PNAS

Supplementary Information for

A Recent Gibbon Ape Leukemia Virus Germline Integration in a Rodent from New Guinea

Saba Mottaghinia [a,b], Saskia Stenzel [c,d], Kyriakos Tsangaras [e], Nikolas Nikolaidis [f], Michael Laue [g], Karin Müller [a], Henriette Hölscher [a], Ulrike Löber [h], Gayle K. McEwen [a], Stephen C. Donnellan [i], Kevin C. Rowe [j], Ken P. Aplin [i,†], Christine Goffinet [c,d] and Alex D. Greenwood [a,k,*]

[a] Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Str. 17, 10315 Berlin, Germany; [b] Centre International de Recherche en Infectiologie, Université de Lyon, Inserm, U1111, Université Claude Bernard Lyon 1, CNRS, UMR5308, École Nationale Supérieure de Lyon, Lyon, France; [c] Institute of Virology, Charité – Universitätsmedizin Berlin, D-10117 Berlin, Germany; [d] Department of Tropical Disease Biology, Liverpool School of Tropical Medicine, Liverpool L3 5QA, United Kingdom; [e] Department of Life and Health Sciences, University of Nicosia, 46 Makedonitissas Avenue, CY-2417 Nicosia, Cyprus; [f] Department of Biological Science, Center for Applied Biotechnology Studies, and Center for Computational and Applied Mathematics, College of Natural Sciences and Mathematics, California State University Fullerton, Fullerton, CA 92834-6850, USA; [g] Advanced Light and Electron Microscopy (ZBS 4), Centre for Biological Threats and Special Pathogens, Robert Koch Institute, D-13353 Berlin, Germany; [h] Max-Delbrük Center for Molecular Medicine in the Helmholtz Association (MDC), 13125 Berlin, Germany; [i] South Australian Museum, North Terrace, Adelaide SA 5000, Australia; [j] Sciences Department, Museums Victoria, Melbourne, VIC 3001, Australia; [k] School of Veterinary Medicine, Freie Universität Berlin, Robert-von Ostertag-Str. 7-13, 14163 Berlin, Germany.

[†] Deceased

* Corresponding author

Email: greenwood@izw-berlin.de

This PDF file includes:

SI Methods

Samples and DNA extraction

A total of 278 rodent (*n* = 122) and bat (*n* = 156) samples from the South Australian Museum (SAM) were analyzed. These samples were collected between 1981 and 2017, and represent seven bat families (Emballonuridae, Hipposideridae, Miniopteridae, Molossidae, Pteropodidae, Rhinolophidae and Vespertilionidae) from 37 genera and ca. 120 species, three rodent genera (*Hydromys*, *Melomys* and *Rattus*) from the family Muridae, representing ca. 38 species with six of them (*R. argentiventer*, *R. exulans*, *R. nitidus*, *R. norvegicus*, *R. rattus* and *R. tanezumi*) found on both sides of the Wallace Line (Fig. 1 and Dataset S1) (1). DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen, Germany) according to the manufacturer's protocol for frozen or ethanol-preserved blood, hair and tissue samples.

GALV-KoRV PCR screening

Degenerate primer set KOGAWM-1F 5'-CCCCTYAATCGACCTCASTGG-3' and KOGAWM-1R 5'-RTATCTCCTATARGCCTCCAT-3' (product size ~200 bp) were designed using Geneious R9.1 (https://www.geneious.com) and synthesized (Sigma-Aldrich, Germany) to amplify part of the *gag* gene (from 1,945 to 2,145 bp) of aligned GALV (KT724048), KoRV (AB721500) and MelWMV (KX059700), using a touchdown Polymerase-Chain-Reaction (PCR): (i) 94°C for 15 min; (ii) 35 cycles consisting of 94°C for 30 s, 70°C (-0.5°C/cycle) for 40 s, 72°C for 1 min; and (iii) 72°C for 6 min. Reactions consisted of 12.5 µl 2x MyFi™ Mix (Bioline, Australia), 3 µl (10 mM) of primer set, 1.5 µl of template and the added water to volume 25 µl. 4 µl of PCR products, including KoRV positive controls from koala spleen DNA were mixed with 1 µl of DNA loading buffer red (Bioline, Australia) and were visualized on 1.5% w/v agarose gel stained with GelGreen® Nucleic Acid Gel Stain (Biotium, USA). To clean-up the amplified PCR products for sequencing, the volume was adjusted to 100 µl with 1xTE buffer and transferred to 384-well multiscreen PCR plates for vacuum-drying the wells. Dried DNA was re-suspended in a 20 µl 1xTE buffer and sent to Australian Genome Research Facility (AGRF, Australia) for Sanger sequencing. BLASTn (2) was used to confirm that the sequences were related to GALV or KoRV.

Illumina library construction

Genomic DNAs were quantified using a Quantus Fluorometer (Promega, USA) and fragmented to an average size of 250 bp with a Bioruptor® Pico sonication device (Diagenode, Belgium) for 15 sec at 7 cycles followed by 90 sec of rest. The size distribution and molarities were measured with an Agilent 2200 TapeStation, using D1000 ScreenTape and reagents (Agilent Technologies, USA). Illumina sequencing libraries were generated for 9 fragmented DNA samples and one control (blank) according to Meyer and Kircher (3) with the modifications of Alfano et al. (4). Each library was ligated to a unique combination of

P5-P7 oligonucleotide index adapters (5) and amplified for 7 cycles with the same cycling conditions described in Alfano et al. (6). 0.8x Agencourt AMPure XP beads (Beckman Coulter, USA) were used to clean the libraries, by binding and eluting with 1.2x AMPure beads. Agilent high-sensitivity tapes and reagents were used to check molarity and fragment size of the libraries.

