ORFs as the subjects, we
69 conducted a sequence similarity search using BLASTp. We obtained nine ORF pairs with
70 high amino acid similarity and the same synteny between platypus and echidna
71 (supplementary table S2). For six pairs among these, we found respective homologs in
72 the human genome, indicating that they were either ERV genes acquired in the common
73 mammalian ancestor or host genes with high similarity to ERVs. Indeed, one of the six
74 genes is *ASPRV1* that is a known ERV-derived protease gene acquired in the common
75 ancestor of mammals and is responsible for skin maintenance (Matsui et al. 2011). The
76 remaining three genes were not found in the human genome. They were located tandemly
77 in the intron of the *GRIP2* gene in the opposite direction (fig. 1B). All three ORFs showed
78 high similarity to the reverse transcriptase (RT) of spumaretrovirus in GyDB
79 (supplementary table S3). Therefore, we designated these genes as *RTOM* [RT-like ORF

80 in Monotreme], and three genes were named as *RTOM1*, *RTOM2*, and *RTOM3* in order
81 of their location from the 5′ direction (fig. 1B). The *RTOM* coding sequences were
82 searched in the genomes of 6 mammals, 2 birds, 8 reptilians, and 2 amphibians
83 (supplementary table S4); however, significant hits were not obtained other than in
84 platypus and echidna (BLASTn: E-value < 1E-5). Therefore, the *RTOM* genes could be
85 monotreme-specific and by originated more than 55 million years ago, the divergence
86 time of platypus and echidna (Zhou et al. 2021).
87
88 We found that the gene structures of *RTOM* genes in the platypus genome were annotated
89 in the RefSeq database (fig. 2A). *RTOM1*, *2*, and *3* genes of platypus contained two
90 introns in the 5′ UTR, and the entire *RTOM* ORFs are expressed as mRNA excluding a
91 second splicing variant of *RTOM3* that partially lost its ORF (fig. 2A). In echidna,
92 *RTOM2* and *RTOM3* gene structures were annotated in the RefSeq transcripts; however,
93 *RTOM1* was not annotated. By conducting transcriptome assemblies of RNA-seq data of
94 echidna tissues (supplementary table S5), we reconstructed the *RTOM1* transcript (fig.
95 2B; supplementary data S1). As a result, all echidna *RTOM* transcripts have two introns
96 in the 5′ UTR, which was similar to observations for platypus. According to the alignment
97 of the six amino acid sequences of platypus and echidna *RTOM* genes, RTOM2 lacks a
98 region shared by RTOM1 and RTOM3, but the C-terminal region was conserved among
99 the RTOM proteins without insertion or deletion (fig. 2C). To investigate the tissue-
100 specific expression of *RTOM* genes, we analyzed the RNA-seq data of platypus and
101 echidna (supplementary table S5). In platypus, *RTOM1*, *2*, and *3* were all highly expressed
102 in the testis (fig. 2D). *GRIP2* was expressed not only in the testis but also in the brain,
103 and its expression level was lower than that of the *RTOM* genes. This suggests that the

104 *RTOM* expression was not a result of the *GRIP2* expression. We further investigated the
105 mapped reads using Interactive Genome Viewer (Thorvaldsdóttir et al. 2013) and found
106 that *RTOM3* showed a splicing variant with an intron in the coding region, as shown in
107 the RefSeq transcript (supplementary fig S1). In echidna, we found that all *RTOM*
108 transcripts were specifically expressed in the testis, similar to platypus. Expression of
109 *GRIP2* in echidna testis was also relatively low, strengthening the idea that the *RTOM*
110 expression is independent of *GRIP2* expression (fig. 2E). Given the higher expression
111 level of *RTOM2* in both platypus and echidna, this gene may play a central role of the
112 RTOM proteins. It is still possible that the relative expression levels of three genes may
113 change according to tissues and developmental stages that were not examined in this study.
114
115 To obtain insights into the viral origin of the *RTOM* genes, we performed a BLASTp
116 search of the amino acid sequence of platypus RTOM1 against the NCBI virus database.
117 We found that retrovirus Pol proteins from various distinct lineages, namely
118 gammaretrovirus, deltaretrovirus, epsilonretrovirus, and spumaretrovirus, are similar to
119 the RTOM1 proteins (BLASTp: E-value < 1E-20). In all hits, the retroviral Pol proteins
120 showed high similarity to the latter half of RTOM1 (about 370-607aa). Domain search
121 against the Pfam database (Mistry et al. 2021) in the HMMER web service (Finn et al.
122 2011) revealed that the latter half of RTOM1 and RTOM3 contain RT domains (fig. 3A;
123 supplementary fig. S2). A phylogenetic tree was constructed from the RT regions of the
124 RTOM proteins and the retroviral Pol proteins (fig. 3B). The RTOM proteins appear to
125 be more related to class III retroviruses, including spumaviruses or spumavirus-related
126 MuERV-L (Llorens et al. 2009). The tree topology suggested that RTOM1 emerged at
127 first, and RTOM2 and RTOM3 were then generated by tandem gene duplications before

