Evolutionarily Young African Rhinoceros Gammaretroviruses

Article *in* Journal of Virology · April 2023

DOI: 10.1128/jvi.01932-22

Kyriakos Tsangaras
Cyprus Institute of Neurology and Genetics
129 PUBLICATIONS **920** CITATIONS
SEE PROFILE

Jens Mayer
Universität des Saarlandes
123 PUBLICATIONS **3,475** CITATIONS
SEE PROFILE

Anisha Dayaram
Charité Universitätsmedizin Berlin
154 PUBLICATIONS **1,123** CITATIONS
SEE PROFILE

Michelle Campbell-Ward
New South Wales Department of Primary Industries
19 PUBLICATIONS **48** CITATIONS
SEE PROFILE

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

GENE DELIVERY

Evolutionarily Young African Rhinoceros Gammaretroviruses

Kyriakos Tsangaras,[a] Jens Mayer,[b] Omar Mirza,[c] Anisha Dayaram,[d] Damien P. Higgins,[e] Benn Bryant,[f] Michelle Campbell-Ward,[f] Cheryl Sangster,[g] Andrea Casteriano,[e] Dirk Höper,[h] Martin Beer,[h] Alex D. Greenwood[c,i]

[a]Department of Life and Health Sciences, University of Nicosia, Nicosia, Cyprus
[b]Institute of Human Genetics, Medical Faculty, University of Saarland, Homburg, Germany
[c]Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research (IZW), Berlin, Germany
[d]Institute of Neurophysiology, Charité—Universitätsmedizin Berlin, Berlin, Germany
[e]Sydney School of Veterinary Science, The University of Sydney, Sydney, New South Wales, Australia
[f]Taronga Western Plains Zoo, Dubbo, New South Wales, Australia
[g]Taronga Conservation Society Australia, Mosman, New South Wales, Australia
[h]Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Greifswald, Germany
[i]School of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany

ABSTRACT High-throughput sequences were generated from DNA and cDNA from four Southern white rhinoceros (*Ceratotherium simum simum*) located in the Taronga Western Plain Zoo in Australia. Virome analysis identified reads that were similar to *Mus caroli* endogenous gammaretrovirus (McERV). Previous analysis of perissodactyl genomes did not recover gammaretroviruses. Our analysis, including the screening of the updated white rhinoceros (*Ceratotherium simum*) and black rhinoceros (*Diceros bicornis*) draft genomes identified high-copy orthologous gammaretroviral ERVs. Screening of Asian rhinoceros, extinct rhinoceros, domestic horse, and tapir genomes did not identify related gammaretroviral sequences in these species. The newly identified proviral sequences were designated SimumERV and DicerosERV for the white and black rhinoceros retroviruses, respectively. Two long terminal repeat (LTR) variants (LTR-A and LTR-B) were identified in the black rhinoceros, with different copy numbers associated with each (n = 101 and 373, respectively). Only the LTR-A lineage (n = 467) was found in the white rhinoceros. The African and Asian rhinoceros lineages diverged approximately 16 million years ago. Divergence age estimation of the identified proviruses suggests that the exogenous retroviral ancestor of the African rhinoceros ERVs colonized their genomes within the last 8 million years, a result consistent with the absence of these gammaretroviruses from Asian rhinoceros and other perissodactyls. The black rhinoceros germ line was colonized by two lineages of closely related retroviruses and white rhinoceros by one. Phylogenetic analysis indicates a close evolutionary relationship with ERVs of rodents including sympatric African rats, suggesting a possible African origin of the identified rhinoceros gammaretroviruses.

IMPORTANCE Rhinoceros genomes were thought to be devoid of gammaretroviruses, as has been determined for other perissodactyls (horses, tapirs, and rhinoceros). While this may be true of most rhinoceros, the African white and black rhinoceros genomes have been colonized by evolutionarily young gammaretroviruses (SimumERV and DicerosERV for the white and black rhinoceros, respectively). These high-copy endogenous retroviruses (ERVs) may have expanded in multiple waves. The closest relative of SimumERV and DicerosERV is found in rodents, including African endemic species. Restriction of the ERVs to African rhinoceros suggests an African origin for the rhinoceros gammaretroviruses.

KEYWORDS endogenous retrovirus, gammaretrovirus, rhinoceros, perissodactyl, evolution

Editor Frank Kirchhoff, Ulm University Medical Center

Address correspondence to Alex D. Greenwood, greenwood@izw-berlin.de.

The authors declare no conflict of interest.

[This article was published on 6 April 2023 with incorrect information in three author affiliations. The affiliations were updated in the current version, posted on 27 April 2023.]

Received 16 December 2022
Accepted 17 March 2023
Published 6 April 2023

Retroviruses are enveloped, single-stranded RNA (ssRNA) viruses that reverse transcribe their genome into double-stranded DNA, which subsequently integrates into the host cell DNA (1). The resulting provirus encodes all the necessary sequences to direct the production of progeny virions that can infect a new cell (2). Exogenous retroviruses infect and integrate into the genomes of somatic cells. A subset of infections may occur in the host germ line, which can result in vertical transmission of the retrovirus in a Mendelian fashion (2–4). Vertically transmitted retroviruses are called endogenous retroviruses (ERVs) and comprise up to 10% of known vertebrate genomes (1, 4, 5).

Most ERVs represent proviral integrations that occurred millions of years ago (1, 5). ERVs are often rendered inactive by the accumulation of mutations over time, with only a small proportion of proviruses capable of producing functional retroviruses (2, 4). Frequent cross-species transmission of retroviruses and formation of ERVs creates a discordance between a large portion of the genome and the host phylogeny where, in some cases, very closely related species do not share ERVs. The best characterized ERVs are those of humans and mice, with other species being relatively poorly characterized (3). High-throughput sequencing (HTS) and genome assemblies from a large number of vertebrates in recent years have enabled characterization of novel ERVs. However, the sequencing coverage of genomes varies enormously between species, and depending on the sequencing technology used (long versus short reads), the number of gaps, and the number of ERV copies, ERVs can be missed when comparing genomes (1, 6).

The *Perissodactyla* order represents a diverse and widespread group of mammals found on all continents except Australia and Antarctica (excluding domestic horses, which are broadly distributed worldwide). Perissodactyls are further divided into two suborders, the *Hippomorpha* and *Ceratomorpha*. The *Hippomorpha* include all eight extant equid species, while the *Ceratomorpha* includes the four extant tapir species and the five extant species of rhinoceros (7). Endogenous gammaretroviruses have not been described in *Perissodactyla*, and a recent study of ERVs in this order indicated that they are absent (8).

Four white rhinoceros died following severe neurological abnormalities at the Taronga Western Plains Zoo, New South Wales, Australia. In an effort to identify the causative infectious agent, HTS, DNA, and cDNA data were generated for the four white rhinoceros. Analysis of the data identified gammaretroviral sequences similar to *Mus caroli* endogenous gammaretrovirus (McERV). Screening additional perissodactyl genome assemblies revealed several similar ERV loci in the assemblies of white rhinoceros and black rhinoceros. The identified sequence was absent from all other *Perissodactyla* genome assemblies screened. The white rhinoceros full-length assembled endogenous gammaretrovirus was designated SimumERV, and its close relative from black rhinoceros was designated DicerosERV. We describe the species distributions, copy numbers, evolutionary ages, and phylogenetic relationships of SimumERV and DicerosERV with other gammaretroviral sequences.