Target enrichment hybridization capture and sequencing

The curated 70-mer biotinylated oligonucleotide meta-viral-baits (probes) list which was previously modified by Alfano et al. (7), was used with further customization. This bait set was based on the retrieved viral oligonucleotides in the generation 5 of the ViroChip (Viro5) (8), available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13323. Further 437 baits tiles as 70-mer oligonucleotides were added to cover full genomes of GALV (KT724048), KoRV (AB721500), MelWMV (KX059700) and WMV (KT724051). The final capture panel consisted of 13,735 unique sequences that can tolerate ~5% sequence divergence from the target without reducing the capture efficiency. The capture panel was synthesized (ArborBiosciences, USA).

To ensure balanced baits were hybridized to each library the indexed libraries were pooled into 3 groups of high (samples 89, 249 and 300), medium (samples 246, 290, 291 and 292) and low (samples 201, 204 and the control) molar concentrations. The pooled libraries were hybridized with customized 70-mer biotinylated oligonucleotide meta-viral-baits (ArborBiosciences, USA) at 61°C for 30 hours following myBaits®-Hybridization Capture for Targeted NGS-manual version 4.01. The captured libraries were measured on an Agilent 2200TapeStation and re-amplified for 22 cycles with the same cycling conditions as in Alfano et al., 2016 with exception of a capture temperature to 61°C. The re-amplified enriched libraries were purified once more with Agencourt AMPure XP beads, pooled to equimolar amounts with a final library concentration of 21.8 nM and 300 bp paired-end sequencing on the Illumina MiSeq platform with v2 reagent kit. The gammaretroviruses enrichment factor was calculated based on the following formula:

$$\text{Enrichment efficiency} = \frac{\text{number of reads mapped to the target region}}{\text{total number of reads}}$$

Bioinformatics analysis and virus classifications

The raw sequencing reads were demultiplexed, adaptor sequences, low-quality reads (quality cutoff 20 and minimum read length of 30 nt) and duplicates removed and merged using Cutadapt v1.15 (9), Trimmomatic v0.27 (10), Picard v1.4 (http://broadinstitute.github.io/picard), and BBMerge (11) respectively. Two pipelines were applied for identification and assembly of viral reads (Fig. S1): Virus

Integrated Pipeline (VIP) (12) in sense mode and Genome Detective (13), a web-based bioinformatics pipeline.

VIP uses Bowtie2 (14) to remove background reads by searching in a human nucleotide database (human DB) which is constructed from a combination of human genomic DNA (GRCh38/hg38), RefSeq (rRNA, RNA and mtDNA) and GOTTCHA bacterial database. Due to the high copy number of ERVs in rodents which are the host species here, the reference database was not modified from default human DB to maintain a modest homology cutoff without losing too many target sequences from the distantly related ERVs. VIP then tries to identify the remaining reads from NCBI RefSeq (viral genomic DNA/RNA and their protein products) and NCBI GenBank viral neighbor genomes, sorting them on genus level into separate bins. Velvet (15) with various k-mer lengths *de novo* assembled each bin as contigs. The best match for the *gammaretrovirus* bin was baboon ERV (NC_022517) for samples 89 and 246 with 76.19% and 94.44% nucleotide identity, respectively. For 249 (74.53%), 290 (81.50%), and 291 (84.16%) was WMV (KT724051), 204 (77.71%) was MelWMV (KX059700) and for 201 (76%), 292 (86.69%) and 300 (74.27%) was KoRV (AB721500).

In the second approach, reads were assembled as contigs via Genome Detective. This workflow employs DIAMOND (16), a protein based alignment method to search the Swissprot UniRef90 database, and sorts viral reads into bins without the lowest common ancestor (LCA) algorithm. These viral bins are *de novo* assembled with metaSPAdes (17) then BLASTn and BLASTx are used to search for virus genotyping against the NCBI RefSeq virus database.

To make a homozygous consensus sequence for each sample, gammaretroviral contigs from the two pipelines imported into Geneious Prime 2020. For detecting and selectively removing redundancy, consensus calling was selected with a 75% threshold. The viral consensus were translated to amino acid sequences and aligned to the annotated KoRV-A and WMV protein sequences in Geneious Prime 2020.1.2. The consensus sequences of the cMWMV cluster (204, 290, 291, 292 and 300) were aligned by Geneious mapper (medium sensitivity/fast and iterate up to 5 times) to 249 consensus sequence as the most complete representative. The resulting multiple sequence alignments were manually curated towards the 3' termini where mis-incorporations tend to cluster (18). Based on majority consensus sequence, the yielded 8459 bp was used for virus synthesis and subsequent infection assays (Text file S2).

Variable regions A and B as the determinant for receptor specificity were examined for cMWMV. The envelope sequences of cMWMV, GALVs and the recently identified bat gammaretroviruses were extracted and aligned by MAFFT v. 7.017 (19). Sequence alignment visualized and annotated with Jalview v2.11.1.7 (20).