128 the divergence of platypus and echidna (fig. 3C). In the non-RT region of RTOM1
129 (approximately 1-369aa), no significant hits for retroviruses were obtained (fig. 3A). We
130 performed a BLASTp search for all non-redundant proteins in the GenBank database for
131 the non-RT region of RTOM1; however, no similar proteins were found except for
132 RTOM2 and 3 (E-value < 0.05). Therefore, the non-RT region of the *RTOM* genes dons
133 not seem to be derived from ERV genes or conserved host genes. Considering the
134 structural divergence of the non-RT region, such as deletion of RTOM2 and splicing
135 variant of platypus RTOM3 (fig. 2C), the RT region is a core domain of the RTOM
136 proteins, and the non-RT region may provide functional modifications specific to each
137 *RTOM* protein.
138
139 During the 187-million-years history after diverging from monotremes (Zhou et al. 2021),
140 therians have acquired many ERV genes and evolved their unique features, especially the
141 placenta (Imakawa and Nakagawa 2017). Our work revealed that monotremes also
142 domesticated ERV genes. The functional inference of the RTOM proteins is difficult as
143 co-opted RT genes, such as RTOMs, have not been reported in other vertebrates to the
144 best of our knowledge. One possibility is that RTOM proteins may function as restrictive
145 factors against ERVs and retrotransposons. For example, gag-derived *Fv1* (Best et al.
146 1996) and env-derived *Fv4* (Ikeda and Sugimura 1989) inhibit retroviral infection in mice.
147 It is possible that the RT domains in the *RTOM* genes compete with retrotransposition as
148 antagonists. Another possibility is that RTOM proteins are involved in physiological
149 functions unique to monotremes. In future studies, it would be important to clarify which
150 cells, viz. germ cells or somatic cells, in testis express the *RTOM* genes. Further studies
151 pertaining to *RTOM1*, *2*, and *3* in platypus and echidna will expand our understanding of

ERV co-option during the evolution of mammals.

Materials and Methods

Identification of conserved ERV genes

The platypus genome (mOrnAna1.p.v1, GCF_004115215.1) and the echidna genome (mTacAcu1.pri, GCF_015852505.1) were used for the ERV gene screening (please see fig. 1). The 240-nt ORF flanked by stop codons were retrieved using the getorf program in the European Molecular Biology Open Software Suite (Rice et al. 2000). For HMM-based sequence search, hmmscan was used (Expected threshold: 1E-5) in HMMER3 v3.2.2 (Eddy 2011). ORFs were clustered using CD-HIT v4.8.1 (Li and Godzik 2006) with 50% amino acid identity. The sequence search for platypus ORFs against echidna ORFs was conducted using BLASTp v2.10.0+ with an e-value < 1E-50 (Camacho et al. 2009). Hits with a bitscore > 400 were retrieved and checked the synteny was checked using the UCSC genome browser (https://genome.ucsc.edu/index.html).

