Comprehensive phylogenomic analysis reveals a novel cluster of simian endogenous retroviral sequences in Colobinae monkeys

Masaki Ikeda[1], Kazuhiro Satomura[1], Tsuyoshi Sekizuka[2], Kentaro Hanada[3], Toshinori Endo[1], Naoki Osada[1,4*]

[1]Department of Information Science and Technology, Hokkaido University, Hokkaido, Japan

[2]Pathogen Genomics Center, National Institute of Infectious Diseases, Tokyo, Japan

[3]Department of Biochemistry & Cell Biology, National Institute of Infectious Diseases, Tokyo, Japan

[4]Global Station for Bid Data and Cybersecurity, GI-CoRE, Hokkaido University, Hokkaido, Japan

Short title: Molecular evolution of SRVs and SERVs

*address correspondence to Naoki Osada, nosada@ist.hokudai.ac.jp

Research Highlights

- SERV sequences were comprehensively searched from seven draft genome sequences of the Old World monkeys.
- The SERVs of Colobinae monkeys formed a distinct cluster and more closely related to exogenous SRVs than the SERVs of Cercopithecinae monkeys.

Abstract

Simian retrovirus (SRV) is a type-D betaretrovirus infectious to the Old World monkeys causing a variety of symptoms. SRVs are also present in the Old World monkey genomes as endogenous forms, which are referred to as Simian endogenous retroviruses (SERVs). Although many SERV sequences have been identified in Cercopithecinae genomes, with potential of encoding all functional genes, the distribution of SERVs in genomes and evolutionary relationship between exogeneous SRVs and SERVs remains unclear. In this study, we comprehensively investigated seven draft genome sequences of the Old World monkey genomes, and identified a novel cluster of SERVs in the two *Rhinopithecus* (*R. roxellana* and *R. bieti*) genomes, which belong to the Colobinae subfamily. The *Rhinopithecus* genomes harbored higher copy numbers of SERVs than the Cercopithecinae genomes. A reconstructed phylogenetic tree showed that the Colobinae SERVs formed a distinct cluster from SRVs and Cercopithecinae SERVs, and more closely related to exogenous SRVs than Cercopithecinae SERVs. Three radical amino acid substitutions specific to Cercopithecinae SERVs, which potentially affect the infectious ability of SERVs, were also identified in the proviral envelope protein. In addition, we found many integration events of SERVs were genus- or species-specific, suggesting the recent activity of SERVs in the Old World monkey genomes. The results suggest that SERVs in Cercopithecinae and Colobinae monkeys were endogenized after the divergence of subfamilies and do not transmit across subfamilies. Our findings also support the hypothesis that Colobinae SERVs are direct ancestors of SRV-6, which has a different origin from the other exogenous SRVs. These findings shed novel insight into the evolutionary history of SERVs, and may help to understand the process of

endogenization of SRVs.

Keywords: SRV, SERV, retrovirus, evolution, Old World monkey

Introduction

The retroviral life cycle consists of integration of proviral DNA sequences into a host genome with subsequent production of active infectious particles. Proviral sequences that integrate into the genomes of host germ cells are referred to as endogenous retroviruses (ERVs) and are vertically transmitted to offspring. Over evolutionary time, most ERV sequences accumulate deleterious mutations and become inactive, although some remain active and produce potentially infectious viral particles (Stoye, 2012). Numerous ERV families are distributed within the genomes of diverse organisms, including nonhuman primates (Johnson, 2015). Nonhuman primates are widely used as model organisms for studies on retroviral infectious diseases.

Simian retrovirus (SRV), a type D betaretrovirus, was first isolated in rhesus macaque (*Macaca mulatta*) and was initially referred to as Mason–Pfizer monkey virus (MPMV) or SRV-3 (Chopra & Mason, 1970; Lerche & Osborn, 2003). Subsequent studies identified eight different serotypes of exogenous SRVs (SRV-1–8). Compared with SRV-6 and -7, SRV-1–5 and SRV-8 have been well characterized at the molecular level, and their complete genome sequences have been determined (Montiel, 2010; Power et al., 1986; Sonigo, Barker, Hunter, & Wain-Hobson, 1986; Takano, Leon, Kato, Abe, & Fujimoto, 2013; Thayer et al., 1987; Zao et al., 2010; Zao et al., 2016). SRV infections sometimes result in simian immunodeficiency diseases (SAIDS) and can also cause more acute symptoms (Yoshikawa et al., 2015). Macaque monkeys (genus *Macaca*) are the most widely recognized natural hosts of SRVs; SRV-6, however, has been isolated from a wild colobine monkey, the Hanuman langur (*Semnopithecus entellus*) (Nandi, Bhavalkar-Potdar, Tikute, & Raut, 2000; Nandi, Van Dooren, Chhangani, & Mohnot, 2006). SRV genomes are organized on the basis of the generic ERV structure: 5′-long terminal repeat (LTR), *gag*, *pro*, *pol*, *env*, and 3′-LTR, where *gag* encodes the viral core and DNA-binding proteins, *pro* encodes proteases, *pol* encodes the reverse transcriptase,

and *env* encodes the envelope glycoproteins (Figure 1).

Genomic sequences highly homologous to SRV genes were first identified in a baboon (*Papio cynocephalus*) (van der Kuyl, Mang, Dekker, & Goudsmit, 1997) and termed as simian endogenous retroviruses (SERVs). Although the sequences isolated in this study contained frameshift mutations, the basic genomic architecture of SRVs was preserved in SERVs, suggesting that SERVs are fossils of ancient SRVs. More recently, the Vero cell line, which is derived from African green monkeys (*Chlorocebus sabaeus*), was found to express mRNAs homologous to SERVs as well as another endogenous retrovirus, BaEV (Ma et al., 2011; Onions et al., 2011); this result strongly implied that African green monkey genomes also harbor SERV sequences. Confirmation was provided by Sakuma et al., who performed whole-genome sequencing of multiple Vero cell samples and showed that their genomes contained SERV sequences encoding full-length proteins without premature stop codons (Osada et al., 2014; Sakuma et al., 2018). These intact SERVs were present in the Vero genome in a heterozygous state, suggesting that they are evolutionary young ERVs that are still segregating within populations of African green monkeys. They also surveyed the draft genome sequences of four Cercopithecinae monkeys (*M. mulatta*, *M. fascicularis*, *P. anubis*, and *C. sabaeus*) and identified many additional SERV-like sequences in their genomes. Phylogenetic analysis demonstrated that exogenous SRVs (SRV-1–5 and SRV-8) and all of SERVs formed a distinct cluster, suggesting that exogenous SRVs and Cercopithecinae SERVs (cer-SERVs) had distinct evolutionary origins. However, the presence of SERVs that are more closely related to exogeneous SRVs than cer-SERVs remains unclear.

The Old Word monkeys (family Cercopithecidae) consist of two subfamilies, Cercopithecinae and Colobinae, that diverged approximately 10–18 Mya (Perelman et al., 2011; Zhou et al., 2014). Although most phylogenetic studies of SRV and SERV have been conducted in Cercopithecinae monkeys, there are less characterized classes of SRVs

and SERVs. The exogenous provirus PO-1-Lu was first isolated by Benveniste and Todaro from the Colobinae monkey *Trachypithecus obscurus* (Benveniste & Todaro, 1977; Todaro et al., 1978). Exogenous SRVs similar in sequence to PO-1-Lu were later isolated from another Colobinae monkey, *S. entellus*, and classified as SRV-6 (Nandi et al., 2000; Sommerfelt, Harkestad, & Hunter, 2003). A phylogenetic analysis using partial nucleotide sequences of SRV-6 and PO-1-Lu showed that they formed a cluster that was distinct from that formed by the other SRVs and SERVs (Sommerfelt et al., 2003). However, further studies for SRV-6 and PO-1-Lu have not been conducted, and hence the presence of SERV-like sequences in the other Colobinae monkey genomes has remained unclear. The genomic characterization of SERV-like sequences in Colobinae monkeys is therefore necessary to understand when SRVs endogenized and how they have propagated in the genomes of Old World monkeys.