Phylogenetic analysis

Thirty seven genome sequences of gammaretroviruses (Table S2) from GenBank (NCBI-GenBank Flat File Release 240) were retrieved and manually curated to include the exogenous bat gammaretroviruses (21, 22). Multiple nucleotide alignments of the consensus sequences with the curated database were performed using default settings in MUSCLE (23). Statistical selection of the best-fit model for the phylogenetic analysis performed using jModelTest (24). Phylogenetic relationships were depicted based on the nucleotide alignments of 46 full genomes, using reticuloendotheliosis virus (REV strain SDAUR-S1) (MF185397) as an outgroup. Bayesian phylogenetic inference produced using Markov Chain Monte Carlo (MCMC) for 1,000,000 iterations in MrBayes v3.2.7 (25). A Maximum likelihood (ML) tree was constructed with rapid bootstrapping (1000 replicates) and GTRGAMMA substitution rate in Randomized Axelerated Maximum Likelihood (RAxML v8.2.11) (25). The polytomy in the internal-node of cMWMV clade indicates an erroneous alignment (soft) or simultaneous divergence of several lineages (hard), which by definition cannot be resolved (26). To distinguish between the two notions, it has been proposed to expand the relationship to independent gene trees, testing whether a bifurcating relationship can be obtained. To validate the phylodynamic of the tree, the alignment ambiguities were removed with Gblocks (27) allowing a combination of three filtering parameters; (i) smaller final blocks, (ii) gap positions within the final blocks and (iii) less strict flanking positions. Further, phylogenetic trees were constructed independently for *gag*, *pol*, *env* genes and protein sequences, using RAxML v8.2.11.

Mapping retroviral integration sites

For a virus to become endogenous, a copy of the provirus is integrated at exactly the same specific site in all cells of the host genome. Due to limitations in sample quantity, only one tissue sample per individual animal was possible. Therefore, to determine if the novel retroviral sequences were endogenous, the integration sites among individuals were identified by aligning the merged sequencing reads to WMV in Geneious mapper (default settings). If endogenous, we would expect some or all of the integration sites to be identical among individuals. The target-site duplication (TSD)- which is formed during retroviral integration into the host genome- of 5 bp (ATAAT) flanking LTR on either side of one sample was identified by manual search. The 5' and 3' flanking sequences could be aligned in all *M. leucogaster* (249, 290, 291, 292 and 300) (Text file S1). The flanking sequences were confirmed as host genomic sequences by BLASTn search, matching the *Melomys* sequences to homologous rat genome sequences. The flanking sequences were extracted and aligned to display the shared integration site. No flanking sequence was identified for 89, 201, 204 and 246.

Retroviral protein structure modeling

The structure characteristics of the cMWMV viral genome were examined in comparison to the WMV genome. SWISS-MODEL server (28) was used for the prediction of the three-dimensional (3D) structures

of WMV and cMWMV (249, a representative sequence with high sequence coverage). From the output structures predicted, only high quality protein models as defined by QMEAN4 (29) values were considered for further analysis. Both WMV and cMWMV genomes produced high quality structures in various domains for all three viral polypeptides (GAG, POL and ENV). Pairwise structural alignment, superimposition, and figure design were performed using PyMol v2.4 (30, 31).

To predict the functional effect of non-synonymous mutations, five major criteria were used (32–34). The assumption was that if a mutation is predicted by the majority of these criteria (at least three) then it would be an ideal candidate for functional validation. First, the mutations were categorized based on whether they change an amino acid of known function, because a mutation on a site of established function would most probably have a functional impact. This analysis was performed by collecting known motifs and amino acid positions of known functions from the literature and the Conserved Domain Database (CDD) of NCBI, and comparing them with the collected mutations. Second, the mutations were categorized based on whether they occur in a highly conserved amino acid position by determining the amino acid conservation level of each position. It was assumed that highly conserved amino acids would have a higher probability of causing a functional change. This analysis was performed by using BLASTp and determining the conservation level of each position that carried a particular mutation for the first 100 unique hits. Third, the identified mutations were classified based on whether the amino acid change was predicted to be radical (different amino acid class; negative or zero scores in both BLOSUM65 and BLOSUM80 substitution matrices). The rationale of the latter criteria relies on the fact that radical changes may alter the function with a higher probability than non-radical amino acid changes. Fourth, the mutations were categorized based on whether a particular mutation is predicted to alter the local conformation or the molecule surface by generating 3D models of the wild-type (WT) and mutated proteins. This step was performed by generating 3D structures of the mutated and non-mutated versions of the proteins, and determining perturbations in the local 3D and topography. Lastly, the mutations were categorized based on the outputs of SIFT (scale-invariant feature transform) (35), SNAP (screening for non-acceptable polymorphisms) (36) and PROVEAN (protein variation effect analyzer) (37).

Immunofluorescence staining and microscopy of cells

The crucial region of PiT-1 receptor that allows for GALV infection in a variety of mammals, including humans, gibbons, koalas and the flying fox, is quite divergent from the same region of the *M. musculus* and *M. terricolor* (previously known as *M. dunni*) proteins, granting NIH3T3 and MDTF (*M. dunni* tail fibroblasts) cells a natural resistance to GALVs infection (22, 38, 39). To determine if tropism of cMWMV is comparable to GALVs or, similar to WMV, is restricted to PiT-1, HEK293T and NIH3T3 cells were stained for PiT-1 and PiT-2 proteins. HEK293T and NIH3T3 cells were seeded at $9x10^4$ cells/ml density and grown on uncoated µ-slide 8-well high glass bottom slide (Ibidi, Germany). At ~70% confluency, the medium was removed and the following steps performed at room temperature; 2x brief wash with