Expression analysis

RNA-seq data of platypus (20 samples from 6 tissues) (Marin et al. 2017) and echidna (11 samples from 7 tissues) (Zhou et al. 2021) were used (supplementary file 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 the 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 data S1) to the RefSeq

176 gene coordinates. We then calculated the expression levels for 20 platypus and 11 echidna
177 RNA-seq samples using the Stringtie2 program with default options (Kovaka et al. 2019).
178
179 *Phylogenetic analysis*
180 Representative retroviral Pol amino acid sequences were retrieved from the GyDB
181 collection
182 (https://gydb.org/index.php/Alignment?alignment=POL_retroviridae_Biology_Direct_4
183 _41_2009&format=txt) (Llorens et al. 2009). A multiple alignment was generated using
184 MAFFT v7.487 (Katoh and Standley 2013), and poorly aligned regions were removed
185 using trimAl v1.4.rev15 (Capella-Gutiérrez et al. 2009). A phylogenetic tree was
186 constructed using IQ-TREE2 v2.0.8 (Minh et al. 2020) with 1000 replicates of ultrafast-
187 bootstrap (Hoang et al. 2018). The tree was visualized using FigTree v1.4.4
188 (http://tree.bio.ed.ac.uk/software/figtree/).
189
190

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, Le Tissier P, Towers G, Stoye JP. 1996. Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature*. 382:826–829.

Blaise S, de Parseval N, Bénit L, Heidmann T. 2003. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc Natl Acad Sci U S A.* 100:13013–13018.

Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL. 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. *J Virol.* 74:3321–3329.

Boso G, Fleck K, Carley S, Liu Q, Buckler-White A, Kozak CA. 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. *Mol Biol Evol.* 38:5453–5471.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890.

219 Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T.
220 2009. Syncytin-A knockout mice demonstrate the critical role in placentation of a
221 fusogenic, endogenous retrovirus-derived, envelope gene. *Proc Natl Acad Sci U*
222 *S A.* 106:12127–12132.

223 Dupressoir A, Vernochet C, Harper F, Guégan J, Dessen P, Pierron G, Heidmann T.
224 2011. A pair of co-opted retroviral envelope syncytin genes is required for
225 formation of the two-layered murine placental syncytiotrophoblast. *Proc Natl*
226 *Acad Sci U S A* 108:E1164–E1173.

227 Eddy SR. 2011. Accelerated profile HMM Searches. *PLoS Comput Biol.* 7:e1002195.

228 Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence
229 similarity searching. *Nucleic Acids Res.* 39:W29–W37.

230 Heidmann O, Béguin A, Paternina J, Berthier R, Deloger M, Bawa O, Heidmann T.
231 2017. HEMO, an ancestral endogenous retroviral envelope protein shed in the
232 blood of pregnant women and expressed in pluripotent stem cells and tumors.
233 *Proc Natl Acad Sci U S A.* 114:E6642–E6651.

234 Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2:
235 Improving the ultrafast bootstrap approximation. *Mol Biol Evol.* 35:518–522.

236 Ikeda H, Sugimura H. 1989. Fv-4 resistance gene: a truncated endogenous murine
237 leukemia virus with ecotropic interference properties. *J. Virol.* 63:5405–5412.

238 Imakawa K, Nakagawa S. 2017. The phylogeny of placental evolution through dynamic
239 integrations of retrotransposons. *Prog Mol Biol Transl Sci.* 145:89–109.

240 Imakawa K, Nakagawa S, Miyazawa T. 2015. Baton pass hypothesis: successive
241 incorporation of unconserved endogenous retroviral genes for placentation during
242 mammalian evolution. *Genes Cells.* 20:771–788.

243 Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous
244 retroviruses. *Nat Rev Microbiol.* 17:355–370.

245 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
246 improvements in performance and usability. *Mol Biol Evol.* 30:772–780.

247 Kovaka S, Zimin A V., Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019.

248 Transcriptome assembly from long-read RNA-seq alignments with StringTie2.
249 *Genome Biol.* 20:278.

250 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of
251 protein or nucleotide sequences. *Bioinformatics.* 22:1658–1659.

252 Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A. 2009. Network dynamics of
253 eukaryotic LTR retroelements beyond phylogenetic trees. *Biol Direct* 4:41.

254 Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-
255 Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, et al. 2011. The Gypsy
256 Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res*.
257 39:D70–D74.

258 Marin R, Cortez D, Lamanna F, Pradeepa MM, Leushkin E, Julien P, Liechti A, Halbert
259 J, Brüning T, Mössinger K, et al. 2017. Convergent origination of a Drosophila-
260 like dosage compensation mechanism in a reptile lineage. *Genome Res.* 27:1974–
261 1987.

262 Matsui T, Miyamoto K, Kubo A, Kawasaki H, Ebihara T, Hata K, Tanahashi S,
263 Ichinose S, Imoto I, Inazawa J, et al. 2011. SASPase regulates stratum corneum
264 hydration through profilaggrin-to-filaggrin processing. *EMBO Mol Med.* 3:320–
265 333.