RESULTS

Identification of SimumERV sequence in white rhinoceros HTS data. Four white rhinoceros died at the Taronga Western Plain Zoo after showing neurological abnormalities. HTS was performed on total RNA extracts from lung tissue samples and DNA extracts from blood and placenta. HTS RNA and DNA data were analyzed using the viral identification pipeline (VIP) (9) to identify viral sequences in the samples. VIP analysis identified gammaretroviral reads similar to Gibbon ape leukemia virus (GaLV) and *Mus dunni* endogenous retrovirus (MDEV) in both DNA and RNA sequenced samples. The longest generated contigs from each sample were aligned, and an 8,536-bp majority-rule consensus sequence was generated. The consensus sequence was searched against the NCBI database using BLASTn. The BLASTn search results indicated highest similarity to McERV, with a query coverage of 53% and 70.12% identity. RetroTector (10) analysis verified that the generated consensus sequence belongs to the gammaretrovirus genus and identified proviral *gag*, *pro*, *pol*, and *env* genes and long terminal

TABLE 1 Rhinoceros NCBI assembly genomes used for screening of SimumERV_cons

Species	Subspecies	ERV BLASTn hits[a]	Full-length proviral hits[a]	LTR-A BLASTn hits[a]	LTR-B BLASTn hits[a]	Accession no.
Ceratotherium simum (white rhinoceros)	*Ceratotherium simum simum*	44	0	369	0	GCA_000283155.1
	Ceratotherium simum cottoni	2,680	0	245	0	GCA_004027795.1
	Ceratotherium simum cottoni	1,255	98	467	0	GCA_021442165.1
Diceros bicornis (black rhinoceros)	*Diceros bicornis bicornis*	83	1	54	130	GCA_004027315.2
	Diceros bicornis minor	353	153	101	373	GCA_020826835.1
	Diceros bicornis minor	65	1	52	134	GCA_013634535.1
Dicerorhinus sumatrensis (Sumatran rhinoceros)	*Dicerorhinus sumatrensis sumatrensis*	0	0	0	0	GCA_002844835.1
	Dicerorhinus sumatrensis harrissoni	0	0	0	0	GCA_014189135.1
Rhinoceros unicornis (Indian rhinoceros)	*Rhinoceros unicornis*	0	0	0	0	GCA_019022865.1
	Rhinoceros unicornis	0	0	0	0	GCA_018403435.2
Coelodonta antiquitatis (woolly rhinoceros)	*Coelodonta antiquitatis*	0	0	0	0	ERX3761614, ERX3761620, SRX9737591, SRX9737592
Stephanorhinus kirchbergensis (Merck's rhinoceros)	*Stephanorhinus kirchbergensis*	0	0	0	0	SRX9738793

[a] For each genome, we indicate the number of high-significance BLASTn hits, the number of full-length proviruses, and the number of LTR sequences that were identified.

repeat (LTR) sequences. The complete proviral consensus sequence was also compared to entries in the NCBI Conserved Domain Database (CDD) (11) and identified motifs for all retroviral genes. The consensus of the novel proviral sequence identified from the four white rhinoceros was named SimumERV_cons (NCBI accession number OP081083).

SimumERV integration site identification in extant and extinct rhinoceros genomes. The SimumERV consensus sequence generated from the four sequenced rhinoceros samples was used as a seed to screen the available genome assemblies of extant rhinoceros downloaded from the NCBI Assembly database. BLASTn searches using the SimumERV_cons sequence, excluding LTRs, were performed on all scaffolds of white rhinoceros and black rhinoceros draft genomes and identified a total of 3,979 and 501 hits, respectively (Table 1) (12). The genome assemblies of Sumatran rhinoceros (*Dicerorhinus sumatrensis*) and Greater Indian rhinoceros (*Rhinoceros unicornis*) did not produce any positive hits (Table 1). The SimumERV_cons sequence was also used as a seed to screen extinct rhinoceros species genomes. Raw sequence data for extinct rhinoceros were obtained from the NCBI SRA database (13) and analyzed using the same methodology as indicated above. The screening of Woolly rhinoceros (*Coelodonta antiquitatis*) and Merck's rhinoceros (*Stephanorhinus kirchbergensis*) did not produce any positive hits (Table 1).

The resulting proviral BLASTn hits plus ~8,000 bp of both up- and downstream scaffold regions were extracted from the respective draft genomes. The scaffold regions extracted from white rhinoceros were analyzed using RetroTector, identifying 98 full-length proviral sequences from the long-read (Oxford Nanopore MinION) GCA_021442165.1 genome assembly. The other two available white rhinoceros NCBI assemblies failed to produce full-length provirus hits (Table 1), most likely due to limitations for assembly of repetitive elements that arise in shorter read-length sequencing approaches, such as Illumina sequencing by synthesis (SBS) (10, 14). Gammaretroviral sequences with high identity to the seed sequence were identified in all white rhinoceros extracted scaffold regions. The extracted gammaretroviral full-length proviral sequences from GCA_021442165.1 were aligned using MAFFT (15) (Fig. S1 in the supplemental material). From the resulting alignment, a consensus sequence was generated that we designated SimumERV. The newly generated consensus sequence SimumERV has a 99.3% pairwise identity to the SimumERV_cons seed sequence used above (Fig. 1). White rhinoceros assembly GCA_021442165.1 screening also revealed another 56 proviral sequences with partial SimumERV *env* sequence matches with a nucleotide identity of, on average, 96.5% and LTRs nearly identical in sequence. Recombination analysis of SimumERV proviral sequences using recombination analysis tool (RAT) and

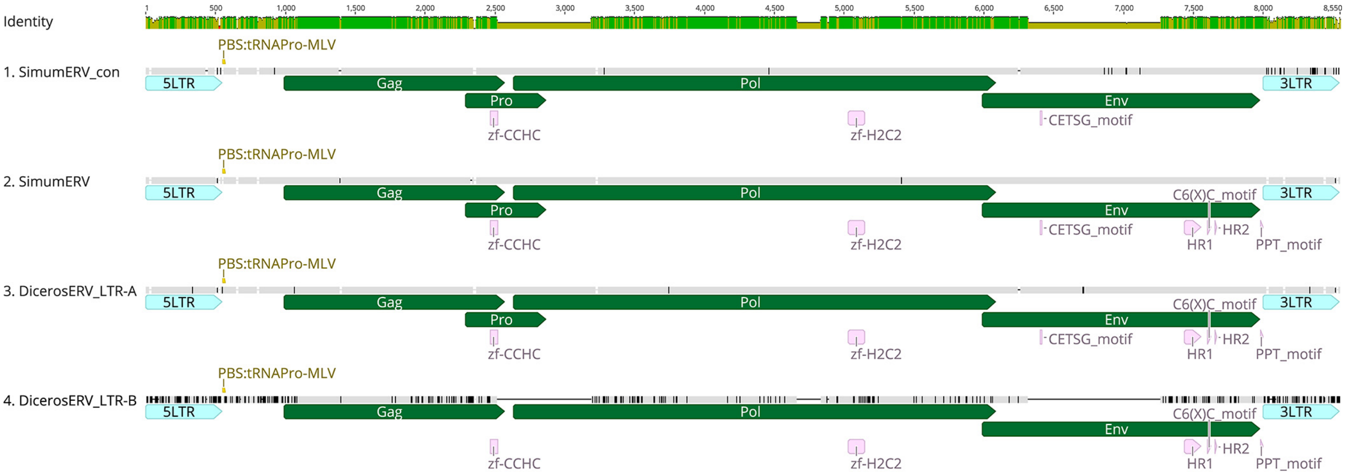

FIG 1 Multiple alignment of SimumERV and DicerosERV consensus sequences from white rhinoceros next-generation sequencing data (SimumERV_cons), white rhinoceros (SimumERV), and black rhinoceros (DicerosERV) genomes. Sequences are annotated indicating the location of LTRs in light blue, primer binding sites (PBSs) are in yellow, proviral genes *gag*, *pro*, *pol*, and *env* are illustrated in green, and identified pathogenicity and conserved domain motifs are illustrated in pink. The four sequences have an average pairwise nucleotide identity of 86.3%. Sequence differences among the sequences are indicated by black vertical lines.

RECCO scripts (16, 17) indicated a potential recombination event in the above-mentioned subset of 56 sequences, with the first and second breakpoint located within the region surrounding nucleotides 1099 and 7498 of the 56 aligned sequences used as input (Fig. S2). Further analysis of the recombinant subset using Repbase suggests that SimumERV recombined with another rhinoceros gamma endogenous retroelement (18). The SimumERV recombinant appears to be a single recombinant event that subsequently expanded in the *C. simum* genome. The 56 recombinant proviral sequences were excluded from the data set since they most likely represent postgenome colonization events, and we focused on the nonrecombinant sequences identified to define the original colonization process.