Dulbecco's Phosphate-Buffered Saline (DPBS; Biowest, Nuaillé, France), 30 min fixation with 4% fresh paraformaldehyde solution, 1 h blocking and permeabilizing with 1% BSA, 0.6% Triton X-100 in DPBS followed by 3x 5 min DPBS wash. For a direct immunofluorescence staining, cells were incubated for 4 h with Santa Cruz Biotechnology (Dallas, USA) anti-PiT-1 (sc-393943 AE546) or anti-PiT-2 (sc-377326 AE546) primary antibodies conjugated with AlexaFluor®546 (1:50) and 10 min with membrane permeable Hoechst 33342 (Thermo Fisher Scientific, USA) as nuclear counterstain (1:40), followed by 3x 5 min DPBS wash. Slides were kept in 100 µl DPBS. Microscopy was performed on the same day using an inverted Olympus confocal laser scanning microscope IX-81 (40x objective) and the related software FluoView1000 (Olympus, Tokyo, Japan). Alexa-546 was recorded in the red channel (emission band pass 560-660 nm) after excitation with a HE/Ne-laser at 543 nm. Nuclear staining was recorded in the blue channel (emission band pass 430-470 nm) after excitation with a 405 nm laser diode.

Consensus sequences and viral sequence synthesis

In Geneious R9.1, the individual consensus sequences of cMWMV and the *M. burtoni* sequence from 204 were aligned to the near complete 249 sequences as the reference genome, using medium sensitivity/fast and iterate up to 5 times. The resulting alignments were manually curated towards the 3' termini where mis-incorporations tend to cluster (18). While, minor sequence variation was observed among the consensus sequences used to generate the overall consensus sequence, none altered any amino acids and therefore, the functional results resulting from the use of the overall consensus sequence should be representative for the individual isolates. The 8459 bp cMWMV (Text file S2) and KoRV-A (AB721500) genomes were chemically synthesized and sub-cloned in pUC57 vector (GenScript, China). These constructs were used to transfect NIH Swiss mouse embryonic fibroblasts (NIH3T3) and Human embryonic kidney (HEK293T) cells.

Cell cultures

HEK293T and NIH3T3 cells (ATCC) were maintained in Dulbecco's modified Eagle's medium (DMEM) (Thermo Fisher Scientific, USA), supplemented with 10% fetal bovine serum (FBS), 1% L-glutamine, 1% antibiotics penicillin, and 1% streptomycin. Cells were cultured at 37°C, 5% CO_2.

Transfection of cMWMV and KoRV-A proviral DNAs

KoRV-A and cMWMV stocks were produced through transient transfection of HEK293T cells. Cells were transfected with pUC57 plasmid vector encoding cMWMV and KoRV-A genomes by calcium-phosphate precipitation using CalPhos Mammalian Transfection kit (Takara, Japan). Medium was changed at 16 hours post transfection (hpi). Virus-containing supernatant was harvested at 40 and 64 hpi and sterile-filtered using a filter with pore sizes of 0.45 µm. The supernatant was ultracentrifuged through a

20% sucrose/PBS solution at 153,400 x g at 4°C for 90 minutes. Virus-containing pellets were resuspended in medium and aliquots were stored at -80°C.

Infection

HEK293T and NIH3T3 cells were seeded at a cell density of 2 x 10^5 cells/ml and 1.6 x 10^5 cells/ml, respectively, in a 48 well plate. Infection of cells was performed with different volumes of concentrated virus-containing supernatant (1 µl, 10 µl and 100 µl) overnight at 37°C with 5% CO_2. Upon 24 h infection, medium was changed and viral supernatant harvested at 24, 48, 72 and 96 hpi. Lastly, 150 µl virus-containing supernatant was mixed with 600 µl RAV1 buffer and stored at -80°C for subsequent viral RNA extraction.

Viral RNA extraction and Taqman RT-qPCR

The PrimerQuestTool from Integrated DNA Technologies (https://www.idtdna.com/) was used to design fluorescent primers and probe on *pol* gene of cMWMV and *env* gene of KoRV-A. TaqMan primers and probes with 5'-6-FAM and 3'-BBQ650 modifications were synthesized (Biomers, Germany). Viral RNA was extracted using NucleoSpin RNA Virus kit (Macherey-Nagel, Germany). The complementary DNA (cDNA) constructed using dNTPs (Thermo Fisher Scientific, USA), random hexamers (Jena Bioscience, Germany) and Moloney Murine Leukemia Virus (M-MLV, MMLV) Reverse Transcriptase (New England Biolabs) with buffer. Quantification of absolute viral copies was performed with the LightCycler 480 II system (Roche, Germany) in technical duplicates using Taq-Man PCR technology. Viral replications of cMWMV and KoRV-A were assessed by quantifying viral copies using primer set cMWMV_2F 5´-GATCCATGCTTCTCACCTCAA-3´, cMWMV_2R 5´-CGAATACGCAGCTTAAGAGGAT-3´ and specific probe cMWMV_2P with 5´-CAGATGAGTCCTGGGAGCTGGAAA-3´ sequence and product size 107 bp, K_env_F 5´-GAGTCCTGGGAACTGGAAAAG-3´, K_env_R 5´-TAGTGGGGCTATTCCTTTTA-3´ and specific probe K_env_P 5´-TCCTCTTAAGTTGCGTGTTCGGCG-3´ and product size 95 bp. DNA concentrations calculated using standards of defined DNA concentrations, consisting of plasmid dilutions that contained a defined plasmid copy number.