266 Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard
267 P, Howes S, et al. 2000. Syncytin is a captive retroviral envelope protein
268 involved in human placental morphogenesis. *Nature* 403:785–789.

269 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A,
270 Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for
271 phylogenetic inference in the genomic era. *Mol Biol Evol.* 37:1530–1534.

272 Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL,
273 Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein
274 families database in 2021. *Nucleic Acids Res.* 49:D412–D419.

275 Nakagawa S, Takahashi MU. 2016. gEVE: a genome-based endogenous viral element
276 database provides comprehensive viral protein-coding sequences in mammalian
277 genomes. *Database.* 2016:1–8.

278 Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression
279 analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat*
280 *Protoc.* 11:1650–1667.

281 Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open
282 Software Suite. *Trends Genet.* 16:276–277.

283 Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer
284 (IGV): high-performance genomics data visualization and exploration. *Brief*
285 *Bioinform.* 14:178–192.

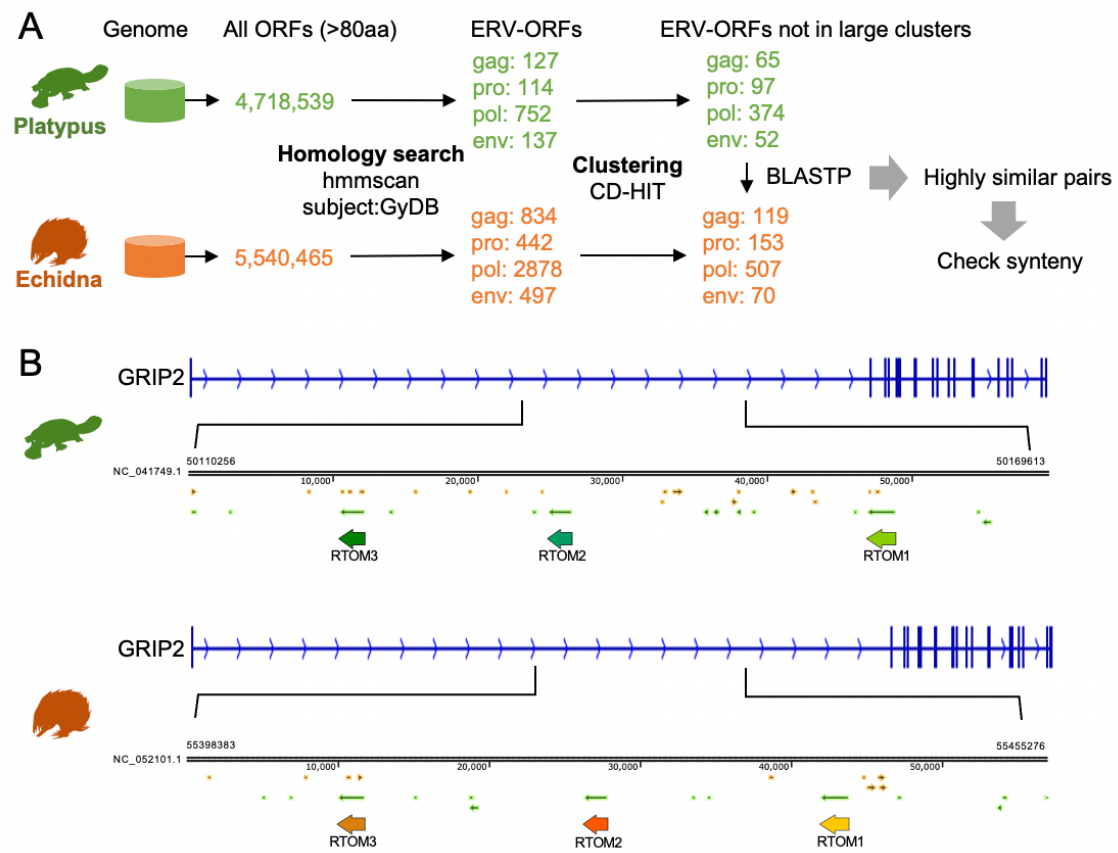
286 Ueda MT, Kryukov K, Mitsuhashi S, Mitsuhashi H, Imanishi T, Nakagawa S. 2020.
287 Comprehensive genomic analysis reveals dynamic evolution of endogenous
288 retroviruses that code for retroviral-like protein domains. *Mob DNA* 11:29.