The scaffold regions extracted from all black rhinoceros genomes revealed 155 full-length proviral sequences with all genes in the sense orientation (Table 1). From the remaining 351 extracted scaffolds, 2 appeared to be full proviral sequences but with gene duplications and partial LTRs, while the remaining extracted scaffolds contained partial proviral genes and LTRs. All proviral sequences extracted from black rhinoceros assembly GCA_020826835.1 were aligned using MAFFT (15) (Fig. S1). Alignment of the extracted proviral sequences revealed the presence of two groups of LTRs in the black rhinoceros extracted scaffolds. The first group was designated LTR-A, with a high sequence similarity (~98%) to the identified SimumERV LTR sequences. The second group was called LTR-B, with a sequence similarity of ~80% and including sequences that are unique to black rhinoceros proviruses (Fig. S3). The black rhinoceros proviral sequences were divided according to the LTR type and realigned using MAFFT (15). LTR-A sequences were observed in 20 proviral sequences, while the remaining 133 proviral sequences had LTR-B sequences. The LTR-A proviruses had fewer truncations affecting their genes than the LTR-B proviruses. LTR-B proviruses had larger proportions of proviruses with truncated *gag*, *pro*, *pol*, and *env* genes and a higher copy number (Fig. S1). Majority-rule consensus sequences of the two alignments were generated and named DicerosERV LTR-A and DicerosERV LTR-B. DicerosERV LTR-A and DicerosERV LTR-B sequence comparisons to the SimumERV sequence revealed pairwise identities of 99.7% and 73.3%, respectively (Fig. 1).

Analysis of SimumERV and DicerosERVs using the NCBI CDD resulted in the identification of retroviral motifs for all retroviral genes (11). RetroTector (10) analysis of consensus sequences further verified the presence of all retroviral genes and LTRs (Fig. 1). RetroTector was able to generate ERV reconstructed protein sequences for each gene from the majority-rule consensus sequences. The CETTG pathogenicity motif that is conserved in exogenous and highly infectious gammaretroviruses appears to contain a threonine-to-serine mutation in SimumERV and DicerosERV sequences, resulting in a CETSG motif (Fig. 1) (19). The CETSG motif was recently observed in an exogenous replication-competent retrovirus called Hervey pteropid gammaretrovirus (HPG) that is circulating among bats as well as in a significant percentage (27%) of koala retrovirus-D proviruses (KoRV-D). The amino acid change resulting in the CETSG motif is hypothesized to attenuate syncytium formation (20). Further analysis of the majority-rule consensus sequences identified tRNA-Pro as the primer binding site for viral replication initiation. The generated consensus sequences were submitted to Repbase and are also provided in the Supplementary Information (18).

Target site duplication (TSD) sites flanking the proviral LTRs were also examined. Retroviral integration generates 4 to 6 bp TSD sites directly flanking the proviral sequence (21). White and black rhinoceros proviral sequences displayed a 4-bp TSD pattern (Table S1). Further analysis of the coding potential of SimumERV and DicerosERV proviral loci revealed that the majority of sequences were heavily mutated or had partial internal coding regions missing, a result that indicates that most of the proviruses are unlikely to produce any functional proteins. The DicerosERV LTR-B proviral sequences JAJIAZ010000042.1 (15264763 to 15296245), JAJIAZ010000006.1 (3526731 to 3550240), and JAJIAZ010000006.1 (9540268 to 9563788) are an exception as they may have coding potential and the ability to produce full-length protein sequences.

LTR sequences from the SimumERV and DicerosERV consensus sequences were used as a seed for BLASTn searches in the white and black rhinoceros genome assemblies, identifying 1,081 and 844 unique hits, respectively (Table 1). LTR-A sequences were present in both black and white rhinoceros genomes. LTR-B sequences were only identified in black rhinoceros genome data (Fig. 2). To determine whether the identified proviruses and solo LTRs share the same integration sites within black and white rhinoceros, extended LTR sequences (±500 bp) were extracted from white and black rhinoceros genomes (GCA_021442165.1 and GCA_020826835.1). Extended LTR sequences were used as queries for BLASTn searches on HTS data from the four white rhinoceros and all other available NCBI white and black rhinoceros genomes. Integration site comparisons revealed identical insertion sites within different black and white rhinoceros individuals. Integration site comparison across white and black rhinoceros failed to identify common integration sites. This conclusion is based on the current quality of the assembled genomes currently available.

Age estimation of DicerosERV and SimumERV. Evolutionary ages of DicerosERV and SimumERV sequences were estimated using three different approaches. First, LTR sequences identified by BLASTn search were aligned using MAFFT (12, 15). The three resulting alignments were manually curated and separated into subgroups based on sequence similarity (Fig. S4), and Kimura-2-parameter (K2P)-corrected distances for LTR sequences compared to the majority-rule consensus sequences were calculated (22). CpG dinucleotide positions were excluded from the analysis, as they are prone to higher mutation rates due to 5-methyl spontaneous deamination (5, 23). Using calculated distances for each LTR alignment and a previously published African rhinoceros mutation rate of 0.00088/nucleotide/year (24), the SimumERV subgroup 1 age was estimated to be ~12.97 (±13.28) million years (Myr) old, the SimumERV subgroup 2 age was estimated to be ~16.28 (±6.92) Myr old, the DicerosERV LTR-A-1 age was estimated to be ~2.72 (±5.50) Myr old, the DicerosERV LTR-A-2 age was estimated to be ~8.36 (±6.22) Myr old, and the DicerosERV LTR-B age was estimated to be ~2.72 (±5.5) Myr old (Fig. 3; Tables S2 to S4).

The second approach was based on the same principle as the first approach but used proviral gene regions instead of LTR sequences. Briefly, multiple alignments were created for each proviral gene using MAFFT, and each alignment was manually curated and separated into subgroups based on sequence similarity (Fig. S5 to S7), and K2P-corrected distances for each gene sequence compared to the corresponding gene's majority-rule consensus were calculated (15, 22). For the second approach, the same African rhinoceros mutation rate was applied, and CpG dinucleotides were excluded.