Thin section electron microscopy (EM)

HEK293T and NIH3T3 cells were seeded in culture-insert 2 well 35 mm µ-Dish (Ibidi, Germany) and infected with cMWMV and KoRV-A as described above. At 48 hpi, the medium was discarded, the virus-infected cells were fixed with 2.5% glutaraldehyde in 0.05 M Hepes buffer (pH:7.2), and then incubated at room temperature for 2h. Afterwards, µ-Dishes were filled with the fixative buffer and cells were embedded in the chambers by using Epon resin (protocol with tannic acid and uranyl acetate block contrasting; (40). Thin sections (60-70 nm thick) were produced with an ultramicrotome, contrasted with uranyl acetate and lead citrate and investigated with a transmission electron microscope (JEM-2100,

Jeol) operated at 200 kV. Images were recorded using a side-mounted CCD camera (Veleta, EMSIS) with 2048x2048 pixels.

PiT-1 sequencing from NIH3T3 cells

RNA was isolated from NIH3T3 cells using the NucleoSpin® RNA Plus kit for DNA, RNA and protein purification using the RNA isolation protocol (Macherey-Nagel, Düren, Germany). The isolated RNA was reverse transcribed into cDNA using SuperScript™ IV (Thermo Scientific). For PCR, each approach contained a final volume of 25 µl using MyTaq™Mix (Bioline, BioCat, Heidelberg, Germany). Primer combinations used were; MusPit1A_F1 5‘-CAGTGGCTCTCTTGAAGAATGGT-3‘ and MusPit1A_R1 5‘-AATACTGAAACCACTGGAGGGTG-3‘, MusPit1A_F2 5‘-CCTGCTTTGGGTCATTTGCC-3‘ and MusPit1A_R2 5’-CACAACAGAGCCCACTTTGC-3‘, VRA_F3 5‘-GCTTTCTGGTATTATGTCTGG-3‘ and VRA_R3 5‘-GTTTGACTGAACTGAACAAGG-3‘, MusPit1B_F1 5‘-CCTTGTTCGTGCGTTCATCC-3‘ and MusPit1B_R1 5‘-ATCCTTGTGCACGGTGTGAT-3‘, MusPit1B_F2 5‘-TTGTTCGTGCGTTCATCCTC-3‘ and MusPit1B_R2 5‘-TGGCACACACTACCTCAGAC-3‘, VRB_F3 5‘-TGGTATGACCAGGATAAGCC-3‘ and VRB_R3 5’-TGTTTCGAAACAGTCGCCAG-3‘ were used. Thermal cycling conditions for PCR were as follows: 94°C for 4 min, followed by 35 cycles 94°C for 30 sec, 54°C for 30 sec, 72° for 1 min and 72°C for 5 min. Products were directly Sanger sequenced (LGC Genomics) and aligned to the *Mus musculus* GenBank entry for PiT-1 (Fig. S8).

DNA sonication inverse PCR to identify cMWMV and KoRV-A integration sites in HEK293T and NIH3T3 cell line infection experiments

Sonication based inverse PCR was conducted as previously described in Löber et al. (2018) (41), Alquezar-Planas et al. (2021) (42) and McEwen et al. (2021) (43) with the following modifications: a total of 240 ng of blunt ended DNA was added for circularization. The blunt ended DNA was diluted to 40 ng/µl. To avoid a loss of product, no purification was performed after the circularization. The undiluted circularized DNA was directly used as template for inverse PCR. PCR was performed under the following thermal cycling conditions: 95°C for 1 min, 30 sec for denaturation, followed by 40 cycles at 95°C for 20 sec, 57°C for 30 sec and 72°C 3 min. The primers used for amplification of cMWMV were primers cMWMV_1 5’-ATTTGCATCCGAAGCCGTGG-3’ and cMWMV_2 5’-GGGGCACCCTGGAAACTGC-3’. The primers for KoRV-A are described in Alquezar-Planas et al. 2021 (42). The samples were PacBio sequenced by the DSMZ, Braunschweig, Germany. Identification of the integration sites was done as described in Löber et al. (41) and Alquezar-Planas et al. (42). Integration site locations compared to the human (HEK293T) or mouse (NIH3T3) genomes are shown in *SI Appendix* Dataset S2.