289 Wang J, Han GZ. 2020. Frequent retroviral gene co-option during the evolution of
290 vertebrates. *Mol Biol Evol.* 37:3232–3242.

291 Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov
292 K, Miller W, Clarke L, Chinwalla AT, et al. 2008. Genome analysis of the
293 platypus reveals unique signatures of evolution. *Nature* 453:175–183.

294 Zhou Y, Shearwin-Whyatt L, Li J, Song Z, Hayakawa T, Stevens D, Fenelon JC, Peel
295 E, Cheng Y, Pajpach F, et al. 2021. Platypus and echidna genomes reveal
296 mammalian biology and evolution. *Nature* 592:756–762.

297 **Figure legends**



298

299 FIG. 1. Identification of RTOM1, 2, and 3. (A) Schematic representation of the in silico

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

301 context of RTOM1, 2, and 3.

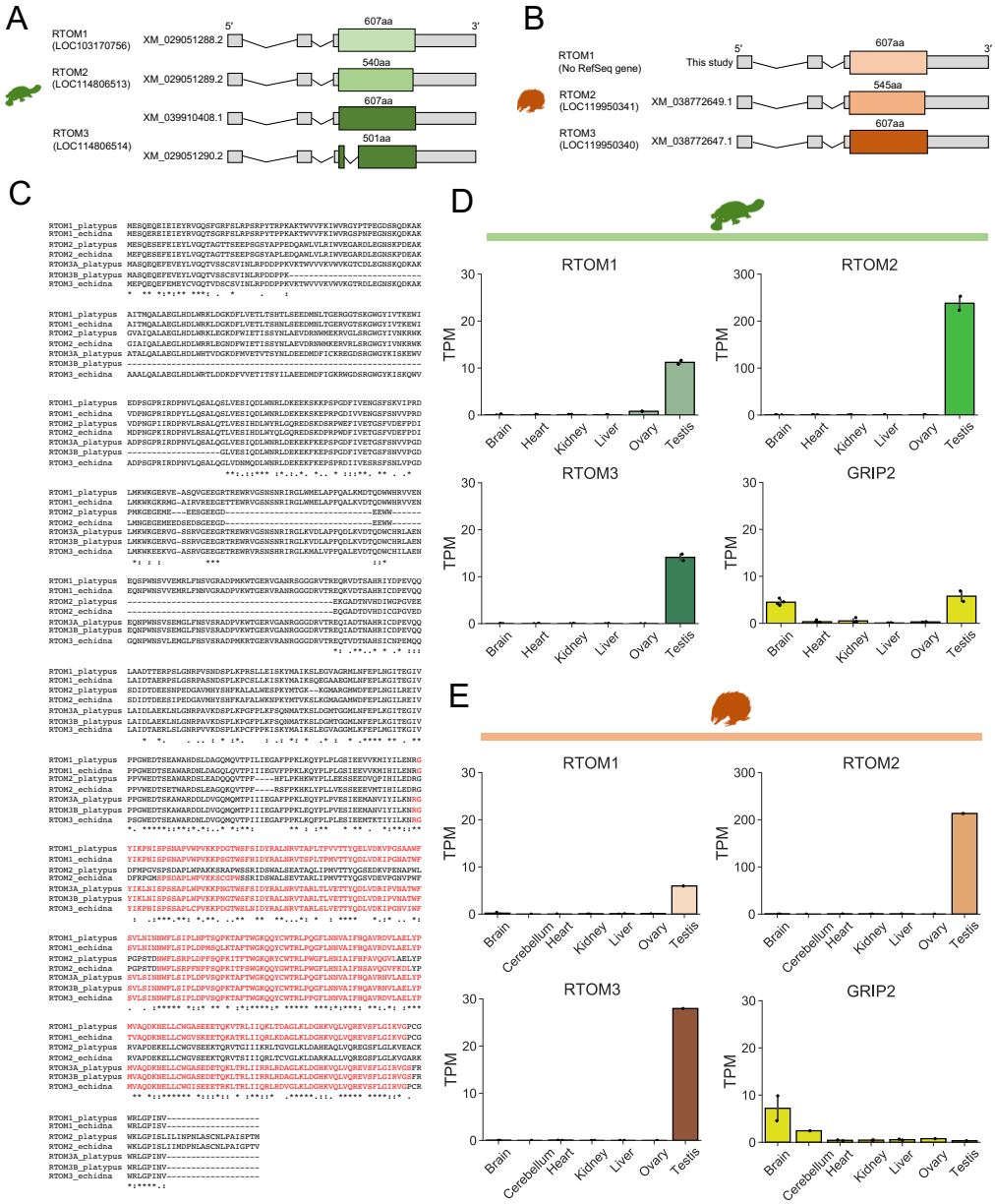


FIG. 2. Expression of *RTOM1*, *2*, and *3*. (A) Schematic representation of the RefSeq transcripts of the *RTOM* genes in platypus. (B) Schematic representation of the reconstructed *RTOM1* transcript and RefSeq transcripts of the *RTOM2* and *3* genes in echidna. (C) Multiple alignment of the amino 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

309 “XM_039910408.1” and “XM_029051290.2,” respectively. The regions showing
310 similarity to the HMM of spumaretrovirus RT domain in GyDB are indicated in red. (D
311 and E) Tissue-specific expression of *RTOM* genes and *GRIP2* in (D) platypus and (E)
312 echidna. Normalized expression levels are presented as transcript per million (TPM).
313

