Genome-wide characterization ofendogenous retroviruses in snub-nosed monkeys

Xiao Wang [1] , **Boshi Wang** [2] , **Zhijin Liu** [2] , **Paul A Garber** [3] , **Huijuan Pan** [Corresp. 1]

[1] Beijing Forestry University, School of Nature Conservation, Beijing, China

[2] Chinese Academy of Sciences, Institute of Zoology, CAS Key laboratory of Animal Ecology and Conservation Biology, Beijing, China

[3] University of Illinois, Department of Anthropology and Program in Ecology and Evolutionary Biology, Urbana, Illinois, America

Corresponding Author: Huijuan Pan
Email address: phjjanine@bjfu.edu.cn

Background. Endogenous retroviruses (ERV) are remnants of former exogenous retroviruses that had previously invaded the germ line of the host that can be vertically transmitted across generations. While the majority of ERVs lack infectious capacity due to the accumulation of deleterious mutations, some ERVs remain active and produce potentially infectious viral particles. ERV sequences have been reported in all mammals, however, the distribution and diversity of ERVs in several primate taxa remains unclear. The aim of this study was to identify and classify the ERV sequences in the genomes of the golden snub-nosed monkey (*Rhinopithecus roxellana*) and the black and white snub-nosed monkey (*Rhinopithecus bieti*), two Endangered primate species that exploit high altitude (2,500-4,500m) temperate forests in southern and central China.

Methods. We used a TBLASTN program to search the ERV sequences of golden snub-nosed monkey genome and the black and white snub-nosed monkey genome. We retrieved all complete accession sequences from the homology search and then used the program, RetroTector, to check and identify the ERV sequences.

Results. We identified 284 and 263 endogenous retrovirus sequences in *R. roxellana* and *R. bieti* respectively. The proportion of full-length sequences of all ERV was 30% in *R. roxellana* and 21% in *R. bieti* and they were described as class I and class Ⅱ or gamma-retrovirus and beta-retrovirus genera. The truncation pattern distribution in the two species was virtually identical. By analyzing and comparing ER V orthologues among 6 primate species, we identified the co-evolution of ERVs with their host. We also examined ERV-like sequences and found 48 such genes in *R. roxellana* and 63 in *R. bieti*. Some of those genes are associated with diseases, suggesting that ERVs might have involved the abnormal expression of certain genes that have contributed to deleterious consequences for the host.

Conclusions. Our results indicate that ERV sequences are widely distributed in snub-nosed monkeys, and their phylogenetic history can mirror that of their hosts over long evolutionary time scales. In addition, ERV sequences appear to have an important influence on the evolution of host pathology.

1 **Genome-wide characterization of endogenous retroviruses in**

2 **snub-nosed monkeys**

4 Xiao Wang[1], Boshi Wang[2], Zhijin Liu[2], Paul A Garber[3], Huijuan Pan[1]

6 [1] School of Nature Conservation, Beijing Forestry University, Beijing, China
7 [2] CAS Key laboratory of Animal Ecology and Conservation Biology, Institute of Zoology,
8 Chinese Academy of Sciences, Beijing, China
9 [3] Department of Anthropology and Program in Ecology and Evolutionary Biology, University of
10 Illinois, Urbana, Illinois, USA

12 Corresponding Author:
13 Huijuan Pan[1]
14 Beijing Forestry University, No.35 East Qinghua Road, Haidian District, Beijing, 100083, China
15 Email address: phjjanine@bjfu.edu.cn

Abstract

Background. Endogenous retroviruses (ERVs) are remnants of former exogenous retroviruses that had previously invaded the germ line of the host that can be vertically transmitted across generations. While the majority of ERVs lack infectious capacity due to the accumulation of deleterious mutations, some ERVs remain active and produce potentially infectious viral particles. ERV sequences have been reported in all mammals, however, the distribution and diversity of ERVs in several primate taxa remains unclear. The aim of this study was to identify and classify the ERV sequences in the genomes of the golden snub-nosed monkey (*Rhinopithecus roxellana*) and the black and white snub-nosed monkey (*Rhinopithecus bieti*), two Endangered primate species that exploit high altitude (2,500-4,500m) temperate forests in southern and central China.

Methods. We used a TBLASTN program to search the ERV sequences of golden snub-nosed monkey genome and the black and white snub-nosed monkey genome. We retrieved all complete accession sequences from the homology search and then used the program, RetroTector, to check and identify the ERV sequences.

Results. We identified 284 and 263 endogenous retrovirus sequences in *R. roxellana* and *R. bieti* respectively. The proportion of full-length sequences of all ERV was 30% in *R. roxellana* and 21% in *R. bieti* and they were described as class Ⅰ and class Ⅱ or gamma-retrovirus and beta-retrovirus genera. The truncation pattern distribution in the two species was virtually identical. By analyzing and comparing ERV orthologues among 6 primate species, we identified the co-evolution of ERVs with their host. We also examined ERV-like sequences and found 48 such genes in *R. roxellana* and 63 in *R. bieti*. Some of those genes are associated with diseases, suggesting that ERVs might have involved the abnormal expression of certain genes that have contributed to deleterious consequences for the host.

Conclusions. Our results indicate that ERV sequences are widely distributed in snub-nosed monkeys, and their phylogenetic history can mirror that of their hosts over long evolutionary time scales. In addition, ERV sequences appear to have an important influence on the evolution of host pathology.

Introduction

44

45 Endogenous retroviruses (ERVs) are remnants of ancient retrovirus infections that have entered a

46 species germ line and can be transmitted vertically to that host's offspring (Wu et al. 2016).

47 These sequences, have been detected in all vertebrates genomes including humans, and they

48 often occupy a intermediate fraction of the genome (Blikstad et al. 2008; Gifford & Tristem

49 2003). For example, ERVs contribute to approximately 8% of the human genome and 10% of the

50 mouse genome (Griffiths 2001; Subramanian et al. 2011; Wu et al. 2016).

51

52 Following recent classification, the International Committee on the Taxonomy of Viruses

53 (ICTV) distinguishes seven genera within the retrovirus family: i.e., alpha-retrovirus, beta-

54 retrovirus, delta-retrovirus, gamma-retrovirus, epsilon-retrovirus, spuma-virus and lenti-virus

55 (Gifford & Tristem 2003). The Retroviridae viral family can be divided into three classes, with

56 gamma-retrovirus and epsilon-retrovirus are assigned to class Ⅰ, beta-retrovirus, alpha-

57 retrovirus, delta-retrovirus and lenti-virus are assigned to class Ⅱ, and those clustering with the

58 spuma-virus are assigned to class Ⅲ (Stocking & Kozak 2008; Vargiu et al. 2016; Villesen et al.