Owing to the development of genome sequence databases, the assessment of endogenized retroviral sequences from genomic scan has become a powerful tool to elucidate the endogenization process of retroviruses (Diehl, Patel, Halm, & Johnson, 2016; Magiorkinis, Blanco-Melo, & Belshaw, 2015). To infer the evolutionary relationships amongst SRVs and SERVs and address the potential origins of SRVs, we performed a comprehensive search for SERV sequences within four Cercopithecinae genomes (*M. mulatta*, *M. fascicularis*, *P. anubis*, and *C. sabaeus*) and three recently sequenced Colobinae genomes (*Rhinopithecus roxellana*, *R. bieti*, and *Nasalis larvatus*) (Omar et al., 2017; Zhou et al., 2014). A phylogeny of these species inferred from the results of previous studies is shown in Figure 2.

In this study, we address four major questions about the evolutionary process of SRVs and SERVs. (1) Do Colobinae monkeys other than *T. obscurus* harbor multiple SERVs in their genomes? If so, what is the phylogenetic relationships to exogenous SRVs and Colobinae SERVs (col-SERVs)? In order to depict the pattern of SRV/SERV molecular

128 evolution, we identify SERV sequences in the seven Old World monkey genomes and
129 conduct a phylogenetic analysis. (2) Did recombination between exogeneous SRVs and
130 SERVs occur in the past? Because SERVs potentially serve as a genetic source for
131 exogeneous SRVs by genetic recombination, we investigate the pattern of recombination
132 among exogenous SRVs and SERVs. (3) When were SERVs endogenized in the Old
133 World monkey genomes? In order to answer the question, we examine the
134 presence/absence of SERVs in each monkey genome and infer the timing of integration.
135 In addition, we perform molecular dating of LTR sequences. When a retrovirus integrates
136 into the host genome, both LTRs share identical nucleotide sequences. After the
137 integration event, the 5′ and 3′ LTR sequences gradually accumulate mutations and can
138 evolve independently; therefore, the sequence divergence between pairs of LTRs should
139 be proportional to the time elapsed since SERV integration under a neutrally evolving
140 model (Johnson & Coffin, 1999). (4) Do SERV protein-coding sequences exhibit
141 detectable signatures of purifying selection? If the propagation of SERV copies have not
142 been related to reinfection or retrotransposition, the ratio of nonsynonymous and
143 synonymous substitution rates (d_N and d_S, respectively) becomes close to 1 (Li, Wu, &
144 Luo, 1985). We estimated the value of d_N and d_S, for each external branch of SERVs and
145 inferred recent selective pressure on SERV proteins.

146 **Methods**

147 *SERV sequence search in Old World monkey genomes*

148 The draft genome sequences of *M. mulatta*, *M. fascicularis*, *P. Anubis*, *C. sabaeus*, *R*
149 *roxellana*, *R. bieti*, and *Nasalis larvatus* were downloaded from the NCBI Genbank
150 database. For *M. mulatta*, the Y chromosome sequence was also obtained from a literature
151 (Hughes et al., 2012). The list of assembly IDs and sample source IDs for these sequences
152 are shown in Supplementary Table 1.

153 Sequences homologous to SERVs in Old World monkeys genomes were searched using
154 the publicly available SERV sequence fully encoding all four proteins, a consensus SERV
155 sequence identified in the Vero cell line (DDBJ/EMBL/Genbank accession number:
156 AB935214), as a query. In order to minimize the bias produced by an arbitrary selected
157 query sequence and maximize the number of potential SERV sequences to be identified,
158 the searches were designed to identify distantly related SERV sequences, which contain
159 two LTRs and all four retroviral genes (*gag*, *pro*, *pol*, and *env* irrespective of their protein-
160 coding potential). The query-derived amino acid sequences of gag, pro, pol, and env were
161 searched against the seven genomes using the tblastn program (Altschul et al., 1997) with
162 a cut-off E-value of 1×10^{-30}. We identified 104 regions with high identity to *gag* (>800
163 bp), *prot* (>300 bp), *pol* (>800 bp), and *env* (>800 bp) were identified in the correct order
164 within a contiguous stretch of DNA <10 kbp, and further retrieved upstream and
165 downstream sequences of 2 kbp and assessed whether these sequences contained direct
166 sequence repeats using self-blastn. If the repeat sequences were longer than 300 bp, the
167 entire region (5′ LTR-*gag-pro-pol-env*-3′ LTR) was selected as a candidate SERV
168 sequence. In total, 91 SERV sequences longer than 8 kbp that included LTRs and did not
169 have consecutive ambiguous nucleotide sequences of >40 bp (typical assembly gap length
170 of unknown size), were used for further analyses.
171 In addition to full-length SERV sequences, we also searched nearly full-length protein-
172 coding genes in the draft genome sequences. Amino acid sequences encoding >90%
173 length of the original query were regarded as nearly full-length. The result is shown in
174 Table 1.

175 *Estimation of SERV copy numbers using short read fragments*

176 We retrieved publicly available short-read fragments of *M. fascicularis*, *M. mulatta*, *C.*
177 *sabaeus*, *R. roxellana, R. bieti* (sequence IDs are provided in Supplementary Table 1). In
178 order to estimate relative copy numbers of SERVs, we first mapped the short reads on

each draft genome sequence using the bwa mem algorithm with default parameters (Li, Ruan, & Durbin, 2008). The genome sequence coverage was estimated from the mode of coverage histogram, accounting for the overestimation attributable to repetitive sequences and copy number variations. We subsequently mapped the same short reads on the SERV sequences with LTRs identified in this study using the same method. In order to reduce the mapping bias at both ends of the SERV sequences, we counted the number of mapped reads only in the non-LTR regions. The coverage in the SERV sequences was estimated by measuring the number of mapped bases divided by the length of non-LTR regions of SERV (8181 bp). In both processes, we did not consider whether the mapped reads were properly paired. The genome coverages of short read sequences were estimated as 35, 34, 86, 75, and 52 in *M. fascicularis*, *M. mulatta*, *C. sabaeus*, *R. roxellana*, and *R. bieti*, respectively. The mean coverages in SERV sequences were 2898, 2250, 6711, 11 519, and 11 116 in *M. fascicularis*, *M. mulatta*, *C. sabaeus*, *R. roxellana*, and *R. bieti*, respectively. The copy number of SERVs was estimated as the coverage in the SERV sequences normalized by the coverage in the draft genome sequence.

Phylogenetic analyses

We conducted phylogenetic analyses of SRVs and SERVs using both nucleotide and amino acid sequences. Squirrel monkey type D retrovirus (SMRV) was used as the outgroup for both types of analyses (accession number: M23385.1). We also included six exogenous SRV sequences (accession numbers: M11841.1, AF12647.1, M16605, M12349, FJ971077.1, and AB611707.1) as well as the previously reported SERV sequence identified from the Vero cell line (SVL4d; accession number: LC310755) in all tree reconstructions (Sakuma et al., 2018). The consensus SERV sequences from Vero cells was not used for the phylogenetic analysis, because it was highly similar to the SVL4d sequence. The partial sequences of SRV-6 and PO-1-Lu (accession numbers: AF187057 and AY282754) were also retrieved from the Genbank database.