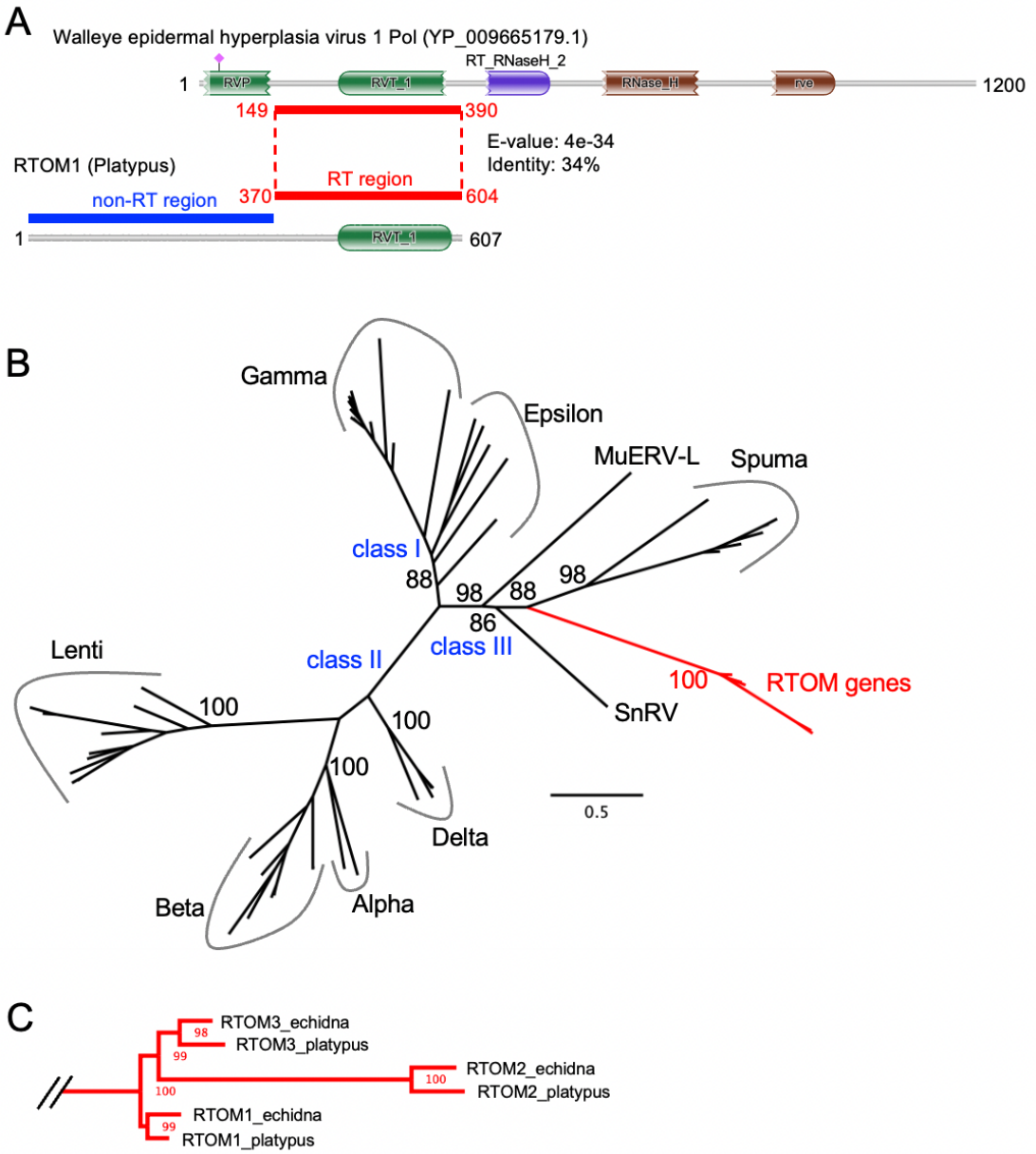


FIG. 3. Evolution of RTOM1, 2, and 3. (A) Comparison between platypus RTOM1 and retroviral Pol protein. Walleye epidermal hyperplasia virus 1 is represented as an example. A region showing similarity to the Pol protein by BLASTp was designated as “RT region.” A region that did not show similarity to 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

321 is available in supplementary data S2. Ultrafast-bootstrap values obtained from 1000
322 times replication are shown in major branches. (C) Detailed representation of the clade