59 2004). Generally, a complete ERV element can be represented as 5'LTR-PBS-*Gag-Pro-Pol-Env*-

60 PPT-3'LTR (Tongyoo et al. 2017). Two LTRs contain regulatory sequences that can alter the

61 expression, splicing , and polyadenylation of those host genes located near the ERV insertion site

62 (Hayward & Katzourakis 2015). The LTRs are generated by reverse transcription and are

63 attached to a primer binding site (PBS) and a polypurine tract (PPT). *Gag* genes encode the

64 structural components of the viral core that includes the capsid (CA), matrix (MA) and

65 nucleocapsid (NC) protein; *Pro* genes encode the viral protease (PR); *Pol* genes encode the

66 reverse transcriptase (RT) and integrase (IN) enzymes; and *Env* genes encode the surface (SU)

67 and transmembrane (TM) protein (Grandi & Tramontano 2018b; Stocking & Kozak 2008).

68

69 ERVs represent genomic fossils of past retroviral infections and can inform us of the diversity

70 and history of retroviruses that have infected a species lineage (Zhuo et al. 2013). At present,

71 studies of ERVs in primates have principally focused on the family Hominidae (humans and

72 apes) and the subfamily cercopithecinae (certain African and Asian monkeys including baboons,

73 macaques, guenons). There is growing evidence that ERVs have played an important role in the

74 evolution of many mammalian lineages by providing new functions and evolutionary stimuli

75 (Deininger et al. 2003; Grandi & Tramontano 2018b; Heidmann et al. 2009). In humans, some

76 HERVs are considered to be involved in various diseases. For example, the HERV-W *Env*

77 mRNA was selectively upregulated in brain tissue from individuals with multiple sclerosis

78 (Brutting et al. 2017; Christensen 2017; Tselis 2011). In addition, a significantly higher HERV-

79 K10, MSRV and ERV-FRD activity was detected in the brains of patients with schizophrenia

80 (Brutting et al. 2017). Although several studies have investigated the pathogenic role of HERVs,

81 the precise link between any HERV sequence (and its expressed products) and human diseases

82 still remains unclear (Grandi & Tramontano 2018a). In general, the majority of research has

83 focused on the characterization of ERV sequences and how this relates to the distribution and

84 evolutionary history of primates. For example, BaEV, has been found only in baboons, geladas,
85 mangabeys and in one subspecies of green monkey (*Chlorocebus*), indicating a strong
86 phylogenetic relationship (van der Kuyl et al. 1995). Similarity, the expressed level of one family
87 of ERV, ERV-W Env, is significantly higher than other ERV families in the rhesus macaque
88 (*Macaca mulatta*). This suggests that in the rhesus macaque, retroviral envelope proteins may
89 play a more crucial role than the other ERV families in the evolutionary history of this species
90 (Eo et al. 2014).
91
92 Although nonhuman primates represent model organisms for studies on endogenous retroviral
93 diseases, studies of ERVs in the subfamily Colobinae are uncommon and lag behind other
94 primate radiations. This likely results from the fact that given specializations of their digestive
95 tract (colobines are foregut fermenters and possess a sacculated and low acid stomach that
96 contains a highly diverse microbial community), and their consumption of a diet high in fiber,
97 colobines are difficult to keep in captivity (Guo et al. 2018). Here we examine the distribution
98 and classification of ERVs and function of ERV in two species of Asian colobines of the genus
99 *Rhinopithecus*.
100
101 The Colobinae represent a highly successful radiation of 10 genera and over 76 species of
102 African and Asian monkeys (IUCN, 2016) (Brandon-Jones et al. 2004). The sacculated stomach
103 of colobines enable them to exploit difficult to digest foods that are high in structural
104 carbohydrates and toxins such as tree bark, mature leaves, lichen, and seeds (Ross 1995; Zhou et
105 al. 2014). Snub-nosed monkeys comprise a group of five extant species of Asian colobines that
106 are found in China (4 species), Myanmar (1 species) and Vietnam (1 species). All five species
107 are among the world's rarest and most endangered primates (IUCN, 2000). The Chinese species
108 of snub-nosed monkeys are unique because they are distributed in high altitude mountainous at
109 elevations of from 2,000 to 4,500 meters above sea level. In the case of *R. roxellana* and *R. bieti*,
110 during the winter, snow may cover the ground for several months and nighttime temperatures
111 drop below zero degrees centigrade (Guo et al. 2018; Long et al. 1994). A comparison of these
112 two closely related species offers a framework to examine genetic and evolutionary
113 differentiation that characterizes this primate lineage.
114
115 Recent advancements in genome sequencing offers the strongest method to sample ERV
116 diversity. Complete genome sequence data allow us to investigate the distribution and diversity
117 of ERV sequences within specific genomes in precise detail, and compare these details among
118 different genomes (Gifford & Tristem 2003). Here, we used bioinformatics tools to identify ERV
119 sequences in *R. roxellana* and *R. bieti* using RetroTector (Sperber et al. 2007). We also identified
120 neighboring genes of ERV sequences and discuss the possible effects of ERVs on gene
121 regulation and their potential contribution in understanding how the structure and function of
122 these sequences impact primate biology and health.

Materials & Methods

Identification of endogenous retrovirus

We used Gag protein and Env protein sequences of the gibbon ape leukemia virus (GenBank number: NP_056791.2; NP_056789.1) as a query and the TBLASTN program (Altschul et al. 1997) to search the golden snub-nosed monkey genome (GCA_000769185.1) and the black and white snub-nosed monkey genome (GCA_001698545.1). We obtained an extensive set of endogenous retroviruses in the sequenced genome and found approximately 100 hits for each species (the TBLASTN search threshold was set as 47% identity and total score was between 1710 to 253). Then we uploaded the matched sequences to RetroTector online (ROL) to examine their structure and conducts a detailed analysis (Sperber et al. 2007). RetroTector (ReTe) is a platform-independent Java program for identification and characterization of provirus sequences in the vertebrate genome. The ROL is a light version of ReTe. ROL implementation (http://retrotector.neuro.uu.se/pub/queue.php), allows GenBank accession number, file, and FASTA cut-and-paste admission of sequences (5 to 1000000 kilobases). ROL can identify full integration and estimate the open reading frame. Moreover, it can detect proviruses a *priori* and is not dependent on repetition, giving one the capacity to identify low-copy number retroviral sequences (Sperber et al. 2009; Vargiu et al. 2016). This allowed us to obtain a detailed analysis of retroviral sequences found in the submitted sequences, which we then viewed with the program, RetroTectorViewer.jar.