205 Prior to phylogenetic analyses of nucleotide sequences, we filtered out sequences that
206 fell outside the length range of 6500–7500 bp. In addition, one nucleotide sequence from
207 *R. roxellana*, RRo-29, had an 838-bp tandem duplication in the *gag* coding region and
208 was excluded from the tree reconstruction. Nucleotide sequences were aligned using the
209 MUSCLE algorithm implemented in MEGA6 with default parameters (Edgar, 2004;
210 Tamura, Stecher, Peterson, Filipski, & Kumar, 2013). The aligned nucleotide sequences
211 were subsequently used to reconstruct phylogenetic trees using the maximum likelihood
212 method implemented in MEGA6 (Saitou & Nei, 1987). For the maximum likelihood tree
213 reconstruction, we used the GTR+Γ+I and JTT+F+Γ models for nucleotide and amino
214 acid sequences, respectively. Substitution model selection was performed using MEGA.
215 Bootstrap resampling was performed 1000 times.
216 A phylogenetic tree was also reconstructed from nucleotide sequences using a Bayesian
217 clustering algorithm in BEAST (Drummond, Suchard, Xie, & Rambaut, 2012). We used
218 the GTR+Γ+I with five discrete gamma distribution categories. The fixed local clock
219 model and the Yule model with uniform birth rate for branching was assumed. The
220 exponential distribution were used as prior distributions for the gamma shape parameter.
221 The Markov chain Monte Carlo simulation was run for 20 million iterations with a burn-
222 in of 5 million. The effective sample size for each parameter was at least 800.
223 Phylogenetic analyses of amino acid sequences were performed as described above
224 using nearly full-length amino acid sequences of gag, pol, and env proteins. Because *pro*
225 partially overlaps *gag* and *pol* and premature stop codons in *pro* were found in many
226 SERVs, we did not use pro amino acid sequences for tree reconstruction. Phylogenetic
227 trees were constructed for each protein individually and for concatenated sequences
228 incorporating all three proteins. For the Bayesian analysis using amino acid sequences,
229 we used the JTT+I with five discrete gamma distribution categories. The fixed local clock
230 model and the Yule model with uniform birth rate for branching was assumed. The

exponential distribution was used as prior distributions for the gamma shape parameter. The Markov chain Monte Carlo simulation was run for 10 million iterations with a burn-in of 1 million. The effective sample size for each parameter was at least 1700.

We used the RDP4 software to detect potential recombination events among sequences (Martin & Rybicki, 2000). Outgroup sequences were excluded from recombination detection analyses. To maximize the length of the informative sequence alignment (in particular, to evaluate the recombination in the *gag* gene), seven SERV sequences with 5′-truncated *gag* sequences were also excluded. Default parameters were used, and results supported by >2 different recombination detection algorithms are presented here.

Estimation of selection pressure

PAML was used to estimate nonsynonymous and synonymous substitution rates (Yang, 2007) within the concatenated coding sequences used for amino acid tree reconstruction. Codon frequency was estimated using the F3X4 model. A uniform d_N/d_S ratio was assumed among sites, but variable d_N/d_S ratios were assumed among branches.

Results

SERV sequences in Old World monkey genomes

We first searched for nearly full-length SERV-like sequences containing both 5′ and 3′ LTR sequences in the draft genome sequences of seven Old World monkeys (*M. mulatta*, *M. fascicularis*, *P. anubis*, *C. sabaeus*, *N. larvatus*, *R roxellana*, and *R. bieti*; see Materials and Methods). Multiple candidate SERV sequences were identified in the genomes except for the *N. larvatus* genome (Table 1). The *N. larvatus* genome harbored several fragmented SERV sequences, but did not contain any full-length SERVs and intact protein-coding genes. All cer-SERVs contained premature stop codons in at least one protein-coding gene. In contrast, many col-SERVs had no evidence of premature stop codons. The *R. roxellana* and *R. bieti* genomes had 22 and 9 intact col-SERVs,

256 respectively, potentially encoding all four functional proteins. For example, col-SERV
257 RRo-1 in the *R. roxellana* genome (scaffold NW_010829022, coordinate 153774-
258 161954) has two 386-bp LTRs, *gag* encoding 655 residues, *prot* encoding 315 residues,
259 *pol* encoding 867 residues, and *env* encoding 572 residues. Furthermore, col-SERV RRo-
260 1 had target duplication sequences (ATTTTG) both upstream of the 5′ LTR and
261 downstream of the 3′ LTR. These genetic signatures demonstrate that SERV sequences in
262 Colobinae monkey genomes are ERVs and retain infectious potential. The complete list
263 of SERVs identified in our search is shown in Table 2. Average genetic distances (Kimura,
264 1980) within exogeneous SRVs, cer-SERVs, and col-SERVs were 0.321, 0.087, and 0.043,
265 respectively.
266 We investigated the presence of substitutions specific to SERV clusters that potentially
267 affect the infectious ability of SERVs, particularly focusing on the *env* gene, which plays
268 a key role in viral infection. We identified three cer-SERV specific amino acid changes in
269 functional domains of gp20 subunit (fusion transmembrane protein), which is cleaved
270 from the premature env protein. Two cer-SERV specific changes were in the heptad repeat
271 regions (position 430aa and 512aa in SRV-3, Figure 3) (Song & Hunter, 2003). The site
272 430 was a basic amino acid residue, histidine or arginine residues in all the exogeneous
273 SRVs and col-SERVs (except for one col-SERV, RRo-29), but asparagine or aspartic acid
274 residues in all the cer-SERVs. The site 512 was an acidic amino acid residue, glutamic
275 acid or aspartic acid residues in all the exogeneous SRVs and col-SERVs, but a
276 noncharged amino acid residue, alanine or threonine residues in all the cer-SERVs. The
277 third cer-SERV specific change was in the transmembrane region (position 535aa in SRV-
278 3). The site 535 was a cysteine residue in all the exogenous SRVs and col-SERVs;
279 however, it was substituted to threonine residue in all the cer-SERVs.
280 Because of the number of identified SERVs in Table 1 might be biased by using the
281 draft genome sequences obtained from different assembly processes, we estimated the

copy number of SERVs using the publicly available short-read sequences produced by next generation sequencers (see Methods; note that the estimation does not distinguish whether SERVs are full-length or fragmented). The copy numbers in *P. anubis* and *N. larvatus* were not estimated because the data were unavailable. The estimated copy numbers were 82.8, 66.2, 78.0, 209.5, and 213.8 in *M. fascicularis*, *M. mulatta*, *C. sabaeus*, *R. roxellana*, and *R. bieti*, respectively. The results showed that *R. roxellana* and *R. bieti* harbor a higher number of SERV sequences in the genomes than Cercopithecinae monkeys.

Phylogenetic relationships of genomic SERVs

Phylogenetic trees were reconstructed from the sequences of full-length SERVs identified in Old World monkey genomes as well as exogenous SRVs. We constructed phylogenetic trees using SERV nucleotide and amino acid sequences. Both trees showed consistent topologies; exogenous SRVs, cer-SERVs, and col-SERVs formed distinct clusters. Interestingly, exogenous SRVs appeared to be the sister of col-SERVs (Figure 4). Full phylogenetic trees constructed from nucleotide sequences of SRVs and SERVs are shown in Supplementary Figure 1–2. The bootstrap statistical support for the relationship between exogenous SRVs and col-SERVs was 83 and 47 in the nucleotide and amino acid tree using the maximum likelihood method, respectively.

Because the results of phylogenetic tree reconstruction might be biased by using a relatively distant outgroup sequence, SMRV, we performed two additional analyses excluding the SMRV sequence. Firstly, we constructed a phylogenetic tree using the nucleotide sequences excluding the SMRV sequence, and inferred the root position by minimizing the variance of Root-To-Tip branch lengths, implemented in TempEst software (Rambaut, Lam, Max Carvalho, & Pybus, 2016). The place of best-fitted root supported the sister relationship between exogenous SRVs and col-SERVs. Secondly, we applied a Bayesian clustering analysis implemented in the BEAST software and inferred

308 the tree root. The posterior root probability supporting the sister relationship between
309 exogenous SRVs and col-SERVs was 1.0000, showing the consistency and robustness of
310 the observation. We repeated the same analyses using the amino acid sequence dataset.
311 Although the best-fitted root in the amino acid tree estimated by TempEst showed that
312 the SRV sequences are the outgroup, probably owing to the fast protein evolution in
313 exogeneous SRVs, the root estimated by BEAST showed that the cer-SERV sequences
314 were the outgroup relative to the SRV and col-SERV sequences, with the root posterior
315 probability of 1.0000.
316 We also reconstructed the phylogenetic trees using the partial nucleotide sequences of
317 SRV-6 and PO-1-Lu. The result is shown in the Supplementary Figure 3. In the
318 phylogenetic tree, SRV-6 and PO-1-Lu were clustered with col-SERVs with high
319 statistical confidence. Furthermore, SRV-6 sequence was highly similar to col-SERVs of
320 *R. roxellana* (RRo-13 and RRo-29).