323 of the RTOM genes in (B).
324

325 **Supplementary Materials**

326 Supplementary fig. S1. Screenshots of Interactive Genomic Viewer of RNA-seq reads on
327 the RTOM genes.

328 Supplementary fig. S2. Protein domains in RTOM1, 2, and 3.

329 Supplementary table S1. The HMM profiles in GyDB used in this study.

330 Supplementary table S2. ERV-like ORFs shared between platypus and echidna

331 Supplementary table S3. The GyDB HMMs hit to the RTOM genes

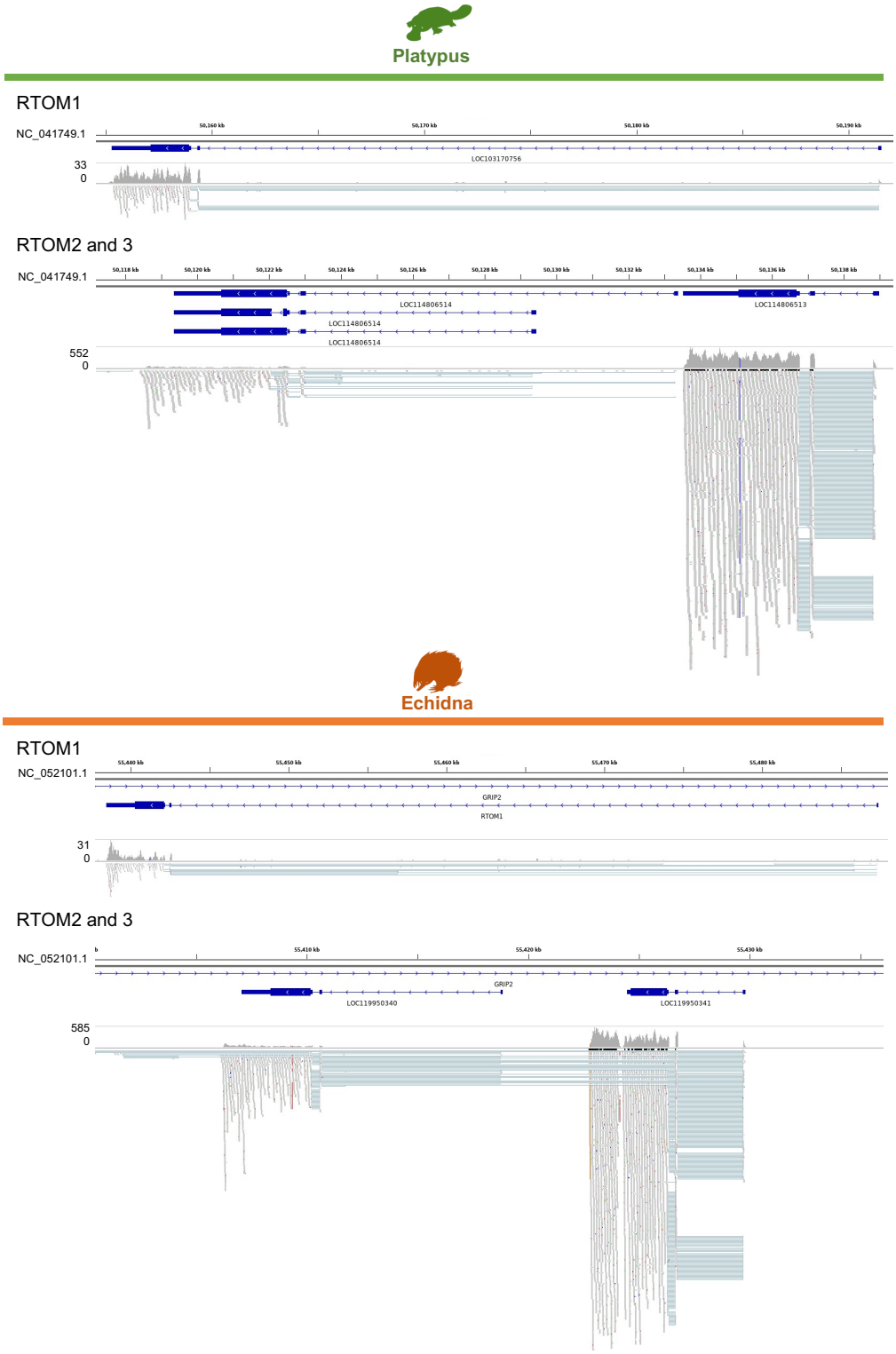
332 Supplementary table S4. Species and genomes used for genes similar to the RTOM genes.