Multiple sequence alignment and phylogenetic analysis

For a complete and accurate analysis, we selected a total of 141 complete full-length ERV sequences that contained 5'LTR, PBS, *Gag*, *Pro*, *Pol*, *Env*, PPT and 3'LTR structure, including 84 GERV sequences and 55 BERV sequences (Supplemental files 1-4). A multiple alignment was constructed from the DNA sequences of the GERV and BERV using BioEdit v7.0.5 (Hall 1999). Maximum likelihood (ML) phylogenies were estimated using nucleotide sequence alignments with Standard RAxML v8.1.17 (Stamatakis 2014) with 1000 bootstrap replicates. All ERV sequences were classified to different genera based on sequences similarity to known ERV sequences for vertebrates (Table 1). Phylogenetic trees were analyzed and adjusted using MEGA7 and iTOL v4 (Letunic & Bork 2016)(http://itol.embl.de/).

Identification of similar ERV sequences and Neighboring genes

Flanking sequences on both sides of the GERV and BERV sequences were selected to identify ERV-like sequences in the genome. The BLAST program was used to measure the similarity of all flanking sequences of GERV and BERV. We detected 13 sequences that were significantly similar with 0 e-value score and query cover⩾85%. These 13 GERV and BERV sequences also had high similarities. To discover additional ERV sequences in the whole genome and to reveal viruses that may have infected primates, we used ERV sequences (RR146/RB237) as a query sequence to query the whole genome sequences in primates with that of available whole genome sequences. For each potential ERV sequences, the RetroTector program was used to check their

163   structure. We found ERV-like sequences in *Homo sapiens* (PRJNA168, GCA_000001405.27),

164   *Macaca fascicularis* (PRJNA215851, GCA_000364345.1), *Macaca mulatta* (PRJNA16397,

165   GCA_000772875.3), *Piliocolobus tephrosceles* (PRJNA419387, GCA_002776525.1),

166   *Theropithecus gelada* (PRJNA477372, GCA_003255815.1) and *Pan troglodytes* (PRJNA10627,

167   GCA_002880755.3). We used the same procedure indicated above to align sequences and build

168   a phylogenetic tree.

169

170   The region information of genes and the position information of ERV sequences were obtained

171   to determine the ERV neighboring genes. The flanking sequences of ERV were used to detect

172   the homology genes in the human genome. Finally, we compared the similarities between the

173   ERV with the flanking sequences as well as the same position of human genes obtained in the

174   previous step by BLAST in NCBI.

Results

175

Endogenous retroviruses identification

176

177 Consistent with the consensus nomenclature used for the human endogenous retrovirus (HERV)
178 (Boeke & Stoye 1997) and the chimpanzee endogenous retrovirus (CERV) (Polavarapu et al.
179 2006), here we refer to the golden snub-nosed monkey endogenous retrovirus by the acronym
180 GERV and black and white snub-nosed monkey endogenous retrovirus by the acronym BERV.
181 Using the procedure described above, we identified a total of 284 GERVs with an average length
182 of 8229 base-pairs and 263 BERVs with an average length of 8068 base-pairs (Supplemental
183 files 3 and 4). For a complete and accurate analysis, we selected full-length ERV sequences
184 which contained four major genes (*gag*, *pro*, *pol* and *env*), two sites critical to replication (PBS
185 and PPT) and flanking with 5'LTR and 3'LTR (Supplemental files 1 and 2). We found 85 full-
186 length GERVs with an average length of 9364 base-pairs and 56 full-length BERVs with an
187 average length of 8867 base-pairs. The proportion of full-length sequences of all endogenous
188 retroviruses was 30% in the golden snub-nosed monkey and 21% in black and white snub-nosed
189 monkey.

190

191 According to the 5'LTR-PBS-*Gag-Pro-Pol-Env*-PPT-3'LTR structure of the endogenous
192 retrovirus, we detected different types of truncation patterns. The proportion of each truncation
193 pattern is shown in Fig. 1. Nine classification types of truncation patterns were detected:
194 complete element, 5'-truncated, PBS truncated, *gag* truncated, *pro* truncated, *pol* truncated, *env*
195 truncated, PPT truncated and 3'-truncated. As indicated in Fig. 1, the truncation pattern
196 distribution in the golden snub-nosed monkey is similar to that in the black and white snub-nosed
197 monkey. While the majority of truncation patterns in the golden snub-nosed monkey were PPT
198 truncations and 3'-truncations, truncations in black and white snub-nosed monkey were enriched
199 with PPT and PBS truncations. Overall, the fewest truncation patterns in each of the two species
200 were *gag* truncations and *pol* truncations.

201

Grouping the endogenous retroviruses into classes and genera

202

203 We used 33 full-length endogenous retroviruses sequences (Table 1) to construct a phylogenetic
204 tree using BioEdit v7.0.5 and RAxML v8.1.17. The resulting 33 endogenous retroviruses were
205 grouped into three classes and seven genera based on the bootstrap values generated in the
206 phylogenetic tree (Fig. 2). Class Ⅰ contains elements related to the gamma-retroviruses and
207 epsilon-retroviruses. Class Ⅱ elements are related to alpha-retroviruses, beta-retroviruses, delta-
208 retroviruses and lentiviruses. Class Ⅲ elements are related to spuma-viruses.

209

210 To better understand the classification of GERVs and BERVs, we further performed a
211 phylogenetic analysis based on 84 full-length GERVs and 33 reference endogenous retrovirus
212 sequences of 7 genera (Fig. 3). Of the 84 GERVs identified in our study, 68 GERVs groups were
213 class Ⅰ which were related to gamma-retroviruses. Only 7 GERVs groups in class Ⅱ were