321 *Detection of recombination events among SERV sequences*

322 Although full-length genomic proviral SERV sequences showed consistent
323 phylogenetic relationships, there might exist heterogeneous evolutionary relationships
324 amongst SERV subsequences owing to recombination. We first separately constructed
325 phylogenetic trees for the LTR, *gag*, *pol*, and *env* regions (Supplementary Figure 4).
326 Although statistical supports were weak for all trees because of insufficient length, the
327 trees showed heterogenous topologies among regions. In LTR regions, cer-SERVs and
328 col-SERVs were clearly separated and they were further divided into three major
329 subclusters. The tree topologies of cer-SERVs were largely consistent between LTR and
330 non-LTR regions. In contrast, col-SERVs showed inconsistent pattern of phylogeny
331 between the LTR and non-LTR regions. For example, in the full-length SERV tree
332 excluding LTRs (Supplementary Figure 1), RRo-8, RRo-13, and RRo-41 formed a
333 distinct cluster from the other col-SERVs, but these three SERVs were placed in different

subclusters at LTR regions, implying that there have been historical recombination events within col-SERVs. To further test the idea, we applied the RDP4 software (Martin & Rybicki, 2000) and the program inferred multiple putative recombination events among the sequences in our dataset. Most recombination events had occurred between two SERVs in a single genome (Supplementary Table 2), particularly among col-SERVs. As expected, RRo-8, RRo-13, and RRo-41 were listed as candidates of recombinant sequences. We also observed potential recombination events between SRV-2, PAn-chrU-2, and RBi-6 in a ~600-bp region near the 3′ end of *gag* as well as between SRV-1, SRV-2, and PAn-chr4-1 in a ~330-bp region near the 3′ end of *pol*. Except for the ancient recombination between SRV-2 and cer-SERVs, we identified no recombination events between exogenous SRVs and SERVs.

Timing of SERV integration

Phylogenetic reconstructions showed that cer-SERVs primarily comprised species-specific clades, except for one clade harboring the recently diverged (~1–2 Mya) *M. mulatta* and *M. fascicularis* SERVs (Supplementary Figures 1 & 2) (Osada et al., 2008; Osada et al., 2010). This observation suggests that these SERVs became genomically integrated after the divergence of each genus. However, it is possible that our sequence search did not identify all SERVs because some SERVs may be highly fragmented after their integrations. To confirm whether these SERVs are indeed species or genus specific, we extracted their flanking sequences and carefully examined whether compatible integration sites were present in other genomes. Except for cases where flanking sequences were ambiguous in draft genome sequences or where flanking sequences were unidentifiable in the other genomes, we confirmed that the cer-SERVs were not integrated into the genomes of the other genera. A schematic illustration of the genus-specific integration sites is shown in Figure 5.

We also examined whether the integration sites of col-SERVs were shared between

Rhinopithecus and *Nasalis*. We found that some of the col-SERV integration sites identified in *R. roxellana* and *R. bieti* were absent in *N. larvatus*. However, the *Nasalis* genome contained solo LTRs at several loci; suggesting that the integrated SERVs were lost in the lineage of *Nasalis*. These shared LTRs were integrated into the genome before the divergence of these genera, but lost in the *Nasalis* genome.

We subsequently performed the molecular dating of integration timing using LTR sequences. Because the substitution rate at SERV LTRs has not been known, we applied the substitution rate range estimated by Johnson and Coffin in Hominidae ERVs ($2–5 \times 10^{-9}$ substitutions per site per year). The results are shown in Table 2. If we applied the average of the substitution rate by Johnson and Coffin (3.53×10^{-9} substitutions per site per year), averaged integration timings of cer-SERV and col-SERV were 3.42 Mya (SD: 2.20) and 6.16 Mya (SD, 3.41), respectively. We also found that, among the SERVs in *R. roxellana*, estimated integration timing of intact SERVs was significantly more recent than that of SERVs with frameshift substitutions ($P = 0.0002$; Mann–Whitney U test).

Selection pressures on SERV sequences

Using the tree reconstructed from amino acid sequences, we estimated d_N and d_S in three SERV protein-coding genes: *gag*, *pol*, and *env* at external branches. On average, exogenous SRVs showed the lowest d_N/d_S ratios in all three genes, reflecting functional constraints on their coding sequences to maintain viral fitness (Table 3). Except for *env*, col-SERV genes showed slightly higher d_N/d_S values than those cer-SERV genes. Although the d_N/d_S ratios in SERV coding sequences were higher than those in exogenous SRV coding sequences, they were considerably below 1 even for *env* genes, which is essential for intercellular viral infection, suggesting that substitutions in *env* had been negatively selected in some of the SERVs (Belshaw et al., 2004).

Discussion

In this study, we investigated the genomic distribution and evolutionary history of exogenous SRVs and SERVs in Old World monkeys using seven draft genome sequences. Studies on col-SERVs lag far behind those on cer-SERVs (particularly macaque SERVs), although a few studies have identified SERVs in Colobinae monkeys (Sommerfelt et al., 2003; Todaro et al., 1978). One reason for the apparent lack of interest in col-SERVs is that an early hybridization-based screening failed to identify SERVs in *Colobus guereza* (van der Kuyl et al., 1997), which suggested that SERVs were integrated after the divergence between Cercopithecinae and Colobinae monkeys. Another reason is that the genomic resources of Cercopithecinae monkeys including draft genome sequences are better established than those of Colobinae monkeys, partly because Cercopithecinae monkeys have been preferentially used for biomedical research. Our results, in contrast, demonstrate that Colobinae monkey genomes also harbor numerous SERVs, some of which might be active. In addition, we found that col-SERVs were more closely related to exogenous SRVs than to any cer-SERV, suggesting that pathogenic exogenous SRVs are derived from col-SERVs. Alternatively, pathogenic exogenous SRVs may descend from to date unidentified or extinct exogenous SRVs. We identified three radical amino acid substitutions specific to cer-SERVs in the gp20 protein, which occurred in the heptad repeat regions and transmembrane region. Interestingly, all col-SERVs harbored the same amino acid residues as the exogenous SRVs. The gp20 proteins form a trimer and play an important role for viral infection. Charged amino acids in the heptad repeat regions are essential to a stable trimer structure during a viral entry process (Aydin, Cook, & Lee, 2014). These three changes might explain the different behavior of cer-SERVs and col-SERVs in their genomes. Altogether, our findings highlight the importance of Colobinae monkeys in the study of

409 SRV, particularly given the relative neglect pertaining to this aspect in biomedical
410 research.
411 As shown in Table 1, *R. roxellana* genome had the highest number of SERVs among
412 the seven genomes analyzed in this study. One important limitation of the draft genome
413 scan approach should be kept in mind; draft genome sequences, by definition, are
414 incomplete and have gaps. Because different draft genome sequences were generated
415 using different sequencing platforms and assembled in different analytical pipelines, we
416 should be careful about the evaluation of the number of SERV sequences estimated by
417 the genome sequence scan. However, our additional analysis using short-read fragments
418 confirmed that the *Rhinopithecus* genomes contain more SERV sequences than the
419 Cercopithecinae genomes, supporting the result of genome sequence scan. Although the
420 estimated SERV copy numbers were fluctuated with different mapping strategies, the
421 pattern of higher SERV copy numbers in *Rhinopithecus* than the Cercopithecinae
422 genomes was robust to the analytical methods (data not shown).
423 The assembly problem of draft genome sequences is particularly of concern in case of
424 recently integrated retroviruses that are still undergoing the process of fixation; such
425 young SERVs are often heterozygous in the host genome and thus tend to be ignored in
426 genome sequence assemblies. Indeed, one previous study identified many heterozygous
427 SERVs that were unrepresented in the draft genome sequence of *C. sabaeus* (Sakuma et
428 al., 2018). The young, heterozygous SERVs were found to cluster with other cer-SERVs
429 (SVL4d in Supplementary Figure 1 and 2) and not with exogenous SRVs. Therefore, we
430 cannot eliminate the possibility that young SERVs directly derived from exogenous SRVs
431 are present in Old World monkey genomes in a heterozygous form. A recent report
432 suggested the existence of a new class of SERVs more closely related to SRV-2 in the
433 pigtailed macaque *M. nemestrina* (Grant et al., 2017). The question of whether there are
434 SERVs directly derived from exogenous SRVs in other Old World monkey genomes will

need to be clarified in future studies.