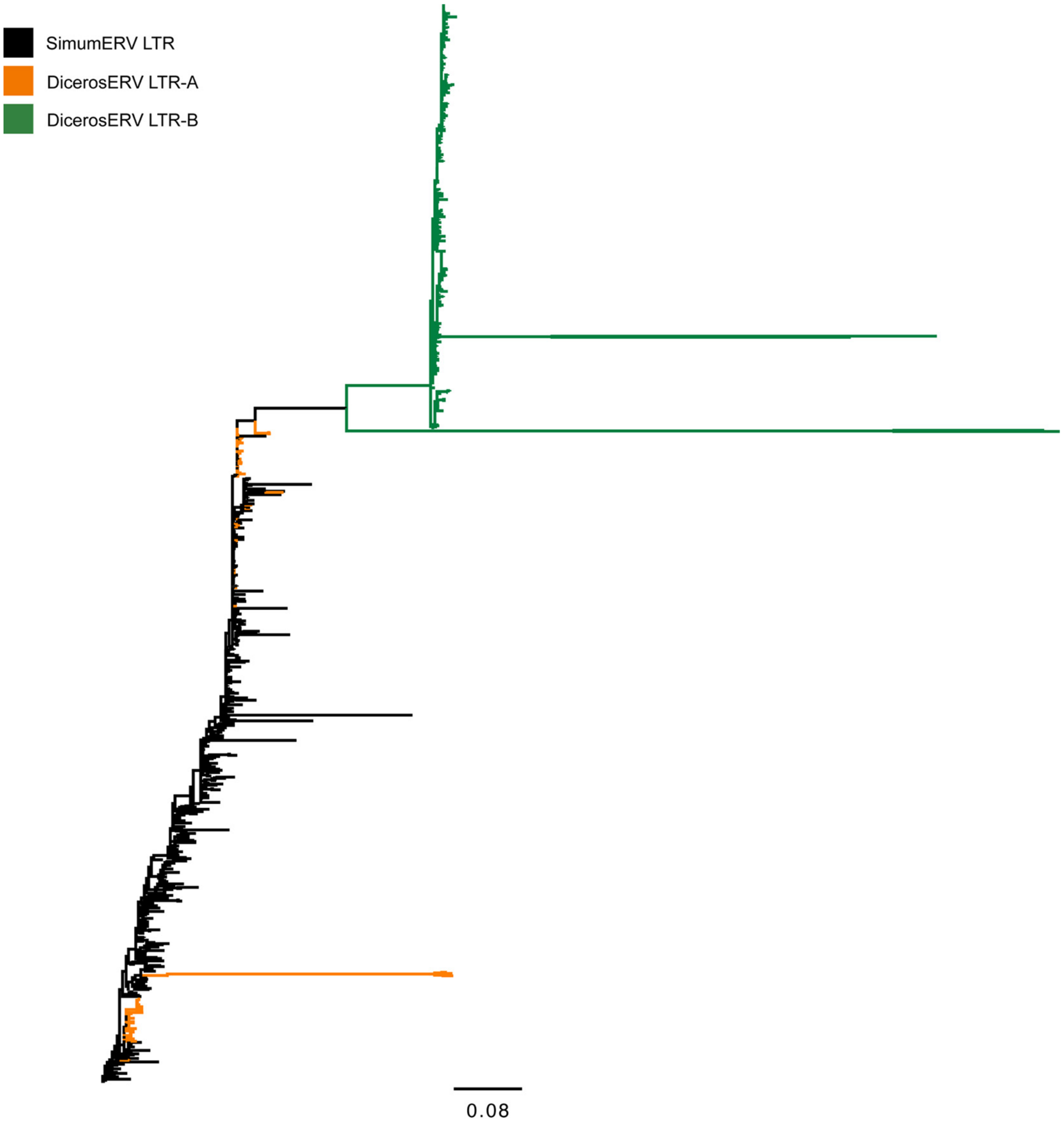

FIG 2 Phylogenetic relationship of SimumERV and DicerosERV LTR sequences. The unrooted maximum likelihood tree illustrates the 463 SimumERV and 470 DicerosERV proviral and solitary LTR sequences identified in white and black rhinoceros genomes excluding sequence outliers. Black clades represent SimumERV LTRs, orange clades represent DicerosERV LTR-A, and green clades represent DicerosERV LTR-B. Sequences that were identified in BLASTn searches as related to the rhinoceros ERVs but that aligned very poorly and may represent false hits were excluded from the analysis. However, sequences that did align well along the proviral sequence but were more divergent from other proviral sequences for shorter internal regions were not excluded as outliers and are visible as long branches.

The *gag* gene-based age estimates were ~12.95 (±11.18) Myr old for SimumERV. The DicerosERV LTR-A *gag* gene was separated into two subgroups with age estimations of 2.90 (±1.28) and 5.22 (±3.89) Myr old (Fig. 3). The DicerosERV LTR-B *gag* gene was grouped into 2 clusters with age estimations of ~5.43 (±3.77) and ~9.52 (±4.21) Myr old. The *pol* gene age estimations were ~13.60 (±4.68) and 7.91 (±7.03) Myr old for the SimumERV subgroups, ~2.90 (±1.35) and 4.97 (±3.45) Myr old for the DicerosERV LTR-A subgroups, and ~4.12 (±2.57), 9.70 (±6.79), 14.20 (±6.15), and 9.99 (±3.07) Myr old for DicerosERV LTR-B (Fig. 3). Age estimations of the *env* gene were ~7.11 (±5.60), ~16.71 (±9.11), and ~8.05 (±5.10) Myr old for SimumERV, ~2.76 (±2.06) and ~4.23 (±2.04) Myr old for the DicerosERV LTR-A subgroups, and ~11.02 (±5.13), ~10.79 (±8.40), and ~3.20 (±1.94) Myr old for DicerosERV LTR-B (Fig. 3; Tables S5 to S13).

The third approach was based on nucleotide divergence of proviral 5′ and 3′ LTRs. Once the provirus is integrated into the genome, the two LTRs have identical sequences

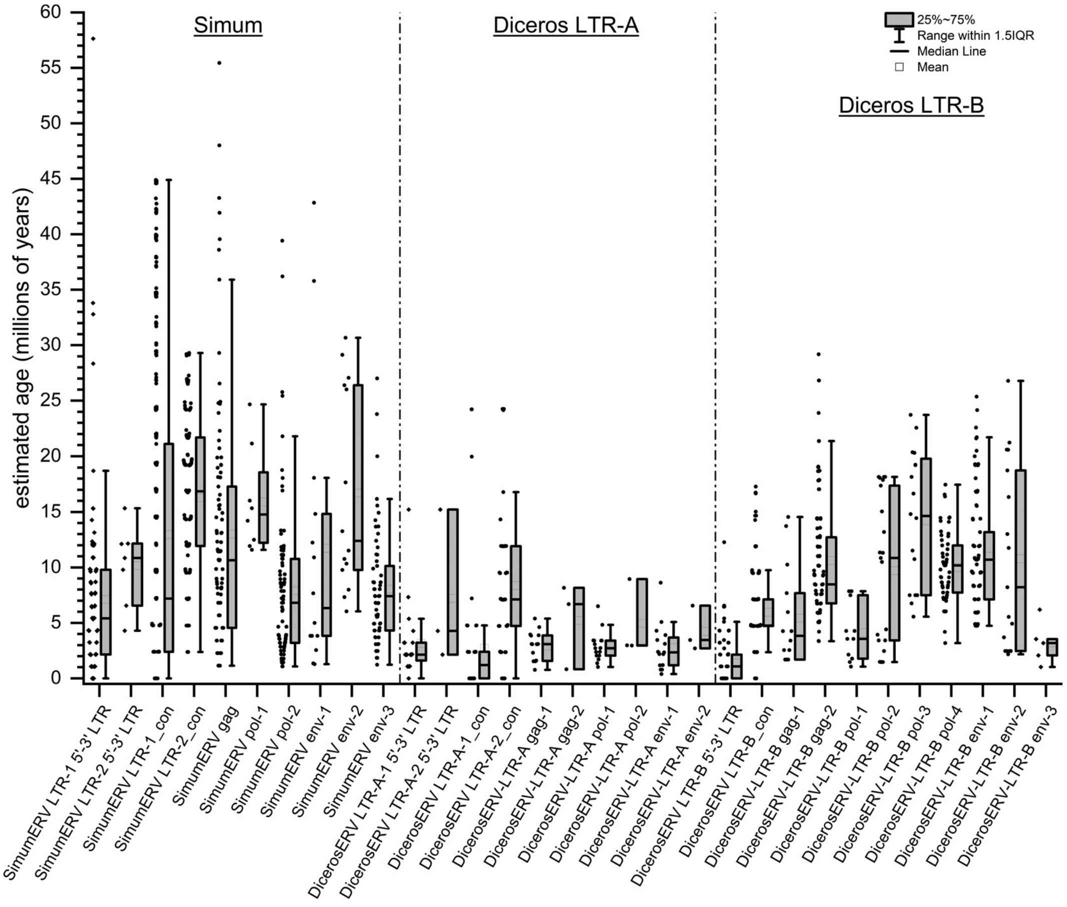

FIG 3 Boxplots illustrating age estimations of SimumERV and DicerosERV loci using multiple approaches. For more accurate age estimations, ERV sequence alignments were separated into subgroups based on sequence similarity. Age estimations were performed comparing the *gag*, *pol*, and *env* gene regions of each provirus as well as LTRs (solitary and proviral; "con") to the consensus sequences of its respective subgroup. Furthermore, provirus ages were estimated via the number of nucleotide differences between individual proviral 5′ to 3′ LTRs. Individual ERV locus (proviruses, solitary LTRs) age estimations are illustrated with black dots next to each subgroup's boxplot. Black lines indicate the median age, squares indicate the mean, and whiskers indicate the 1.5-fold interquartile range.

(1). Mutations in the LTR regions, as in the rest of the provirus, will begin to accumulate at the host mutation rate (5). The age of integration can be determined based on the host mutation rate and number of nucleotide differences between the two proviral LTRs. Using the above-mentioned mammalian mutation rate and calculated nucleotide distances between each proviral LTR pair, evolutionary ages were estimated for each provirus. SimumERV proviruses were estimated to be ~7.10 (±8.28) Myr old, DicerosERV LTR-A proviruses were estimated to be ~3.17 (±3.28) Myr old, and DicerosERV LTR-B proviruses were estimated to be 1.29 (±1.95) Myr old (Fig. 3; Tables S14 to S16).