Supplementary Figures

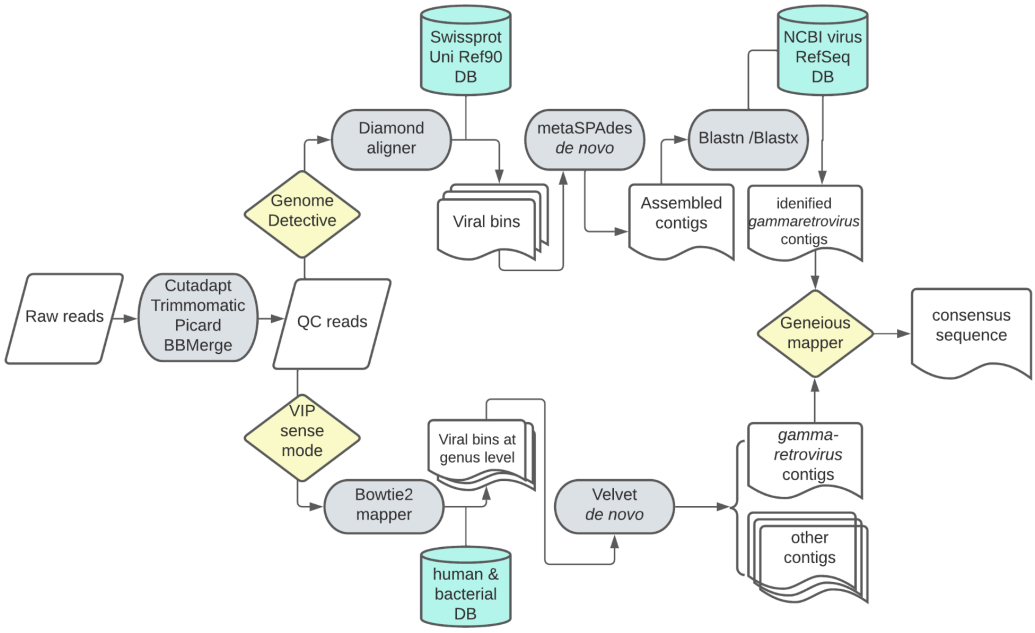
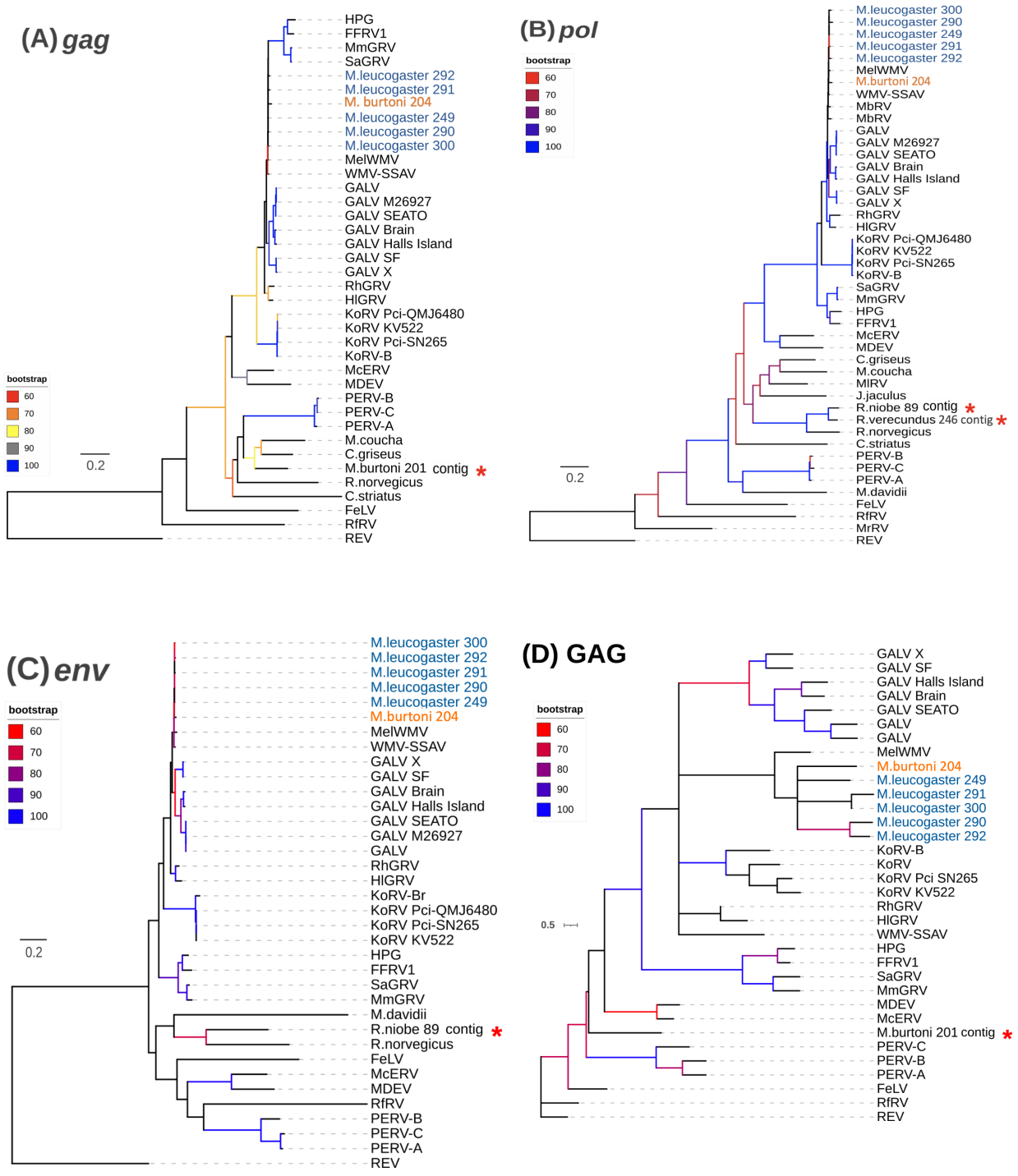
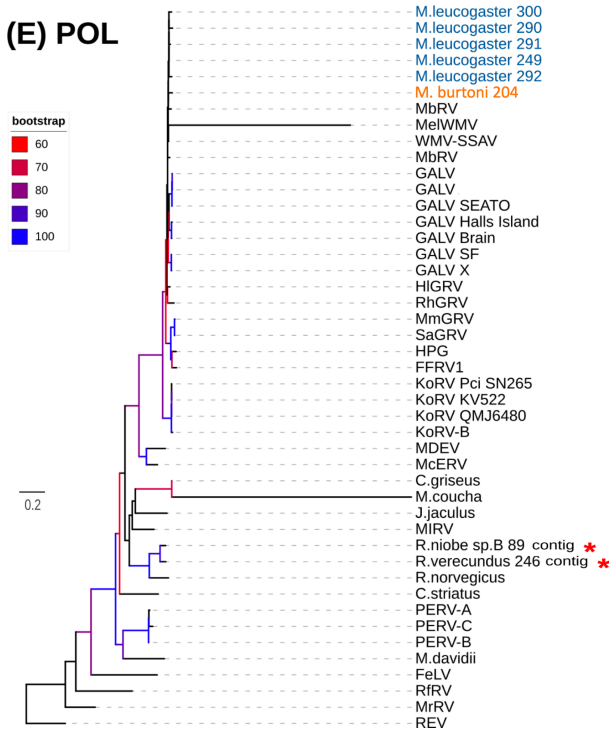
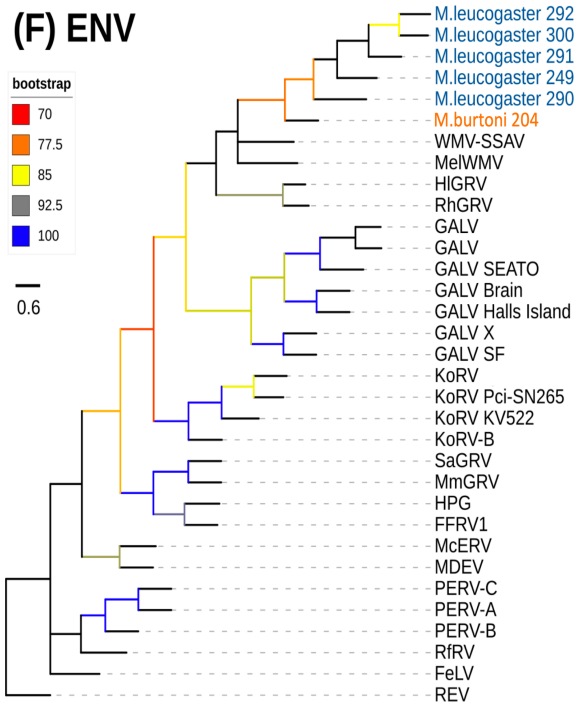