However, the fact that no cer-SERVs were shared among genera revealed from genome sequence analysis strongly suggests that cer-SERVs originated relatively recently, at least within the last 8 million years for the SERVs in macaques and baboons (Perelman et al., 2011). On the other hand, the fact that some col-SERV integration events were species specific, some were genus specific, and some were shared between *Nasalis* and *Rhinopithecus* is suggestive of continuous introduction of SERVs in Colobinae genomes. The estimated integration timings by molecular dating also suggest the above scenarios.

In summary, our comprehensive genome search revealed many novel SERV sequences in Old World monkey genomes, several of which potentially encode functional proteins that are essential for the production of infectious virus. Analysis of integration sites showed that most SERV integration events were not shared between Cercopithecinae and Colobinae monkeys, indicating that cer-SERVs and col-SERVs have independent evolutionary origins. In addition, the distinct phylogenetic clustering of exogenous SRVs, cer-SERVs, and col-SERVs suggests there was no transmission of SERV elements between the two Cercopithecidae subfamilies. These findings shed novel light on the evolutionary history of SERVs and may help understand the process of endogenization of SRVs.

Acknowledgments

We are grateful to We are grateful to Drs. Makoto Kuroda, Chisato Sakuma, Kyoko Saito, and Toshiyuki Yamaji (National Institute of Infectious Diseases), and Dr. Fumio Kasai (National Institutes of Biomedical Innovation, Health and Nutrition) for providing the SERV sequences identified from the Vero cell line. We also thank two anonymous reviewers and Dr. Sébastien Calvignac-Spencer for helpful comments. The study is partly supported by the Japan Society for the Promotion of Science KAKENHI Grant number 17H04003 (to K.H.)

Ethical statement

This study utilized only published genome sequences of non-human primates. Ethical policies for handling living animals were not applied.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389-3402. doi:10.1093/nar/25.17.3389

Aydin, H., Cook, J. D., & Lee, J. E. (2014). Crystal Structures of Beta- and Gammaretrovirus Fusion Proteins Reveal a Role for Electrostatic Stapling in Viral Entry. *Journal of Virology, 88*(1), 143-153. doi:10.1128/jvi.02023-13

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Pačes, J., Burt, A., & Tristem, M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences of the United States of America, 101*(14), 4894-4899. doi:10.1073/pnas.0307800101

Benveniste, R. E., & Todaro, G. J. (1977). Evolution of primate oncornaviruses: An endogenous virus from langurs (Presbytis spp.) with related virogene sequences in other Old World monkeys. *Proceedings of the National Academy of Sciences of the United States of America, 74*(10), 4557-4561.

Chopra, H. C., & Mason, M. M. (1970). A New Virus in a Spontaneous Mammary Tumor of a Rhesus Monkey. *Cancer Research, 30*(8), 2081-2086.

Diehl, W. E., Patel, N., Halm, K., & Johnson, W. E. (2016). Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. *eLife, 5*, e12704. doi:10.7554/eLife.12704

Drummond, A., Suchard, M., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics

with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution, 29*(8), 1969 - 1973.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792-1797. doi:10.1093/nar/gkh340

Grant, R., Keele, B., Kuller, L., Watanabe, R., Perret, A., & Smedley, J. (2017). Identification of novel simian endogenous retroviruses that are indistinguishable from simian retrovirus (SRV) on current SRV diagnostic assays. *Journal of Medical Primatology, 46*(4), 158-161. doi:10.1111/jmp.12297

Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Graves, T., Fulton, R. S., . . . Page, D. C. (2012). Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature, 483*(7387), 82-86. doi:10.1038/nature10843

Johnson, W. E. (2015). Endogenous Retroviruses in the Genomics Era. *Annual Review of Virology, 2*(1), 135-159. doi:10.1146/annurev-virology-100114-054945

Johnson, W. E., & Coffin, J. M. (1999). Constructing primate phylogenies from ancient retrovirus sequences. *Proceedings of the National Academy of Sciences, 96*(18), 10254-10260. doi:10.1073/pnas.96.18.10254

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution, 16*(2), 111-120.

Lerche, N. W., & Osborn, K. G. (2003). Simian Retrovirus Infections: Potential Confounding Variables in Primate Toxicology Studies. *Toxicologic Pathology, 31*(1_suppl), 103-110. doi:10.1080/01926230390174977

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research, 18*(11), 1851-1858. doi:10.1101/gr.078212.108

Li, W. H., Wu, C. I., & Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood

512 of nucleotide and codon changes. *Molecular Biology and Evolution, 2*(2), 150-174.
513 Ma, H., Ma, Y., Ma, W., Williams, D. K., Galvin, T. A., & Khan, A. S. (2011). Chemical
514 Induction of Endogenous Retrovirus Particles from the Vero Cell Line of African Green
515 Monkeys. *Journal of Virology, 85*(13), 6579-6588. doi:10.1128/jvi.00147-11
516 Magiorkinis, G., Blanco-Melo, D., & Belshaw, R. (2015). The decline of human
517 endogenous retroviruses: extinction and survival. *Retrovirology, 12*(1), 8.
518 doi:10.1186/s12977-015-0136-x
519 Martin, D., & Rybicki, E. (2000). RDP: detection of recombination amongst aligned
520 sequences. *Bioinformatics, 16*(6), 562-563. doi:10.1093/bioinformatics/16.6.562
521 Montiel, N. A. (2010). REVIEW ARTICLE: An updated review of simian
522 betaretrovirus (SRV) in macaque hosts. *Journal of Medical Primatology, 39*(5), 303-314.
523 doi:10.1111/j.1600-0684.2010.00412.x
524 Nandi, J. S., Bhavalkar-Potdar, V., Tikute, S., & Raut, C. G. (2000). A Novel Type D
525 Simian Retrovirus Naturally Infecting the Indian Hanuman Langur (Semnopithecus
526 entellus). *Virology, 277*(1), 6-13. doi:10.1006/viro.2000.0567
527 Nandi, J. S., Van Dooren, S., Chhangani, A. K., & Mohnot, S. M. (2006). New Simian
528 β Retroviruses from Rhesus Monkeys (*Macaca Mulatta*) and Langurs (*Semnopithecus*
529 *Entellus*) from Rajasthan, India. *Virus Genes, 33*(1), 107-116. doi:10.1007/s11262-005-
530 0032-x
531 Omar, N., Wong, Y. S., Li, X., Chong, Y. L., Abdullah, M. T., & Lee, N. K. (2017).
532 Enhancer Prediction in Proboscis Monkey Genome: A Comparative Study. *Journal of*
533 *Telecommunication, Electronic and Computer Engineering (JTEC), 9*(2-9), 175-179.
534 Onions, D., Côté, C., Love, B., Toms, B., Koduri, S., Armstrong, A., . . . Kolman, J.
535 (2011). Ensuring the safety of vaccine cell substrates by massively parallel sequencing of
536 the transcriptome. *Vaccine, 29*(41), 7117-7121. doi: 10.1016/j.vaccine.2011.05.071
537 Osada, N., Hashimoto, K., Kameoka, Y., Hirata, M., Tanuma, R., Uno, Y., . . . Takahashi,

I. (2008). Large-scale analysis of *Macaca fascicularis* transcripts and inference of genetic divergence between *M. fascicularis* and *M. mulatta*. *BMC Genomics, 9*, 90. doi:10.1186/1471-2164-9-90

Osada, N., Kohara, A., Yamaji, T., Hirayama, N., Kasai, F., Sekizuka, T., . . . Hanada, K. (2014). The Genome Landscape of the African Green Monkey Kidney-Derived Vero Cell Line. *DNA Research, 21*(6), 673-683. doi:10.1093/dnares/dsu029

Osada, N., Uno, Y., Mineta, K., Kameoka, Y., Takahashi, I., & Terao, K. (2010). Ancient genome-wide admixture extends beyond the current hybrid zone between *Macaca fascicularis* and *M. mulatta*. *Molecular Ecology, 19*(14), 2884-2895.doi:10.1111/j.1365-294X.2010.04687.x

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., . . . Pecon-Slattery, J. (2011). A Molecular Phylogeny of Living Primates. *PLoS Genetics, 7*(3), e1001342. doi:10.1371/journal.pgen.1001342

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry, 25*(13), 1605-1612.doi:10.1002/jcc.20084

Power, M., Marx, P., Bryant, M., Gardner, M., Barr, P., & Luciw, P. (1986). Nucleotide sequence of SRV-1, a type D simian acquired immune deficiency syndrome retrovirus. *Science, 231*(4745), 1567-1572. doi:10.1126/science.3006247

Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution, 2*(1), vew007-vew007. doi:10.1093/ve/vew007

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*, 406-425.