214 related to beta-retroviruses. Class Ⅰ GERVs accounted for 81% and class Ⅱ GERVs accounted
215 for 8.3%. Class Ⅰ GERVs are the most abundant endogenous retroviruses in the golden snub-
216 nosed monkey while only a small number of class Ⅱ endogenous retroviruses were present. We
217 also conducted a phylogenetic analysis based on 55 full-length BERVs and 33 reference
218 endogenous retroviruses sequences characteristic of the 7 genera (Fig. 4). In the tree shown, 21
219 BERVs groups of class Ⅰ belonged to gamma-retroviruses genera, and 33 BERVs groups of
220 class Ⅱ belonged to beta-retroviruses. Class Ⅰ BERVs accounted for 38.2% and class Ⅱ
221 BERVs accounted for 60%.
222
223 **Identification of similar ERV sequences and inferring virus evolutionary history**
224 Golden snub-nosed monkeys and black and white snub-nosed monkeys were found to exhibit
225 several common ERV sequences, i.e., 13 GERV and BERV sequences that were virtually
226 identical, including their flanking sequences. In order to study the evolutionary history of ERVs
227 in primates, we used RR146/RB237 as a query sequence to search all primate genomes in the
228 NCBI database. We found 13 species which had an ERV similar to RR146/RB237. However,
229 only 6 of these species were detected by the RetroTector program. To understand their
230 evolutionary relationships, we established a phylogenetic tree among ERV orthologues from
231 *Homo sapiens*, *Macaca fascicularis*, *Macaca mulatta*, *Piliocolobus tephrosceles*, *Theropithecus*
232 *gelada* and *Pan troglodytes*. The resulting phylogeny revealed two major lineages, the first
233 containing the Old-World monkeys (Cercopithecoidea), including *Macaca fascicularis*, *Macaca*
234 *mulatta*, *Piliocolobus tephrosceles*, *Theropithecus gelada*, *Rhinopithecus roxellana* and
235 *Rhinopithecus bieti* (Fig. 5). The second lineage included *Homo sapiens* and *Pan troglodytes*,
236 which belong to the Hominoidea. Given that our sampling of ERV diversity was limited, these
237 results do not fully reflect the evolutionary history of the primate endogenous retroviruses.
238 Moreover, based on our results (Fig. 5), the evolutionary history of ERVs tends to broadly follow
239 the evolutionary history of their hosts. These data also revealed that virus phylogenetic history
240 can mirror that of their hosts over long evolutionary timescales.
241
242 **Endogenous retroviruses neighboring gene profiles**
243 We found 48 and 63 neighboring genes in the golden snub-nosed monkey and black and white
244 snub-nosed monkey, respectively (Supplemental files 5 and 6). By comparing these ERV
245 neighboring gene sequences with their homologues in humans, we found, for example, the *PEX3*
246 gene, *ALDH1A1* gene and *ARL11* gene contain the ERV sequence in the golden snub-nosed
247 monkeys, but not humans. In contrast, the *PNPT1* gene and the *DOCK5* gene, which contain
248 ERV sequences, are identical in these two species. In the black and white snub-nosed monkey,
249 the *GPC5* gene and the *TLR5* gene contain ERV sequences while in humans, these genes do not.
250 However, the *CHEK2* gene and *DOCK5* gene in humans and in black and white snub-nosed
251 monkey both contain an ERV sequence (Fig. 6). We also found that the *DOCK5* gene is present

252   in the golden snub-nosed monkey, the black and white snub-nosed monkey and humans, and the
253   alignment results are highly similar.

254 **Discussion**

255 As part of the ERV infection process, a viral particle can become part of the germline of its host
256 and then vertically transmitted to host's offspring and future generations. Some elements are able
257 to amplify within the genome and increase their copy number, leading to a larger number of
258 ERV fragments (Dewannieux & Heidmann 2013). Here, we have identified 284 ERVs in the
259 genome of the golden snub-nosed monkey and 263 ERVs in the genome of the black and white
260 snub-nosed monkey. The proportion of ERV fragments of all endogenous retroviruses was 70%
261 in *R. roxellana* and 79% in *R. bieti*. We assume that the accumulation of nonsense mutations,
262 insertions, and deletions of internal coding regions and long terminal repeats resulted in the
263 inactivity of ERV sequences over evolutionary time, and those ERV sequences gradually
264 become incomplete and exhibit many truncation patterns in the genome. Based on phylogenetic
265 relationships (Fig. 3 and Fig. 4), we found that GERV sequences and BERV sequences mostly
266 grouped with class Ⅰ and class Ⅱ ERVs. There was no ERV sequence related to delta-
267 retroviruses or lenti-retroviruses. In addition, there was no ERV sequences in class Ⅲ. This may
268 be due to the fact that delta-retroviruses, spuma-retroviruses and lenti-retroviruses are all
269 complex retroviruses. They code for regulatory proteins with different functions, in addition to
270 gag, pro, pol and env proteins (Geng 2012). A further explanation for why these genera have not
271 been previously identified is that the trans-acting regulatory gene products of these complex
272 genera have precluded germline integration (Gifford & Tristem 2003). Alpha-retroviruses have
273 only been identified previously in avians of the genus *Gallus* (e.g. pheasants) (Gifford & Tristem
274 2003). Secondly, the query sequence we used in Blast is the Gag protein and Env protein
275 sequence of the Gibbon ape leukemia virus, which is a gamma-retrovirus. Gamma-retroviruses
276 and beta-retroviruses both have a simple genome organization. Finally, gamma-retroviruses are
277 the most commonly found of all known retroviruses while alpha-retroviruses, delta-retroviruses
278 and lenti-retroviruses are less common (Vargiu et al. 2016). Therefore, the proportion of gamma-
279 retroviruses and beta-retroviruses in our results is relatively high.
280
281 ERVs are members of the LTR elements, which represent the most complex elements of
282 retroelements (Khodosevich et al. 2002). The LTRs of ERVs contain many regulatory sequences,
283 such as promoters, enhancers, polyadenylation signals and factor-binding sites. Some of these
284 ERVs could have integrated into regulatory regions of the genome, and they may influence the
285 expression of nearby genes, which have consequently contributed to host evolution or associated
286 with diseases (Goodier 2016; Khodosevich et al. 2002). This means that genes near the ERV
287 could be affected by the promoter in the LTR of ERVs. For instance, HERV immunosuppressive
288 functions might contribute to cancer progression by reducing the immune recognition and attack
289 of tumor cells (Grandi & Tramontano 2018b). Thus, ERVs we have found in the golden snub-
290 nosed monkey and black and white snub-nosed monkey genomes may have influenced the
291 expression of nearby genes and therefore associated with increased susceptibility to diseases.
292 However, additional studies are needed to evaluate the potential impact of ERV expression on

293 host health and its role in regulating signaling pathways involved in the manifestation of
294 particular diseases.
295

Conclusions

297 Our analysis of the completed published golden snub-nosed monkey and black and white snub-
298 nosed monkey genomes has identified 284 GERV and 263 BERV sequences, most of which are
299 from genera of gamma-retroviruses and beta-retroviruses. The proportion of full-length
300 sequences of all endogenous retroviruses is 30% in the golden snub-nosed monkey and 21%
301 black and white snub-nosed monkey. However, more data are needed to determine whether these
302 structurally complete ERV sequences remain active. In addition, the relationship between these
303 ERV neighboring genes and diseases is unclear. In summary, our results document a co-
304 evolutionary process between ERVs and their primate hosts that has occurred over the past
305 several million years. Further analysis should allow us to better understand ERVs and their role
306 in host protection.
307