In all dating approaches, sequences with extreme nucleotide distances, thus evolutionary ages, were considered to be outliers, also evident in multiple sequence alignments, and were excluded from age estimations. The different age estimation approaches of the investigated rhinoceros ERVs indicate a higher heterogeneity of the overall LTR collection than age estimates based on 5′ and 3′ proviral LTRs. Proviral 5′ and 3′ LTR sequence homogeneity could indicate gene conversion events that reduce the genetic distance between the two LTR sequences, resulting in underestimation of integration times (25–27). Several proviral 5′ and 3′ LTR sequences also appear to be identical in sequence to one another, indicating that those ERVs might still be actively proliferating or formed very recently (28). Also, the variable dates observed between the two DicerosERV LTR groups may reflect multiple independent infections or integrations into the germ line by the original exogenous retroviral ancestor, as observed for other recent germ line infections (e.g., the koala retrovirus

KoRV) (29). Alternatively, the variable dates may be due to different time periods of expansions in the germ line. The estimated average of all calculations is that the overall colonization of the rhinoceros genome took place in the last 8 Myr.

Subgroup analysis of DicerosERV and SimumERV. Analysis of the different subgroups used in the age estimation indicated that for SimumERV proviral sequences, LTR subgroup 2 was more likely to form proviral sequences containing *pol* subgroup 1 and *env* sequences not belonging to subgroup 1. The DicerosERV LTR-A subgroup analysis indicated that LTR subgroup 2 sequences were not associated with viral genes of subgroup 1, whereas LTR subgroup 1 was associated with viral genes of subgroup 1. Subgroup analysis of DicerosERV LTR-B sequences indicated that *gag* gene subgroup 1 was generally associated with other subgroup 1 sequences (Tables S17 to S19).

SimumERV molecular screening in white rhinoceros samples. Sequence-specific PCRs were also performed in DNA extracted from white rhinoceros samples. Molecular screening of SimumERV was performed to further confirm the presence of the bioinformatically identified ERVs. Genomic DNA was subjected to PCR using specific primers designed to amplify approximately 400 bp in the *gag* and *env* gene region each. PCR products of the expected size were produced for both *gag* and *env* primer sets. Sanger sequencing of the PCR products confirmed the targeted *gag* and *env* sequences, demonstrating little sequence divergence from the three generated consensus sequences (Fig. S8). This approach further verified the presence of the identified endogenous retrovirus in the genomes of African rhinoceros.

Phylogenetic analysis of SimumERV and DicerosERV sequences. SimumERV and DicerosERV consensus proviral nucleotide and protein sequences, the latter generated by RetroTector, were searched using BLAST against the NCBI database. Retroviral and endogenous retroviral sequence matches with significant homology to the query sequences were downloaded from the NCBI database. Sequences were multiply aligned along with the query sequences using MAFFT (12, 15). Maximum likelihood phylogenetic analysis was performed using reticuloendotheliosis virus as an outgroup for both nucleotide and protein analyses. Nucleotide phylogeny placed the identified *Rhinocerotoidea* gammaretroviruses SimumERV and DicerosERV in a monophyletic group that forms a sister clade with several rodent ERVs, including African thicket rats (*Grammomys surdaster*) and African grass rats (*Arvicanthis niloticus*) that are found only in Africa (Fig. 4). Protein phylogenetic analysis also resulted in the same outcome, with *Rhinocerotoidea* gammaretroviruses being in the same clade in all protein phylogenies. In the Env protein phylogeny along with the consensus sequences of SimumERV and DicerosERVs, we also included protein sequences generated from two DicerosERV LTR-B nontruncated proviral sequences, JAJIAZ010000002.1 (94775313 to 94798831) and JAJIAZ010000006.1 (3526731 to 3550240), with the analysis placing them as sister groups with SimumERV and DicerosERV LTR-A (Fig. S9). In a Gag and Pol protein phylogenetic analysis, the *Rhinocerotoidea* gammaretroviruses clade was forming a sister clade with the rodent ERVs as in the nucleotide analysis. The Env protein phylogeny, on the other hand, resulted in a polytomy most likely due to insufficient informative sites (Fig. S9).

DISCUSSION

Previous analysis of perissodactyl genomes suggested that gammaretroviruses are absent from that order (8). However, in that analysis, the only available rhinoceros genomes were generated with short-read sequencing approaches that are known to have several limitations when analyzing repetitive elements, and, therefore, lower-copy or incomplete retroviral sequences could be missed (14). From the genomes screened in Zhou et al., we identified only two DicerosERV full-length proviruses and no full-length proviral sequences of SimumERV. High-quality white and black rhinoceros genomes were released in 2021 and 2022 from the Max Planck Institute of Molecular Genetics and the Vertebrate Genome Project, respectively, using long-read sequencing approaches (30). Screening of those genome sequences identified 98 SimumERV and 153 DicerosERV proviral sequences. The use of the viral search tool GLUE based on structural proteins that was used in Zhu et al. (8) would likely have