Sakuma, C., Sekizuka, T., Kuroda, M., Kasai, F., Saito, K., Ikeda, M., . . . Hanada, K. (2018). Novel endogenous simian retroviral integrations in Vero cells: implications for quality control of a human vaccine cell substrate. *Scientific Reports, 8*(1), 644.

566 doi:10.1038/s41598-017-18934-2
567 Sommerfelt, M. A., Harkestad, N., & Hunter, E. (2003). The endogenous langur type
568 D retrovirus PO-1-Lu and its exogenous counterparts in macaque and langur monkeys.
569 *Virology, 315*(2), 275-282. doi:10.1016/S0042-6822(03)00548-8
570 Sonigo, P., Barker, C., Hunter, E., & Wain-Hobson, S. (1986). Nucleotide sequence of
571 Mason-Pfizer monkey virus: An immunosuppressive D-type retrovirus. *Cell, 45*(3), 375-
572 385. doi:10.1016/0092-8674(86)90323-5
573 Song, C., & Hunter, E. (2003). Variable Sensitivity to Substitutions in the N-Terminal
574 Heptad Repeat of Mason-Pfizer Monkey Virus Transmembrane Protein. *Journal of*
575 *Virology, 77*(14), 7779-7785. doi:10.1128/jvi.77.14.7779-7785.2003
576 Stoye, J. P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary
577 saga. *Nature Reviews Microbiology, 10*, 395. doi:10.1038/nrmicro2783
578 Takano, J.-I., Leon, A., Kato, M., Abe, Y., & Fujimoto, K. (2013). Isolation and DNA
579 characterization of a simian retrovirus 5 from a Japanese monkey (Macaca fuscata).
580 *Journal of General Virology, 94*(5), 955-959. doi:doi:10.1099/vir.0.047621-0
581 Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6:
582 Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution,*
583 *30*(12), 2725-2729. doi:10.1093/molbev/mst197
584 Thayer, R. M., Power, M. D., Bryant, M. L., Gardner, M. B., Barr, P. J., & Luciw, P. A.
585 (1987). Sequence relationships of type D retroviruses which cause simian acquired
586 immunodeficiency syndrome. *Virology, 157*(2), 317-329. doi:10.1016/0042-
587 6822(87)90274-1
588 Todaro, G. J., Benveniste, R. E., Sherr, C. J., Schlom, J., Schidlovsky, G., & Stephenson,
589 J. R. (1978). Isolation and characterization of a new type D retrovirus from the Asian
590 primate, *Presbytis obscurus* (spectacled langur). *Virology, 84*(1), 189-194.
591 doi:10.1016/0042-6822(78)90231-3

van der Kuyl, A. C., Mang, R., Dekker, J. T., & Goudsmit, J. (1997). Complete nucleotide sequence of simian endogenous type D retrovirus with intact genome organization: evidence for ancestry to simian retrovirus and baboon endogenous virus. *Journal of Virology, 71*(5), 3666-3676.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution, 24*(8), 1586-1591. doi:10.1093/molbev/msm088

Yoshikawa, R., Okamoto, M., Sakaguchi, S., Nakagawa, S., Miura, T., Hirai, H., & Miyazawa, T. (2015). Simian Retrovirus 4 Induces Lethal Acute Thrombocytopenia in Japanese Macaques. *Journal of Virology, 89*(7), 3965-3975. doi:10.1128/jvi.03611-14

Zao, C.-L., Armstrong, K., Tomanek, L., Cooke, A., Berger, R., Estep, J. S., . . . Lerche, N. W. (2010). The complete genome and genetic characteristics of SRV-4 isolated from cynomolgus monkeys (*Macaca fascicularis*). *Virology, 405*(2), 390-396. doi:10.1016/j.virol.2010.06.028

Zao, C.-L., Tomanek, L., Cooke, A., Berger, R., Yang, L., Xie, C., . . . Rong, R. (2016). A novel simian retrovirus subtype discovered in cynomolgus monkeys (*Macaca fascicularis*). *Journal of General Virology, 97*(11), 3017-3023. doi:doi:10.1099/jgv.0.000601

Zhou, X., Wang, B., Pan, Q., Zhang, J., Kumar, S., Sun, X., . . . Li, M. (2014). Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. *Nature Genetics, 46*(12), 1303-1310. doi:10.1038/ng.3137

613 **Table 1. Summary of SERVs identified in the seven Old World monkey genomes**

Species	# of SERVs	# of intact SERVs*	# of *gag*†	# of *pol*†	# of *env*†
Macaca fascicularis (MFa)	12	0	20	40	0
Macaca mulatta (MMu)	6	0	9	16	3
Papio Anubis (PAn)	8	0	17	16	17
Chlorocebus sabaeus (CSa)	3	0	2	3	2
Nasalis larvatus (NLa)	0	0	0	0	0
Rhinopithecus roxellana (RRo)	41	22	50	12	51
Rhinopithecus bieti (RBi)	10	8	10	3	12

614 *SERVs without premature stop codons in *gag*, *pro*, *pol*, and *env*

615 †number of *gag*, *pol*, and *env* encoding nearly full-length proteins

616

617 **Table 2 List of SERVs in the six Old world monkey genomes**

Species/Assembly	ID	chromosome/scaffold	strand	start	end	gag*	pol*	env*	T_{int} (Mya) §
Macaca fascicularis/	MFa-chr6-1	6	−	78590774	78599010	+	−	+	0.27–0.68
GCA_000364345.1	MFa-chrU-1	NW_005093489.1	−	61438	69843	+	+	+	4.27–10.68
	MFa-chr6-2	6	+	40765258	40773563	+	−	+	0.49–1.22
	MFa-chrU-2	NW_005093479.1	−	703289	711617	+	−	−	2.33–5.83
	MFa-chr1	1	−	98380607	98388838	+	+	+	2.9–7.24
	MFa-chr3-1	3	−	71798389	71806704	−	−	+	2.33–5.83
	MFa-chr2	2	−	84058787	84067087	+	+	+	1.89–4.73
	MFa-chrU-3	NW_005093490.1	+	15894	24305	*	+	−	3.61–9.02
	MFa-chr5	5	+	3618605	3626912	+	+	+	2.08–5.21
	MFa-chr3-2	3	+	64085155	64093474	−	−	+	1.68–4.2
	MFa-chr4	4	+	55800507	55808825	+	−	+	1.84–4.6
	MFa-chr11	11	+	105870233	105878551	+	+	+	1.15–2.87
Macaca mulatta/	MMu-chr1	1	−	414923	423250	+	+	+	2.08–5.2
GCF_000002255.2	MMu-chr4	4	−	32624247	32632554	−	−	+	2.59–6.49
	MMu-chrY-1	Y	−	3535329	3543643	*	+	+	2.33–5.84
	MMu-chrY-2	Y	−	1201937	1210230	+	+	+	5.64–14.09
	MMu-chrY-3	Y	−	869	9396	+	+	+	1.6–3.99
	MMu-chrY-4	Y	+	3868458	3876757	*	+	+	2.18–5.45
Papio Anubis/	PAn-chr13	13	−	43983316	43991550	−	+	+	1.31–3.27