Acknowledgements

309 PAG wishes to thank Chrissie McKenney, Sara Garber, and Jenni Garber for their support.

References

310
311 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997.
312   Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
313   *Nucleic Acids Research* 25:3389-3402. 10.1093/nar/25.17.3389
314 Blikstad V, Benachenhou F, Sperber GO, and Blomberg J. 2008. Evolution of human
315   endogenous retroviral sequences: a conceptual account. *Cellular and Molecular Life*
316   *Sciences* 65:3348-3365. 10.1007/s00018-008-8495-2
317 Boeke JD, and Stoye JP. 1997. Retrotransposons, Endogenous Retroviruses, and the Evolution of
318   Retroelements: Cold Spring Harbor Laboratory Press.
319 Brandon-Jones D, Eudey AA, Geissmann T, Groves CP, Melnick DJ, Morales JC, Shekelle M,
320   and Stewart CB. 2004. Asian Primate Classification. International Journal of Primatology
321   25:97-164. 10.1023/b:ijop.0000014647.18720.32
322 Brutting C, Emmer A, Kornhuber ME, and Staege MS. 2017. Cooccurrences of Putative
323   Endogenous Retrovirus-Associated Diseases. *Biomed Research International*
324   2017:7973165. 10.1155/2017/7973165
325 Christensen T. 2017. Human endogenous retroviruses in the aetiology of MS. *Acta Neurol Scand*
326   136 Suppl 201:18-21. 10.1111/ane.12836
327 Deininger PL, Moran JV, Batzer MA, and Jr HHK. 2003. Mobile elements and mammalian
328   genome evolution. *Current Opinion in Genetics & Development* 13:651-658.
329   10.1016/j.gde.2003.10.013
330 Dewannieux M, and Heidmann T. 2013. Endogenous retroviruses: acquisition, amplification and
331   taming of genome invaders. *Current Opinion in Virology* 3:646-656.
332   10.1016/j.coviro.2013.08.005
333 Eo J, Cha HJ, Imai H, Hirai H, and Kim HS. 2014. Short communication: expression profiles of
334   endogenous retroviral envelopes in *Macaca mulatta* (rhesus monkey). *Aids Research &*
335   *Human Retroviruses* 30:996-1000. 10.1089/AID.2014.0010
336 Geng Y. 2012. Molecular Biology of Lentivirus and Foamy virus: Science Press.
337 Gifford R, and Tristem M. 2003. The Evolution, Distribution and Diversity of Endogenous
338   Retroviruses. *Virus Genes* 26:291-315. 10.1023/a:1024455415443
339 Goodier JL. 2016. Restricting retrotransposons: a review. *Mobile DNA* 7:16. 10.1186/s13100-
340   016-0070-z
341 Grandi N, and Tramontano E. 2018a. HERV Envelope Proteins: Physiological Role and
342   Pathogenic Potential in Cancer and Autoimmunity. *Frontiers in Microbiology* 9:462.
343   10.3389/fmicb.2018.00462
344 Grandi N, and Tramontano E. 2018b. Human Endogenous Retroviruses Are Ancient Acquired
345   Elements Still Shaping Innate Immune Responses. *Frontiers in Immunology* 9:2039.
346   10.3389/fimmu.2018.02039
347 Griffiths DJ. 2001. Endogenous retroviruses in the human genome sequence. *Genome Biology*
348   2:1-5. 10.1186/gb-2001-2-6-reviews1017
349 Guo ST, Hou R, Garber PA, Raubenheimer D, Righini N, Ji WH, Jay O, He SJ, Wu F, and Li
350   FF. 2018. Nutrient-specific compensation for seasonal cold stress in a free-ranging
351   temperate colobine monkey. *Functional Ecology*. 10.1111/1365-2435.13134
352 Hall TA. 1999. BioEdit : a user-friendly biological sequence alignment editor and analysis
353   program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98. 10.1021/bk-
354   1999-0734.ch008
355 Hayward A, and Katzourakis A. 2015. Endogenous retroviruses. *Current Biology* 25:R644-

356   R646. 10.1016/j.cub.2015.05.041

357   Heidmann T, Dewannieux M, Dupressoir A, Esnault C, Heidmann O, Lavialle C, Louf G, Ribet

358   D, and Vernochet C. 2009. Endogenous retroviruses: from infectious elements to bona

359   fide genes with a physiological role. *Retrovirology* 6:1-1. 10.1186/1742-4690-6-S2-I2

360   Khodosevich K, Lebedev Y, and Sverdlov E. 2002. Endogenous retroviruses and human

361   evolution. *Comparative and Functional Genomics* 3:494-498. 10.1002/cfg.216

362   Letunic I, and Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and

363   annotation of phylogenetic and other trees. *Nucleic Acids Research* 44:W242-W245.

364   10.1093/nar/gkw290

365   Long Y, Kirkpatrick CR, Zhongtai, and Xiaolin. 1994. Report on the distribution, population,

366   and ecology of the yunnan snub-nosed monkey (Rhinopithecus bieti). *Primates* 35:241-

367   250. 10.1007/bf02382060

368   Polavarapu N, Bowen NJ, and Mcdonald JF. 2006. Identification, characterization and

369   comparative genomics of chimpanzee endogenous retroviruses. *Genome Biology* 7:R51.

370   10.1186/gb-2006-7-6-r51

371   Ross C. 1995. Colobine Monkeys: Their Ecology, Behaviour and Evolution. International

372   *Journal of Primatology* 16:1035-1037. 10.1007/BF02696118

373   Sperber G, Lövgren A, Eriksson NE, Benachenhou F, and Blomberg J. 2009. RetroTector online,

374   a rational tool for analysis of retroviral elements in small and medium size vertebrate

375   genomic sequences. *Bmc Bioinformatics* 10:S4-S4. 10.1186/1471-2105-10-S6-S4

376   Sperber GO, Airola T, Jern P, and Blomberg J. 2007. Automated recognition of retroviral

377   sequences in genomic data--RetroTector. *Nucleic Acids Research* 35:4964.

378   10.1093/nar/gkm515

379   Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of

380   large phylogenies. *Bioinformatics* 30:1312-1313. 10.1093/bioinformatics/btu033

381   Stocking C, and Kozak CA. 2008. Murine endogenous retroviruses. *Cellular and Molecular Life*

382   *Sciences* 65:3383-3398. 10.1007/s00018-008-8497-0

383   Subramanian RP, Wildschutte JH, Russo C, and Coffin JM. 2011. Identification,

384   characterization, and comparative genomic distribution of the HERV-K (HML-2) group

385   of human endogenous retroviruses. *Retrovirology* 8:90. 10.1186/1742-4690-8-90

386   Tongyoo P, Avihingsanon Y, Promon S, Mutirangura A, Mhuantong W, and Hirankarn N. 2017.

387   EnHERV: Enrichment analysis of specific human endogenous retrovirus patterns and

388   their neighboring genes. *PLoS One* 12:e0177119. 10.1371/journal.pone.0177119

389   Tselis A. 2011. Evidence for viral etiology of multiple sclerosis. *Seminars in Neurology* 31:307-

390   316. 10.1055/s-0031-1287656

391   van der Kuyl AC, Dekker JT, and Goudsmit J. 1995. Distribution of baboon endogenous virus

392   among species of African monkeys suggests multiple ancient cross-species transmissions

393   in shared habitats. *Journal of Virology* 69:7877-7887. 10.3109/13550289509111034

394   Vargiu L, Rodriguez-Tome P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E,

395   and Blomberg J. 2016. Classification and characterization of human endogenous

396   retroviruses; mosaic forms are common. *Retrovirology* 13:7. 10.1186/s12977-015-0232-y