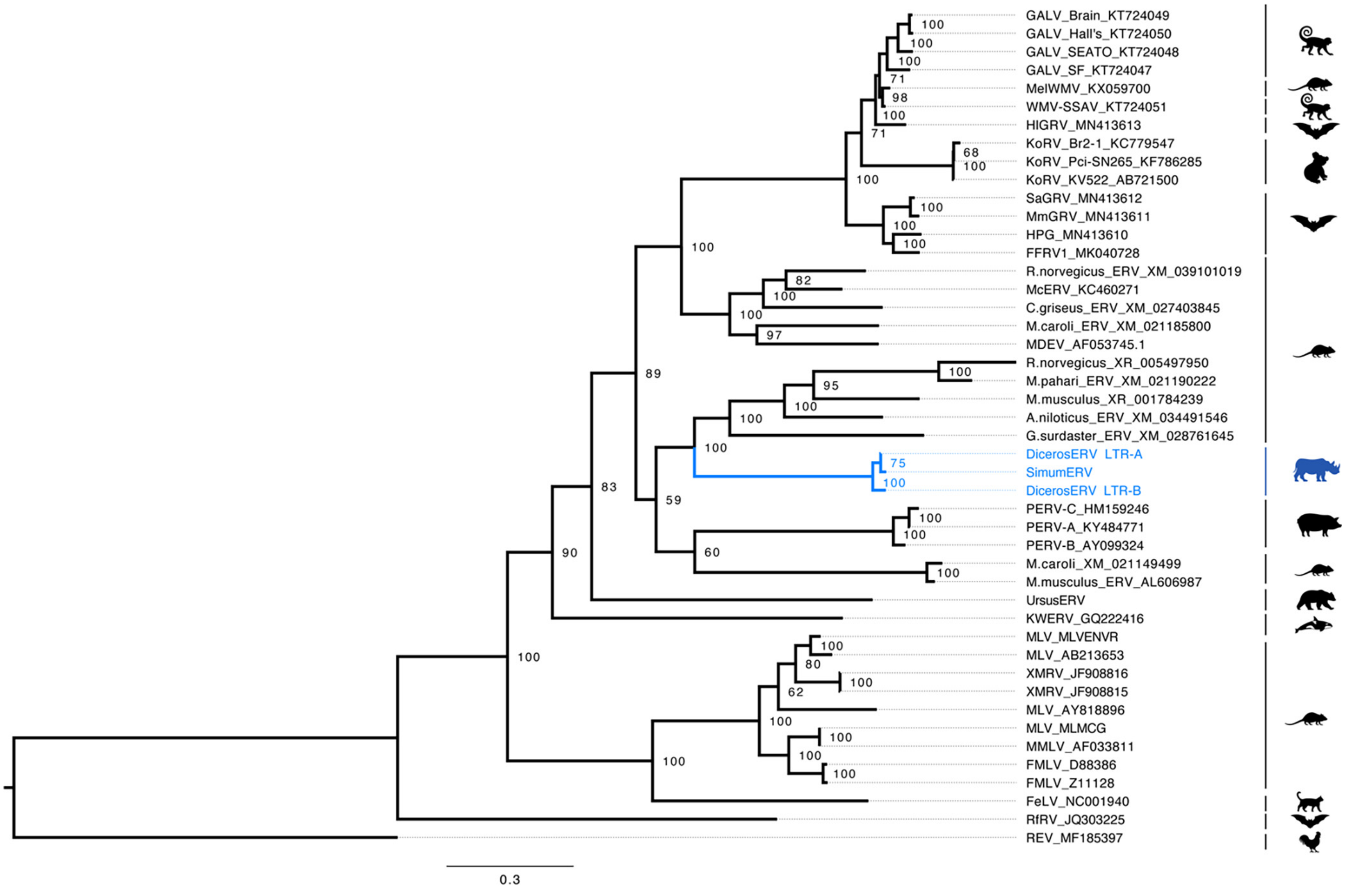

FIG 4 Phylogenetic analysis of SimumERV and DicerosERV consensus nucleotide whole-genome sequences excluding LTRs within the family *Retroviridae*. A phylogenetic tree was constructed using RAxML and the GTR gamma substitution model with 20 maximum likelihood searches and 500 rapid bootstrap replicates. Bootstrap support is given at nodes. Reticuloendotheliosis virus was used as an outgroup. The scale bar represents nucleotide substitutions per site.

therefore missed SimumERV and DicerosERV given the partial coverage of the majority of proviruses in the 2012 genome builds for *Ceratotherium simum* and *Diceros bicornis*. This is clearly the case, as both long-read data assemblies and HTS virome analysis from lung, placenta, and blood samples from four white rhinoceros samples identified a gammaretrovirus most closely related to an African rat ERV that was present in all African rhinoceros samples tested here or in the current genome databases.

Dating of LTR and gene divergence suggests that one or multiple germ line integrations occurred between 1.5 and 15 Mya. This is consistent with the complete absence of this retroviral clade from Asian rhinoceros, as the estimated divergence between Asian and African rhinoceros clades is approximately 16 Mya (31). The lower 1.5-Mya estimate for the LTR-B group that was only identified in black rhinoceros indicates a colonization event by this LTR group after the split of white and black rhinoceros. Given the uncertainty of the age estimates, it is still conceivable that all germ line invasions preceded the divergence of the two African rhinoceros clades, but at least some of the colonization events may have occurred close to the time of host lineage divergence.

The two different LTR groups suggest several possible scenarios for how these ERVs have proliferated, which we cannot currently distinguish. First, it is possible that the exogenous retroviral ancestor or relatives of it remained in circulation in a reservoir host in Africa and that the black rhinoceros ERVs represent two independent genome colonizations separated in time. Second, at some point after the separation of the lineages leading to the white and black rhino clades, an intracellular retrotransposition, perhaps by one of the more intact LTR-A ERVs, produced the LTR-B clade with no exogenous retroviral involvement. Given the high similarity between SimumERV and DicerosERVs, the latter scenario is more likely, as over millions of years, it would be expected that the ancestral retrovirus would have diverged substantially between the genome colonization events.

The most closely related sequences to SimumERV and DicerosERV are found in a clade of rodent ERVs that includes the sympatric African grass rat (*Arvicanthis niloticus*) and the African thicket rat (*Grammomys surdaster*). While it is hard to infer the direct transmission route for events millions of years in the past, the sympatry of both hosts with African rhinos and the lack of SimumERV and DicerosERV from Asian rhinoceros suggests that genome colonization of African rhinos occurred exclusively in Africa without involving any of the extant or extinct Asian rhino lineages (32). Further analysis of genomes from additional sympatric species, particularly rodents, may identify additional, more closely related ERVs that may further clarify the rhinoceros gammaretrovirus origins.

MATERIALS AND METHODS

Samples and nucleic acid extraction. White rhinoceros tissue and blood samples were provided from the Taronga Western Plains Zoo in New South Wales, Australia. Four female white rhinoceros (*Ceratotherium simum simum*) died in an Australian open range zoo over a 4-week period. Disease investigation failed to identify an etiology. No animal experiments were performed, and samples were gathered as a part of standard veterinary care at the Taronga Western Plains Zoo.

DNA and RNA were extracted from lung, placenta, and blood samples using a Qiagen RNeasy kit following the manufacturer's protocol with minor modifications (33). Briefly the tissue samples were lysed overnight, while the blood sample was lysed for 1 h. Eluted DNA/RNA sample integrity and quantity was determined using an Agilent 2200 TapeStation (Agilent Technologies).

Illumina library construction and HTS. RNA samples were reverse transcribed using a RevertAid first-strand cDNA synthesis kit from Thermo Fisher Scientific according to the manufacturer's instructions. Second-strand synthesis was performed using Klenow DNA polymerase I (Thermo Fisher Scientific) as described in Dayaram et al. (33). The resulting cDNA/DNA mix and extracted DNA samples were sonicated to an average size of 300 bp using a Covaris M220. Fragmented samples and negative controls were further processed to generate dual-index Illumina libraries, as previously described (34). Each library was amplified in triplicate reactions to minimize PCR bias. The three reactions of each sample were pooled after amplification and cleaned up using a MinElute purification kit (Qiagen). Quantification and fragment size distribution of each Illumina library was assessed using an Agilent 2200 TapeStation (Agilent Technologies). Dual-index libraries were then pooled equimolarly to a final concentration of 17.5 nM and paired-end sequenced on an Illumina NextSeq platform in the Berlin Center for Genomics in Biodiversity Research (BeGenDiv).

Bioinformatic analysis and viral screening. Generated raw BCL files were demultiplexed and sorted based on the indices of each sample using the bcl2fastq Illumina conversion software (SRA accession number PRJNA862320). Cutadapt v2.6 was used to quality trim and size filter short reads from the resulting fastq files. Viral screening was performed on each paired data set using a modified version of the Viral Identification Pipeline (VIP) v2 (9). VIP analysis was performed using the sense mode that includes both nucleotide and amino acid alignments and as recommended for viral discovery. The pipeline first aligned the next-generation sequencing data to a reference genome and filtered out all the reads that aligned to it. The remaining reads were further filtered using a bacterial database. The data that were not removed in the first two steps were aligned to nucleotide and amino acid viral sequence databases. VIP-identified viral reads were then de novo assembled to create longer contigs with Velvet (35). Resulting contigs were then aligned using MAFFT to generate a majority-rule consensus (15).

Generation of SimumERV provirus and LTR consensus sequence. Retroviral majority rule consensus sequence generated from the next-generation sequencing data were used as a probe to screen the *Ceratotherium simum simum* genome sequence assembly (GCA_021442165.1). Positive hits were extracted and aligned using MAFFT v.7.450 (15). The resulting alignment was manually curated, and proviral and LTR majority-rule consensus sequence was generated using Geneious Prime 2022 v2. RetroTector was used to generate putative and reconstructed retroviral proteins for all retroviral genes based on the proviral consensus sequence (10). Conserved retroviral motifs were identified using NCBI CCDs, further verifying the RetroTector findings (11).

Recombination inference analysis. Recombination analysis was performed for multiple sequence alignments using two computational methods: recombination analysis tool (RAT) and RECCO (16, 17). Both computational methods were used with default parameters.

Genome assembly screening. *Perissodactyla* genomes from the NCBI assembly database were downloaded to determine if the identified transcriptome and white rhinoceros genome proviral sequences were present in other members of the order. The following *Perissodactyla* genomes were downloaded from the NCBI assembly database: *Ceratotherium simum simum* (GCA_000283155.1), *Ceratotherium simum cottoni* (GCA_004027795.1), *Diceros bicornis bicornis* (GCA_004027315.2), *Diceros bicornis minor* (GCA_020826835.1 and GCA_013634535.1), *Dicerorhinus sumatrensis harrissoni* (GCA_014189135.1), *Dicerorhinus sumatrensis sumatrensis* (GCA_002844835.1), *Rhinoceros unicornis* (GCA_019022865.1 and GCA_018403435.2), *Equus caballus* (GCA_002863925.1), *Tapirus indicus* (GCA_004024905.1), and *Tapirus terrestis* (GCA_004025025.1). A custom BLASTn database was constructed for each genome in Geneious

Prime 2022 v2. Transcriptome and genome majority-rule consensus proviral reads were used as query sequences, and each genome was screened using BLASTn script with default parameters. Positive hits from each genome were extended 8,000 bp on each site and aligned using MAFFT v.7.450 (15). Each alignment was then manually curated, and a majority-rule consensus sequence was generated.