GCA_000264685.2	PAn-chr2	2	−	118845707	118854112	−	−	−	0.84–2.09
Papio anubis	PAn-chr4-1	4	−	102091220	102099421	+	−	−	3.26–8.14
	PAn-chrU-1	JH685002	−	38105	46511	+	−	−	4.29–10.72
	PAn-chr3	3	+	78436276	78444599	+	+	−	3.01–7.53
	PAn-chr4-2	4	+	67256607	67265014	−	−	−	1.69–4.23
	PAn-chr6	6	+	30107451	30115852	+	−	+	1.69–4.24
	PAn-chrU-2	JH687348	+	4424	12780	+	+	+	7.61–19.03
Chlorocebus sabaeus/	CSa-chr17-1	17	−	38970823	38979182	−	+	+	0.66–1.66
GCA_000409795.1	CSa-chr17-2	17	+	38737629	38746034	+	−	−	2.96–7.39
	CSa-chr8	8	+	67814295	67822700	−	−	−	1.68–4.2
Rhinopithecus roxellana/	RRo-1	NW_010829022.1	+	153774	161954	−	−	−	1.92–4.79
GCF_000769185.1	RRo-2	NW_010828124.1	+	606447	614866	−	−	−	0.58–1.44
	RRo-3	NW_010789988.1	+	59268	67742	−	−	−	2.3–5.76
	RRo-4	NW_010830674.1	+	391701	399878	−	−	−	4.03–10.08
	RRo-5	NW_010829297.1	+	426461	434641	−	−	−	2.11–5.27
	RRo-6	NW_010828396.1	+	776000	784467	−	−	−	0.76–1.9
	RRo-7	NW_010827817.1	+	113822	122289	−	−	−	0.19–0.47
	RRo-8	NW_010827474.1	+	381905	390236	−	−	−	1.31–3.29
	RRo-9	NW_010824106.1	+	94380	102828	−	−	−	0
	RRo-10	NW_010819874.1	+	261592	270025	−	−	−	4.45–11.14
	RRo-11	NW_010798467.1	+	622698	631110	−	−	−	3.65–9.11
	RRo-12	NW_010827005.1	+	1826448	1834816	−	−	−	7.27–18.17

Rhinopithecus roxellana/	RRo-13	NW_010828720.1	+	89877	98188	−	−	−	3.32–8.29
	RRo-14	NW_010803703.1	−	332847	341316	−	−	−	1.53–3.83
	RRo-15	NW_010803213.1	−	1831928	1840378	−	−	−	0.79–1.97
	RRo-16	NW_010788543.1	−	933455	941904	−	−	−	0
	RRo-17	NW_010830488.1	−	961666	970094	−	−	−	4.49–11.22
	RRo-18	NW_010811370.1	−	401270	409739	−	−	−	4.06–10.15
	RRo-19	NW_010829349.1	−	262952	271379	−	−	−	3.45–8.61
	RRo-20	NW_010793464.1	−	154772	163178	−	−	−	5.62–14.04
	RRo-21	NW_010796474.1	−	1212835	1221102	−	−	−	4.09–10.21
	RRo-22	NW_010809255.1	−	77327	85714	−	−	+	2.72–6.79
	RRo-23	NW_010795985.1	−	17764	25955	−	+	−	2.65–6.62
	RRo-24	NW_010828039.1	−	238671	247011	+	−	−	6.45–16.13
	RRo-25	NW_010830337.1	−	1654244	1662724	−	−	−	8.65–21.62
	RRo-26	NW_010826131.1	−	34151	42578	−	+	−	6.88–17.2
	RRo-27	NW_010796444.1	−	496731	504911	+	+	−	6.28–15.7
	RRo-28	NW_010797013.1	−	266627	275161	+	+	+	6.9–17.25
	RRo-29	NW_010828958.1	−	100994	109183	+	−	−	5.21–13.03
	RRo-30	NW_010814135.1	−	469683	478030	+	−	−	4.88–12.21
	RRo-31	NW_010792178.1	−	849888	858216	−	−	+	6.59–16.48
	RRo-32	NW_010830147.1	+	543693	552069	+	+	+	6.39–15.98
	RRo-33	NW_010798505.1	+	94929	103087	+	+	+	7.36–18.4
	RRo-34	NW_010830301.1	+	1099502	1107875	−	+	−	5.06–12.64

	RRo-35	NW_010818555.1	+	88085	96314	+	+	+	6.31–15.76
Rhinopithecus roxellana	RRo-36	NW_010830147.1	+	440357	448565	+	+	−	7.42–18.56
	RRo-37	NW_010830701.1	+	1363491	1371877	+	−	+	5.57–13.92
	RRo-38	NW_010829302.1	+	278775	287103	+	−	−	3.45–8.63
	RRo-39	NW_010829941.1	+	845977	854390	+	−	−	4.21–10.53
	RRo-40	NW_010792819.1	+	83532	91768	−	+	−	8.05–20.13
	RRo-41	NW_010830531.1	+	461335	469529	−	+	*	7.7–19.26
Rhinopithecus bieti/	RBi-1	MCGX01006469.1	−	373228	381704	−	−	−	0.19–0.48
GCA_001698545.1	RBi-2	MCGX01000569.1	−	740591	748873	−	−	−	3.78–9.45
	RBi-3	MCGX01000375.1	−	2085838	2094269	−	−	−	5.73–14.33
	RBi-4	MCGX01002382.1	+	339495	347874	−	−	−	6.3–15.76
	RBi-5	MCGX01004395.1	+	582626	591042	−	−	−	6.81–17.03
	RBi-6	MCGX01011615.1	+	4123447	4131793	−	−	−	7.02–17.54
	RBi-7	MCGX01017424.1	+	57172	65643	−	−	−	2.51–6.28
	RBi-8	MCGX01019892.1	+	806528	814717	−	−	−	3.48–8.7
	RBi-9	MCGX01007578.1	−	759829	768438	+	−	−	5.77–14.42
	RBi-10	MCGX01000863.1	−	22655	31048	+	−	−	5.44–13.6

*presence/absence of premature stop codons in each coding region: +: present, −:absent, *: no premature stop codon but amino acid sequence is truncated by deletion

§estimated integration time, inferred from the divergence of LTR sequences. Substitution rate of $2–5 \times 10^{-9}$ per site per year was assumed.

618

Table 3. d_N/d_S ratios in *gag*, *pol* and *env* coding sequences, averaged within each cluster

Cluster	*gag*	*pol*	*env*
SRV	0.065	0.063	0.220
cer-SERV	0.169	0.106	0.663
col-SERV	0.241	0.124	0.314

*the ratio shown is of d_N averaged across branches to d_S averaged across external branches

Figure Legends

Figure 1

Genomic structure of complete simian endogenous retrovirus (SERV) sequences. The open and gray-shaded boxes represent LTR and target duplication sites, respectively.

Figure 2

Phylogeny of the seven Old World monkeys estimated from the results of previous studies. Divergence times were obtained from a study by Perelman et al. (Perelman et al., 2011), except for the divergence time between *M. fascicularis* and *M. mulatta* (N. Osada et al., 2008) and between *R. roxellana* and *R. bieti* (Zhou et al., 2014).

Figure 3

Cer-SERV specific substitutions in the heptad repeat regions of gp20 protein. A trimeric structure of SRV-3 gp20 is shown (PDB ID:4JF3). One subunit is colored in blue and the other are colored in grey. The heptad repeat regions form two parallel helices. cer-SERV specific changes are shown in red with the structure of side chains. The figure was generated using the UCSF Chimera package (Pettersen et al., 2004).