397   Villesen P, Aagaard L, Wiuf C, and Pedersen FS. 2004. Identification of endogenous retroviral

398   reading frames in the human genome. *Retrovirology* 1:32-32. 10.1186/1742-4690-1-32

399   Wu HL, Leon EJ, Wallace LT, Nimiyongskul FA, Buechler MB, Newman LP, Castrovinci PA,

400   Paul Johnson R, Gifford RJ, Brad Jones R, and Sacha JB. 2016. Identification and

401   spontaneous immune targeting of an endogenous retrovirus K envelope protein in the

402   Indian rhesus macaque model of human disease. *Retrovirology* 13:6. 10.1186/s12977-
403   016-0238-0
404 Zhou X, Wang B, Pan Q, Zhang J, Kumar S, Sun X, Liu Z, Pan H, Lin Y, Liu G, Zhan W, Li M,
405   and Ren B. 2014. Whole-genome sequencing of the snub-nosed monkey provides
406   insights into folivory and evolutionary history. *Nature Genetics* 46:1303-1310.
407   10.1038/ng.3137
408 Zhuo X, Rho M, and Feschotte C. 2013. Genome-wide characterization of endogenous
409   retroviruses in the bat Myotis lucifugus reveals recent and diverse infections. *Journal of*
410   *Virology* 87:8493-8501. 10.1128/jvi.00892-13

List of sequences used for phylogenetic analysis in this study

1 Table 1

2 List of sequences used for phylogenetic analysis in this study

Virus Abbreviation	Retrovirus	GenBank Accession no.	Length
Alpha retrovirus			
ALV	Avian leukosis virus	KU375453	7746bp
RSV	Rous sarcoma virus	AF052428	9396bp
LPDV	Lymphoproliferative disease virus	U09568	7143bp
Beta retrovirus			
RERV	Rabbit endogenous retrovirus	AF480925	6300bp
SRV-1	Simian SRV-1 type D retrovirus	M11841	8173bp
MMTV	Mouse mammary tumor virus	NC_001503	8805bp
Spuma retrovirus			
HFV	Human foamy virus	Y07725	13242bp
FeFV	Feline foamy virus	AJ223851	10479bp
SFV	Macaque simian foamy virus	X54482	12972bp
SMRV	Squirrel monkey virus	GU356394	11684bp
EFV	Equine foamy virus	AF201902	12035bp
Delta retrovirus			
BLV	Bovine leukemia virus	K02120	8714bp
HTLV-1	Human T-lymphotropic virus 1	D13784	8400bp
STLV-1	Siman T-lymphotropic virus 1	Z46900	9025bp
Epsilon retrovirus			
WDSV	Walleye dermal sarcoma virus	L41838	12708bp
WEHV1	Walleye epidermal hyperplasia virus 1	AF133051	12999bp
WEHV2	Walleye epidermal hyperplasia virus 2	AF133052	13125bp
Gamma retrovirus			
BaEV	Baboon endogenous virus	X05470	8018bp
MuLV	Moloney murine leukemia virus	AF033811	8332bp
MDEV	Mus dunni endogenous virus	AF053745	8655bp
GALV	Gibbon ape leukemia virus	M26927	8088bp
FeLV	Feline leukemia virus	M18247	8440bp
PERV	Porcine endogenous retrovirus	AF038601	7333bp
KoRV	Koala type C endogenous virus	AF151794	8431bp
RMLV	Rauscher murine leukemia virus	U94692	8282bp
RaLV	Rat leukemia virus	M77194	8107bp

Lentivirus

HIV-1	Human immunodeficiency virus 1	K03454	9176bp
HIV-2	Human immunodeficiency virus 2	X05291	9671bp
SIV	simian immunodeficiency virus	M33262	10535bp
FIV	Feline immunodeficiency virus	M25381	9474bp
EIAV	Equine infectious anemia virus	AF028232	8229bp
MVV	Ovine lentivirus	M31646	9256bp
BIV	Bovine immunodeficiency virus	M32690	8482bp

Figure 1

ERV truncation patterns distribution in the golden snub-nosed monkey (*R. roxellana*) and the black and white snub-nosed monkey (*R. bieti*)

Percentage of eight ERV truncation patterns and full-length ERV sequences in *R. roxellana* and *R. bieti* .
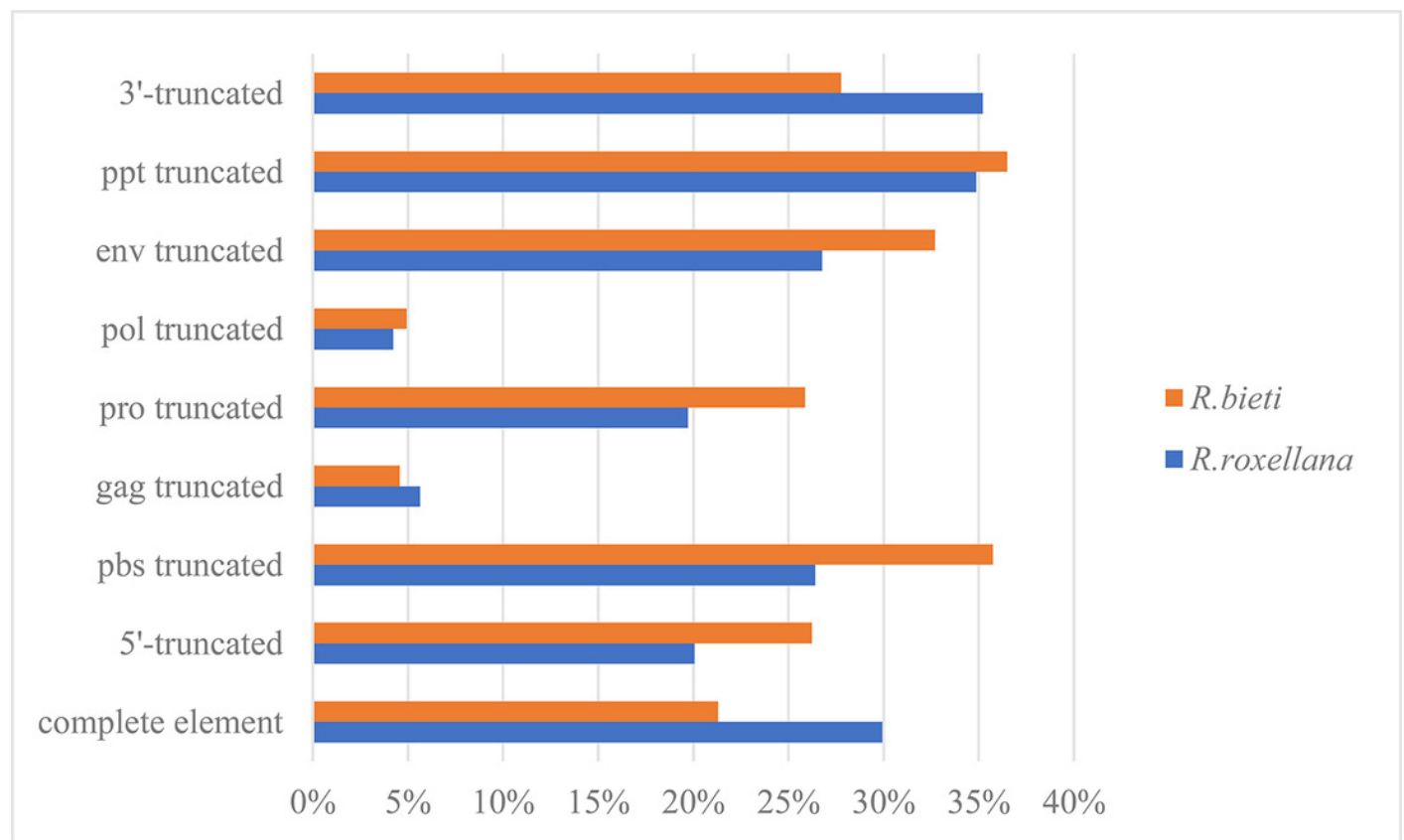
Figure 2