SRA data screening. Woolly rhinoceros (*Coelodonta antiquitatis*; ERX3761614 to ERX3761620 and SRX9737591 to SRX973759) and Merck's rhinoceros (*Stephanorhinus kirchbergensis*; SRX9738793) genomic data were downloaded from the NCBI's SRA (13). The SRA files downloaded were converted to fastq files using SRA toolkit (13). Fastq files from woolly rhinoceros and Merck's rhinoceros were converted into custom BLASTn databases using Geneious Prime 2022 v2. SimumERV consensus sequences from the white rhinoceros genome and transcriptomes were used as query sequences to screen the custom BLASTn databases using BLASTn script with default parameters.

SimumERV PCR. SimumERV regions were amplified from the genomic DNA of a white rhinoceros (Lima-2016-179) extracted from blood using the following primer pairs that target regions in *gag* and *env* genes, respectively: Csim_GAG_ F3 (5′-TGCCATCTTTGCCCAGTAGG-3′), Csim_GAG_ R3 (5′-AGATGAGTCGGG GCTCAGAA-3′), Csim_ENV_F5 (5′-GACTCCGCTGTTCGAGTTGA-3′), and Csim_ENV_R5 (5′-ACCTCATTTGACGG GATGGG-3′). PCRs were performed in 22-μL reaction volumes containing 12.5 μL of MyTaq polymerase mix 2× (Bioline, Meridian Biosciences, Heidelberg, Germany), 454.5 nM each primer pair, and 1 μL of DNA template. Thermocycling conditions were 95°C for 5 min followed by 35 cycles of 95°C for 20 s, 55°C for 20 s, and 72°C for 4 min, with a final extension step at 72°C for 2 min. Amplified products were visualized on 1.5% (wt/vol) agarose gels using 6× Orange loading dye (Life Technologies GmbH, Thermo Fisher, Darmstadt, Germany). The PCR amplification products were Sanger sequenced using the above-mentioned forward and reverse primers (LGC Genomics, Berlin, Germany).

Phylogenetic analysis. Multiple alignments were generated using MAFFT v.7.450 followed by manual curation and refinement (15). Phylogenetic analysis was performed on the majority-rule consensus genome sequences, the major retroviral gene reconstructed proteins (Gag, Pol, and Env), and related sequences obtained from GenBank: REV (MF185397, ASH96780, ASH96781, and ASH96782), RfRV (JQ303225, AFA52558, AFA52559, and AFA52560), FeLV (NC_001940, NP_047255, and NP_047256), FMLV (Z11128, CAA77479.1, and CAA77478.1), FMLV (D88386, BAA22066.1, BAA22065.1, and BAA22064.1), MMLV (AF033811, AAC82568.1, AAC82566.1, and AAC82567.1), MLV (MLMCG; AAB59942.1 and AAB59943.1), MLV (AY818896, AAV68488.2, and AAV68489.1), XMRV (JF908815, AEI59722.1, AEI59723.1, and AEI59724.1), XMRV (JF908816, AEI59725.1, AEI59726.1, and AEI59727.1), MLV (AB213653, BAD98608.1, and BAD98609.1), MLV (MLVENVR; AAA46518.1 and AAA46519.1), *M. musculus* (AL606987), *M. caroli* (XM021149499 and XP_021041459.1), *G. surdaster* (XM_028761645 and XP_028617478.1), *M. musculus* (XR_001784239), *M. pahari* (XM_021190222 and XP_021045881.1), *R. norvegicus* (XR_005497950 and XP_038956947.1), *R. norvegicus* (XM_039101019), *M. caroli* (XM_021185800 and XP_021005158.1), *M. musculus* (AC130672), CPERV (UGO47158), KWERV (GQ222416, ACX69256, and ACX69257), PERV-A (KY484771, ASU50141, and ASU50142), PERV-B (AY099324, AAM29194, and AAM29193), PERV-C (HM159246, ADK35877, ADK35878, and ADK35879), *Arvicanthis niloticus* (XM_034491546, XP_034347437, and XP_034367612), McERV (KC460271, AGP25479, AGP25480, and AGP25481), MDEV (AF053745, AAC31804, AAC31805, and AAC31806), GaLV-SF (KT724047, ALV83299, ALV83300, and ALV83301), GaLV-Hall's Island (KT724050, ALV83308, ALV83309, and ALV83310), GaLV-Brain (KT724049, ALV83305, ALV83306, and ALV83307), GaLV-SEATO (KT724048, ALV83302, ALV83303, and ALV83304), KoRV-KV522 (AB721500, BAM67146, and BAM67147), KoRV Pci-SN265 (KF786285, AHY24814, AHY24815, and AHY24816), KoRV Br2-1 CEETG (KC779547, AGO86849, and AGO86848), WMV-SSAV (KT724051, ALV83311, ALV83312, and ALV83313), MmGRV (MN413611, QJT93249, QJT93250, and QJT93251), SaGRV (MN413612, QJT93252, QJT93253, and QJT93254), FFRV1 (MK040728, QDA02049, QDA02050, and QDA02051), HPG (MN413610, QJT93246, QJT93247, and QJT93248), HlGRV (MN413613, QJT93255, QJT93256, and QJT93257), *Cricetulus griseus* ERV (XM_027403845, XP_027259646, and XP_027275435), and UrsusERV (Repbase reports 15 [11] and 3519 [2015]). Reticuloendotheliosis virus was used as an outgroup in the analysis. Prottest-3.4.2 and jmodelTest-2.1.10 were used to determine the best-fitting evolutionary model for the phylogenetic analysis (36, 37). Prottest analysis of all three gene protein alignments indicated that the best-fit model to be used was Jones-Taylor-Thornton (JTT) with gamma distribution. A general time reversible (GTR) model with gamma distribution and invariable sites was used for the nucleotide alignment. Phylogenetic analysis was performed using RAxML maximum likelihood inference program with 20 maximum likelihood searches and 500 rapid bootstrap replicates for both nucleotide and amino acid alignments (38).

Age estimation of SimumERV and DicerosERV. The age estimation for the identified elements was performed using three different approaches. For all three approaches, hypermutable CpG regions were excluded from the analysis, and the previously reported African rhinoceros mutation rate of 0.00088/nucleotide/Myr (24) was used. All the data used in all three age estimation approaches were grouped into different subgroups based on sequence similarity. Consensus sequences were generated based on the resulting alignments of each subgroup (Fig. S5 to S8 in the supplemental material). For the first method, we determined the sequence divergence of identified LTR sequences from the consensus proviral LTR sequence based on a previously described method (5, 39). Estimated divergence was subsequently corrected using the Kimura-2-parameter (K2P) model (22). The calculated sequence divergences from the consensus were then used to estimate the age of SimumERV and DicerosERV LTRs, assuming a molecular clock (Tables S2 to S4). For the second age estimation method, we determined the sequence divergences of all identified gene sequences for SimumERV and DicerosERV elements from their consensus. K2P-corrected divergences were used to estimate SimumERV and DicerosERV gene ages (Tables S5 to S13). The third age estimation approach used the proviral 5′ and 3′ LTR sequence divergence. Proviral 5′ and 3′ LTR sequences are identical after integration into the host genome, and they acquire mutations time independently based on the host mutation rate. Age estimates for each provirus were calculated using $T = D/(2 \times 0.0022)$, where D is the K2P-corrected sequence divergence between 5′ and 3′ proviral LTR sequences (40) (Tables S14 to S16). Mean and standard deviation values were calculated for all three age estimation approaches.

Data availability. All generated sequence data were submitted to the NCBI SRA and can be accessed with the accession number PRJNA862320.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 8.1 MB.