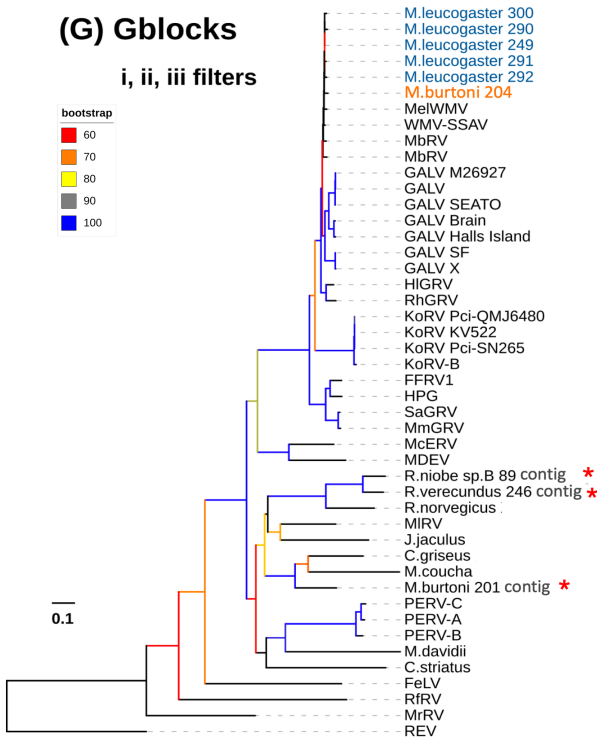
Fig. S1

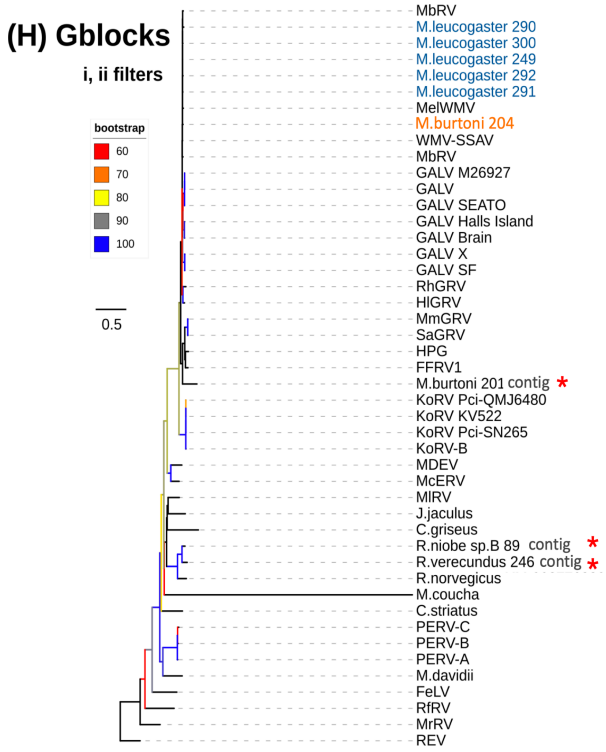
Overview of the workflow applied for identification and classification of viral reads and constructing the consensus sequences shows the curated reads were passed through Genome Detective and Virus Integrated Pipeline (VIP) in parallel. These pipelines use different databases and mappers to search for viral sequences. They also use different *de novo* algorithms for assembling the contigs. The consensus sequence was built in Geneious R.9 from these contigs. This figure is created using Lucidchart, www.lucidchart.com.