Figure 4

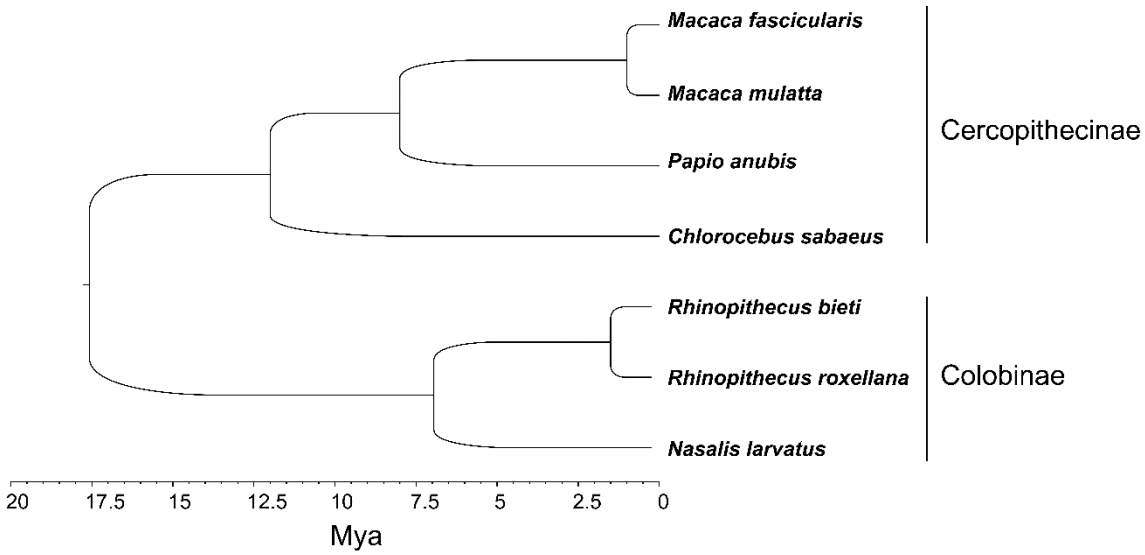
Phylogenetic tree constructed from full-length nucleotide sequences of simian retroviruses (SRVs) and simian endogenous retroviruses SERVs using maximum likelihood methods. Bootstrap values (%) are shown along the branches. The full phylogenetic tree showing labels of all SERV sequences is presented in Supplementary Figure 1.

Figure 5

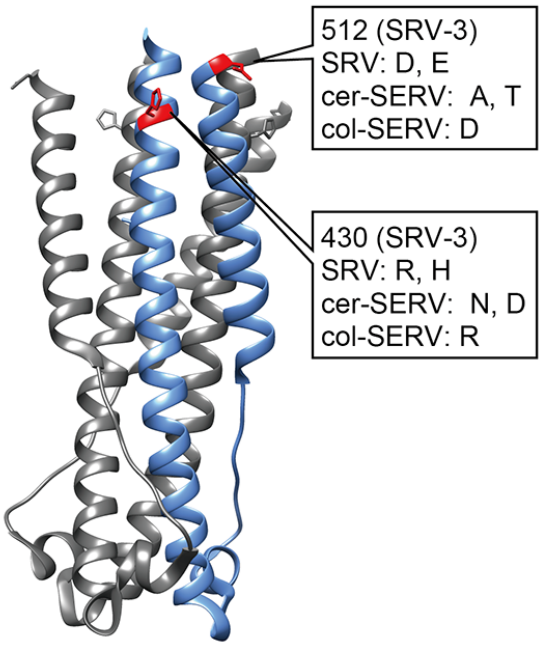
Simian endogenous retrovirus (SERV) integration site at MFa-chr5. The insertion

boundary in MFa and the orthologous regions in PAn and CSa are shown. Target site duplication sequences are shown in boxes. Note that a C→A mutation occurred at the first site in the target site duplication sequence after the divergence between macaques and baboons. The absence of integration in PAn and CSa indicates that SERV integration occurred after the divergence between macaques and baboons.

653



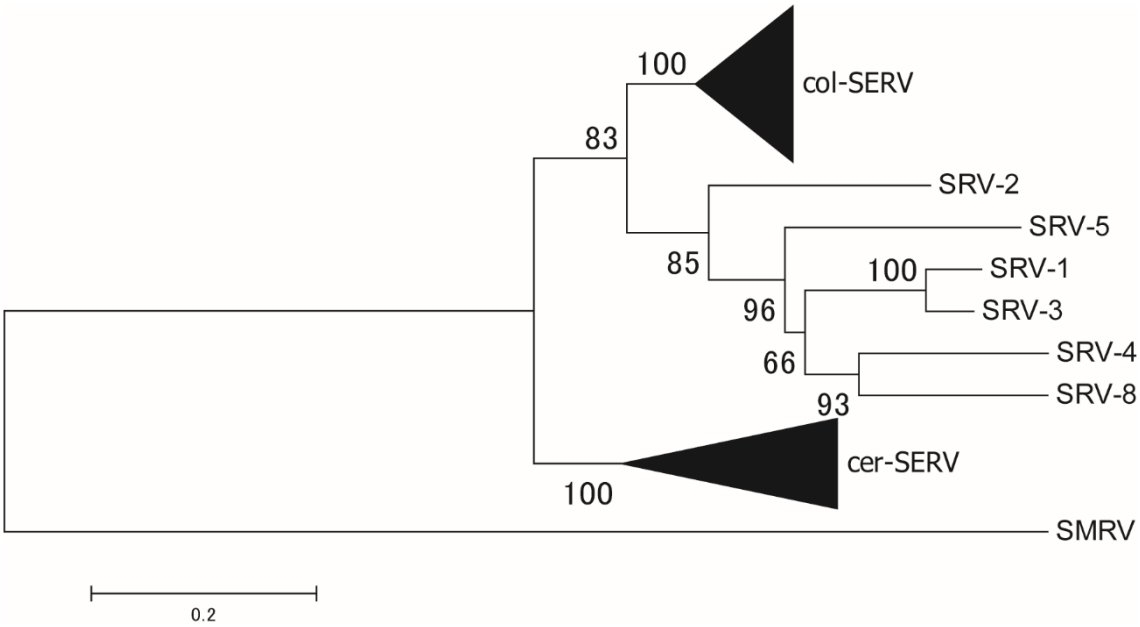
654



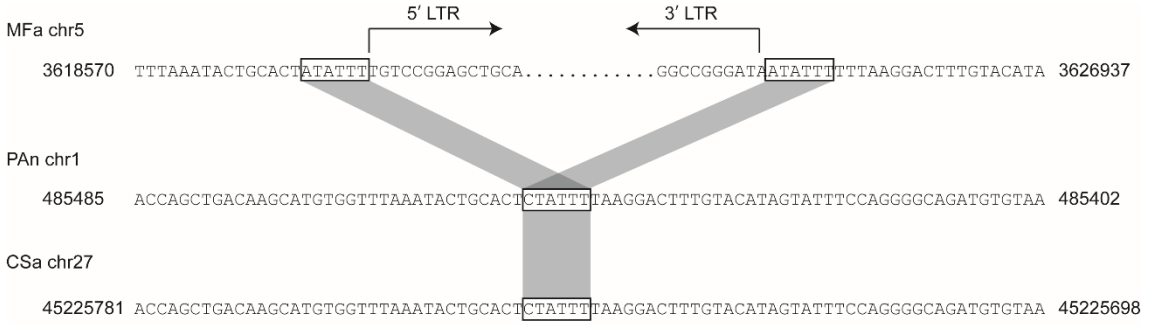
665

666

667


100
col-SERV
83
SRV-2
SRV-5
85
100
SRV-1
96
SRV-3
66
SRV-4
93
SRV-8
cer-SERV
100
SMRV
0.2

668


5′ LTR
3′ LTR
MFa chr5
3618570 TTTAAATACTGCACTATATTTTGTCCGGAGCTGCA.....GGCCGGGATAATATTTTTTAAGGACTTTGTACATA 3626937
PAn chr1
485485 ACCAGCTGACAAGCATGTGGTTTAAATACTGCACTCTATTTTAAGGACTTTGTACATAGTATTTCCAGGGGCAGATGTGTAA 485402
CSa chr27
45225781 ACCAGCTGACAAGCATGTGGTTTAAATACTGCACTCTATTTTAAGGACTTTGTACATAGTATTTCCAGGGGCAGATGTGTAA 45225698