ACKNOWLEDGMENT

The authors wish to thank Karin Hönig for technical support.

REFERENCES

1. Johnson WE. 2015. Endogenous retroviruses in the genomics era. Annu Rev Virol 2:135–159. https://doi.org/10.1146/annurev-virology-100114-054945.
2. Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. Nat Rev Microbiol 17:355–370. https://doi.org/10.1038/s41579-019-0189-2.
3. Tsangaras K, Mayer J, Alquezar-Planas DE, Greenwood AD. 2015. An evolutionarily young polar bear (*Ursus maritimus*) endogenous retrovirus identified from next generation sequence data. Viruses 7:6089–6107. https://doi.org/10.3390/v7112927.
4. Gifford R, Tristem M. 2003. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26:291–315. https://doi.org/10.1023/a:1024455415443.
5. Mayer J, Tsangaras K, Heeger F, Ávila-Arcos M, Stenglein MD, Chen W, Sun W, Mazzoni CJ, Osterrieder N, Greenwood AD. 2013. A novel endogenous betaretrovirus group characterized from polar bears (*Ursus maritimus*) and giant pandas (*Ailuropoda melanoleuca*). Virology 443:1–10. https://doi.org/10.1016/j.virol.2013.05.008.
6. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21:30. https://doi.org/10.1186/s13059-020-1935-5.
7. Steiner CC, Ryder OA. 2011. Molecular phylogeny and evolution of the *Perissodactyla*. Zool J Linn Soc 163:1289–1303. https://doi.org/10.1111/j.1096-3642.2011.00752.x.
8. Zhu H, Gifford RJ, Murcia PR. 2018. Distribution, diversity, and evolution of endogenous retroviruses in perissodactyl genomes. J Virol 92:e00927-18. https://doi.org/10.1128/JVI.00927-18.
9. Li Y, Wang H, Nie K, Zhang C, Zhang Y, Wang J, Niu P, Ma X. 2016. VIP: an integrated pipeline for metagenomics of virus identification and discovery. Sci Rep 6:23774. https://doi.org/10.1038/srep23774.
10. Sperber GO, Airola T, Jern P, Blomberg J. 2007. Automated recognition of retroviral sequences in genomic data—RetroTector. Nucleic Acids Res 35:4964–4976. https://doi.org/10.1093/nar/gkm515.
11. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. 2011. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res 39:D225–D229. https://doi.org/10.1093/nar/gkq1189.
12. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
13. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. 2011. The sequence read archive. Nucleic Acids Res 39:D19–D21. https://doi.org/10.1093/nar/gkq1019.
14. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: their purpose and place. Hum Mol Genet 27:R234–R241. https://doi.org/10.1093/hmg/ddy177.
15. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
16. Etherington GJ, Dicks J, Roberts IN. 2005. Recombination analysis tool (RAT): a program for the high-throughput detection of recombination. Bioinformatics 21:278–281. https://doi.org/10.1093/bioinformatics/bth500.
17. Maydt J, Lengauer T. 2006. Recco: recombination analysis using cost optimization. Bioinformatics 22:1064–1071. https://doi.org/10.1093/bioinformatics/btl057.
18. Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474. https://doi.org/10.1186/1471-2105-7-474.
19. Oliveira NM, Satija H, Kouwenhoven IA, Eiden MV. 2007. Changes in viral protein function that accompany retroviral endogenization. Proc Natl Acad Sci U S A 104:17506–17511. https://doi.org/10.1073/pnas.0704313104.
20. Hayward JA, Tachedjian M, Kohl C, Johnson A, Dearnley M, Jesaveluk B, Langer C, Solymosi PD, Hille G, Nitsche A, Sánchez CA, Werner A, Kontos D, Crameri G, Marsh GA, Baker ML, Poumbourios P, Drummer HE, Holmes EC, Wang LF, Smith I, Tachedjian G. 2020. Infectious KoRV-related retroviruses circulating in Australian bats. Proc Natl Acad Sci U S A 117:9529–9536. https://doi.org/10.1073/pnas.1915400117.
21. Kim S, Rusmevichientong A, Dong B, Remenyi R, Silverman RH, Chow SA. 2010. Fidelity of target site duplication and sequence preference during integration of xenotropic murine leukemia virus-related virus. PLoS One 5:e10255. https://doi.org/10.1371/journal.pone.0010255.
22. Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120. https://doi.org/10.1007/BF01731581.
23. Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol 22:650–658. https://doi.org/10.1093/molbev/msi043.
24. Moodley Y, Westbury MV, Russo IRM, Gopalakrishnan S, Rakotoarivelo A, Olsen RA, Prost S, Tunstall T, Ryder OA, Dalén L, Bruford MW. 2020. Interspecific gene flow and the evolution of specialization in black and white rhinoceros. Mol Biol Evol 37:3105–3117. https://doi.org/10.1093/molbev/msaa148.
25. Zheng J, Wei Y, Han GZ. 2022. The diversity and evolution of retroviruses: perspectives from viral "fossils". Virol Sin 37:11–18. https://doi.org/10.1016/j.virs.2022.01.019.
26. Kijima TE, Innan H. 2010. On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol 27:896–904. https://doi.org/10.1093/molbev/msp295.
27. Jedlicka P, Lexa M, Kejnovsky E. 2020. What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? Front Plant Sci 11:644. https://doi.org/10.3389/fpls.2020.00644.
28. Zheng J, Wang J, Gong Z, Han GZ. 2021. Molecular fossils illuminate the evolution of retroviruses following a macroevolutionary transition from land to water. PLoS Pathog 17:e1009730. https://doi.org/10.1371/journal.ppat.1009730.
29. Ishida Y, Zhao K, Greenwood AD, Roca AL. 2015. Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. Mol Biol Evol 32:109–120. https://doi.org/10.1093/molbev/msu275.
30. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, DiCuccio M, Murphy TD, Pruitt KD, Kimchi A. 2016. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res 44:D73–D80. https://doi.org/10.1093/nar/gkv1226.
31. Liu S, Westbury MV, Dussex N, Mitchell KJ, Sinding MHS, Heintzman PD, Duchêne DA, Kapp JD, von Seth J, Heiniger H, Sánchez-Barreiro F, Margaryan A, André-Olsen R, De Cahsan B, Meng G, Yang C, Chen L, van der Valk T, Moodley Y, Rookmaaker K, Bruford MW, Ryder O, Steiner C, Bruins-van Sonsbeek LGR, Vartanyan S, Guo C, Cooper A, Kosintsev P, Kirillova I, Lister AM, Marques-Bonet T, Gopalakrishnan S, Dunn RR, Lorenzen ED,

Shapiro B, Zhang G, Antoine PO, Dalén L, Gilbert MTP. 2021. Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. Cell 184:4874–4885. https://doi.org/10.1016/j.cell.2021.07.032.

32. Hayward A, Grabherr M, Jern P. 2013. Broad-scale phylogenomics provides insights into retrovirus—host evolution. Proc Natl Acad Sci U S A 110:20146–20151. https://doi.org/10.1073/pnas.1315419110.
33. Dayaram A, Tsangaras K, Pavulraj S, Azab W, Groenke N, Wibbelt G, Sicks F, Osterrieder N, Greenwood AD. 2018. Novel divergent polar bear-associated mastadenovirus recovered from a deceased juvenile polar bear. mSphere 3:e00171-18. https://doi.org/10.1128/mSphere.00171-18.
34. Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010:pdb.prot5448. https://doi.org/10.1101/pdb.prot5448.
35. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. https://doi.org/10.1101/gr.074492.107.
36. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and high-performance computing Europe PMC Funders Group. Nat Methods 9:772. https://doi.org/10.1038/nmeth.2109.
37. Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165. https://doi.org/10.1093/bioinformatics/btr088.
38. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
39. Kapitonov V, Jurka J. 1996. The age of Alu subfamilies. J Mol Evol 42:59–65. https://doi.org/10.1007/BF00163212.
40. Dangel AW, Baker BJ, Mendoza AR, Yu CY. 1995. Complement component *C4* gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. Immunogenetics 42:41–52. https://doi.org/10.1007/BF00164986.