333 Supplementary table S5. RNA-seq data used in this study.

334 Supplementary data S1. Nucleotide sequence of the echidna *RTOM1* transcript.

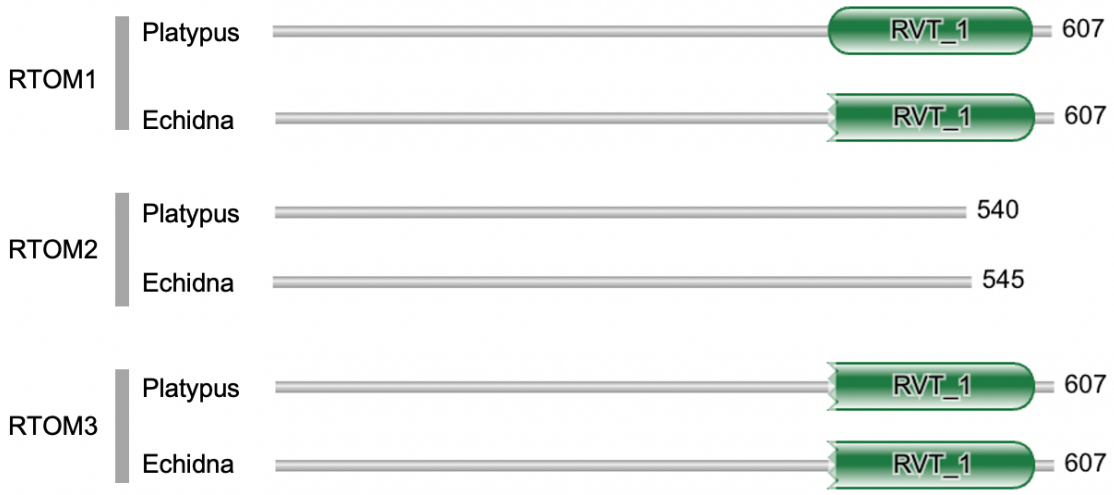
335 Supplementary data S2. Alignment of representative retroviral *pol* genes and the *RTOM*
336 genes.

337



Supplementally FIG. S1. Screenshots of Interactive Genomic Viewer of RNA-seq reads on the *RTOM* genes. The transcript tracks in blue lines display the coordinates from the

341 RefSeq GTF files. Thick blue lines indicate the coding sequences. Since there is no
342 corresponding RefSeq transcript for echidna *RTOM1*, its gene coordinate was manually
343 added from assembled transcripts in this study (Materials and Methods).
344



Supplementally FIG. S2. 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).