Phylogenetic relationship of different ERVs

Phylogenetic tree was based on 33 full-length ERV sequences in seven genera.
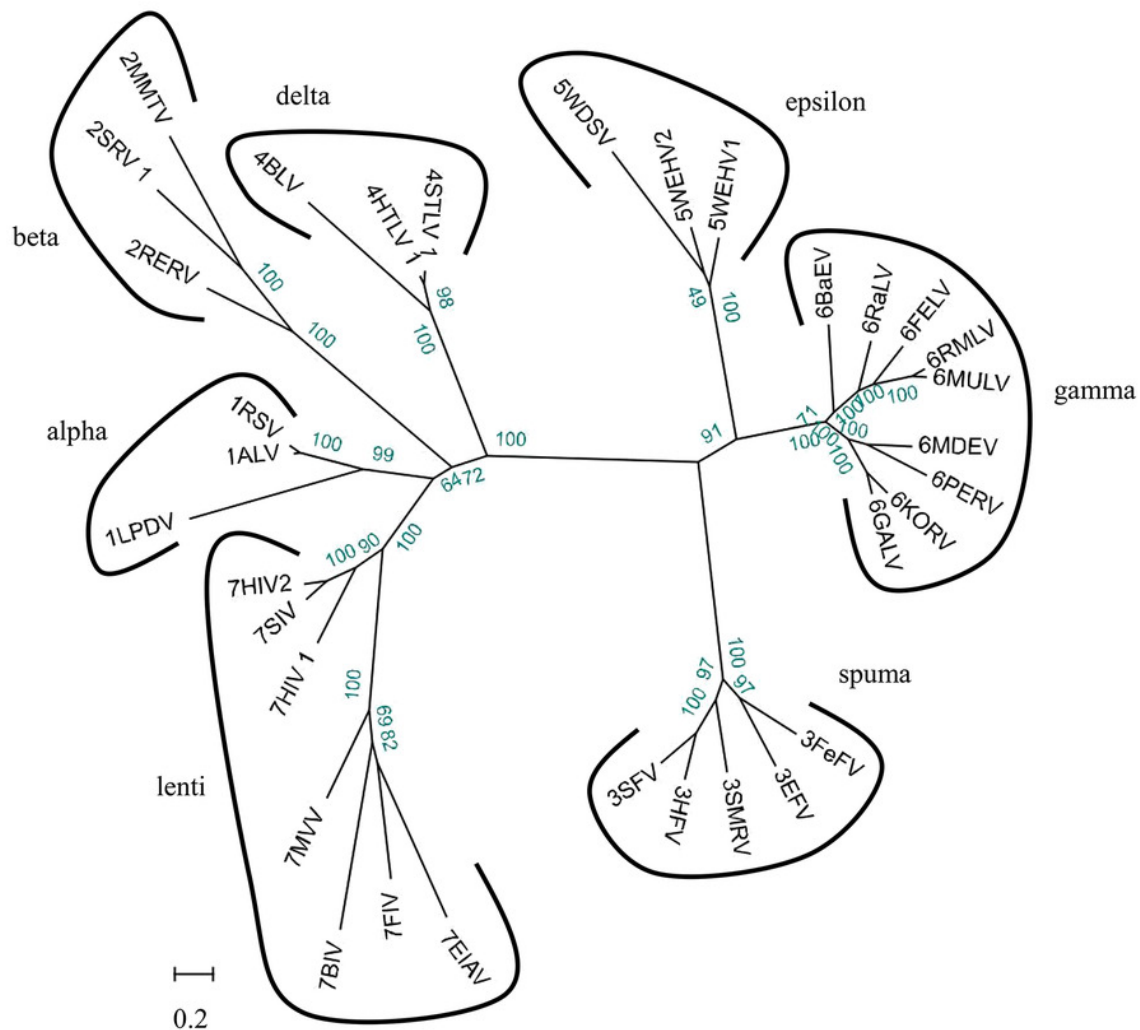
Figure 3

Phylogenetic relationship of different ERVs in golden snub-nosed monkey

Phylogenetic tree was based on the 84 GERV sequences and 33 reference ERV sequences in seven genera. Within the tree, GERV sequences are named with RR (RR stands for the acronym of *Rhinopithecus roxellana*). The major taxonomic names are shown to the right.

gamma

beta

0.50
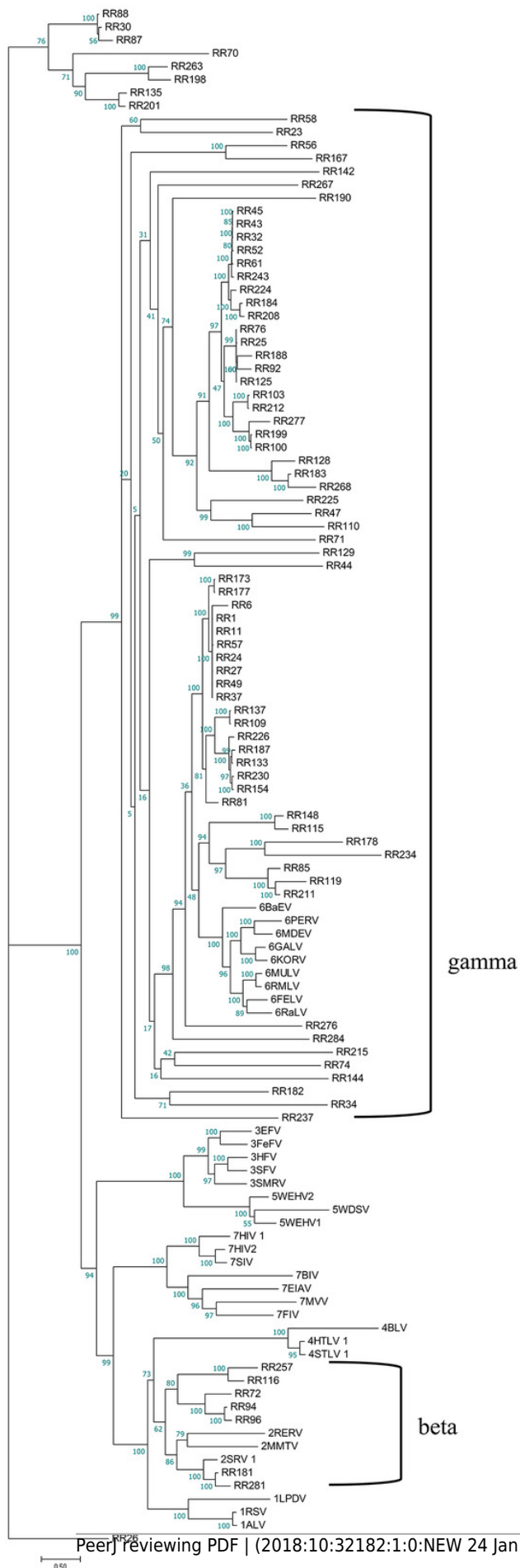
Figure 4

Phylogenetic relationship of different ERVs in black and white snub-nosed monkey

Phylogenetic tree was based on the 55 BERV sequences and 33 reference ERV sequences in seven genera. Within the tree, BERV sequences are named with RB (RB stands for the acronym of *Rhinopithecus bieti*). The major taxonomic names are shown to the right.
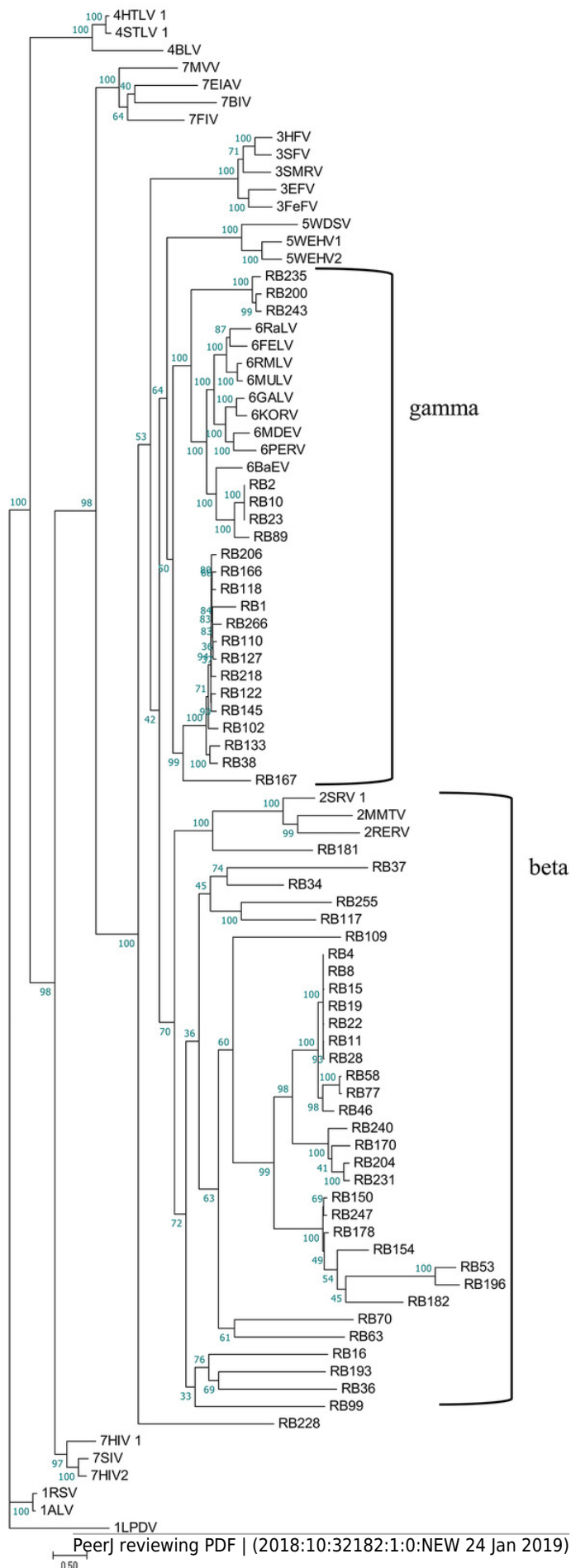
Figure 5

Phylogenetic relationship of similar ERVs

Phylogenetic tree was based on the similar ERV sequences found in eight species and showed the evolutionary relationship of the similar ERV sequences in different primate species.

Note: Auto Gamma Correction was used for the image. This only affects the reviewing manuscript. See original source image if needed for review.
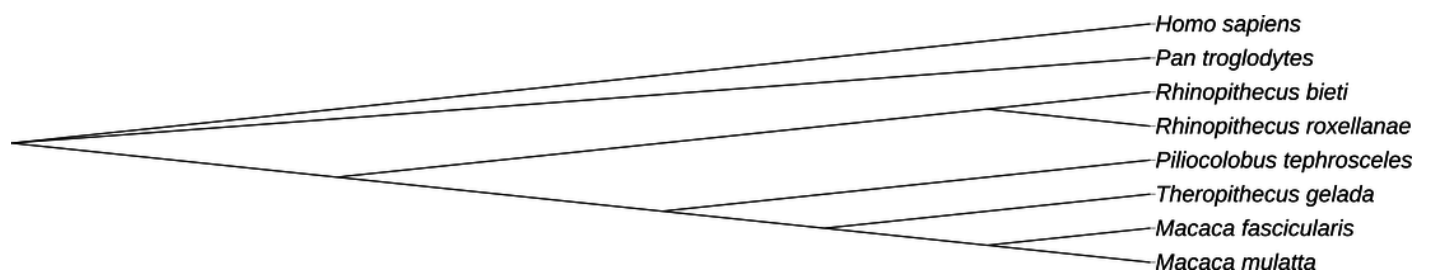
Figure 6

Alignment of human and black and white snub-nosed monkey *GPC5* and *DOCK5* partial genes

Graph showed the results of the black and white snub-nosed monkey and human genes alignment. (A) The partial sequence of the *GPC5* gene was aligned between the black and white snub-nosed monkey and human. Query sequence was the *GPC5* gene of the black and white snub-nosed monkey containing the ERV sequence. Human *GPC5* gene did not contain ERV sequence. (B) Query sequence was the *DOCK5* gene of the black and white snub-nosed monkey containing the ERV sequence. Human *DOCK5* gene also contained ERV sequence.