(A) gag
HPG
FFRV1
MmGRV
SaGRV
M.leucogaster 292
M.leucogaster 291
M. burtoni 204
M.leucogaster 249
M.leucogaster 290
M.leucogaster 300
MelWMV
WMV-SSAV
GALV
GALV M26927
GALV SEATO
GALV Brain
GALV Halls Island
GALV SF
GALV X
RhGRV
HlGRV
KoRV Pci-QMJ6480
KoRV KV522
KoRV Pci-SN265
KoRV-B
McERV
MDEV
PERV-B
PERV-C
PERV-A
M.coucha
C.griseus
M.burtoni 201 contig *
R.norvegicus
C.striatus
FeLV
RfRV
REV
bootstrap
60
70
80
90
100
0.2
(B) pol
M.leucogaster 300
M.leucogaster 290
M.leucogaster 249
M.leucogaster 291
M.leucogaster 292
MelWMV
M.burtoni 204
WMV-SSAV
MbRV
MbRV
GALV
GALV M26927
GALV SEATO
GALV Brain
GALV Halls Island
GALV SF
GALV X
RhGRV
HlGRV
KoRV Pci-QMJ6480
KoRV KV522
KoRV Pci-SN265
KoRV-B
SaGRV
MmGRV
HPG
FFRV1
McERV
MDEV
C.griseus
M.coucha
MlRV
J.jaculus
R.niobe 89 contig *
R.verecundus 246 contig *
R.norvegicus
C.striatus
PERV-B
PERV-C
PERV-A
M.davidii
FeLV
RfRV
MrRV
REV
bootstrap
60
70
80
90
100
0.2
(C) env
bootstrap
60
70
80
90
100
0.2
M.leucogaster 300
M.leucogaster 292
M.leucogaster 291
M.leucogaster 290
M.leucogaster 249
M.burtoni 204
MelWMV
WMV-SSAV
GALV X
GALV SF
GALV Brain
GALV Halls Island
GALV SEATO
GALV M26927
GALV
RhGRV
HlGRV
KoRV-Br
KoRV Pci-QMJ6480
KoRV Pci-SN265
KoRV KV522
HPG
FFRV1
SaGRV
MmGRV
M.davidii
R.niobe 89 contig *
R.norvegicus
FeLV
McERV
MDEV
RfRV
PERV-B
PERV-C
PERV-A
REV
(D) GAG
bootstrap
60
70
80
90
100
0.5
GALV X
GALV SF
GALV Halls Island
GALV Brain
GALV SEATO
GALV
GALV
MelWMV
M.burtoni 204
M.leucogaster 249
M.leucogaster 291
M.leucogaster 300
M.leucogaster 290
M.leucogaster 292
KoRV-B
KoRV
KoRV Pci SN265
KoRV KV522
RhGRV
HlGRV
WMV-SSAV
HPG
FFRV1
SaGRV
MmGRV
MDEV
McERV
M.burtoni 201 contig *
PERV-C
PERV-B
PERV-A
FeLV
RfRV
REV


(E) POL
bootstrap
60
70
80
90
100
0.2
M.leucogaster 300
M.leucogaster 290
M.leucogaster 291
M.leucogaster 249
M.leucogaster 292
M. burtoni 204
MbRV
MelWMV
WMV-SSAV
MbRV
GALV
GALV
GALV SEATO
GALV Halls Island
GALV Brain
GALV SF
GALV X
HlGRV
RhGRV
MmGRV
SaGRV
HPG
FFRV1
KoRV Pci SN265
KoRV KV522
KoRV QMJ6480
KoRV-B
MDEV
McERV
C.griseus
M.coucha
J.jaculus
MlRV
R.niobe sp.B 89 contig *
R.verecundus 246 contig *
R.norvegicus
C.striatus
PERV-A
PERV-C
PERV-B
M.davidii
FeLV
RfRV
MrRV
REV


(F) ENV
bootstrap
70
77.5
85
92.5
100
0.6
M.leucogaster 292
M.leucogaster 300
M.leucogaster 291
M.leucogaster 249
M.leucogaster 290
M.burtoni 204
WMV-SSAV
MelWMV
HlGRV
RhGRV
GALV
GALV
GALV SEATO
GALV Brain
GALV Halls Island
GALV X
GALV SF
KoRV
KoRV Pci-SN265
KoRV KV522
KoRV-B
SaGRV
MmGRV
HPG
FFRV1
McERV
MDEV
PERV-C
PERV-A
PERV-B
RfRV
FeLV
REV


(G) Gblocks
i, ii, iii filters
bootstrap
60
70
80
90
100
0.1
M.leucogaster 300
M.leucogaster 290
M.leucogaster 249
M.leucogaster 291
M.leucogaster 292
M.burtoni 204
MelWMV
WMV-SSAV
MbRV
MbRV
GALV M26927
GALV
GALV SEATO
GALV Brain
GALV Halls Island
GALV SF
GALV X
HlGRV
RhGRV
KoRV Pci-QMJ6480
KoRV KV522
KoRV Pci-SN265
KoRV-B
FFRV1
HPG
SaGRV
MmGRV
McERV
MDEV
R.niobe sp.B 89 contig *
R.verecundus 246 contig *
R.norvegicus
MlRV
J.jaculus
C.griseus
M.coucha
M.burtoni 201 contig *
PERV-C
PERV-A
PERV-B
M.davidii
C.striatus
FeLV
RfRV
MrRV
REV


(H) Gblocks
i, ii filters
bootstrap
60
70
80
90
100
0.5
MbRV
M.leucogaster 290
M.leucogaster 300
M.leucogaster 249
M.leucogaster 292
M.leucogaster 291
MelWMV
M.burtoni 204
WMV-SSAV
MbRV
GALV M26927
GALV
GALV SEATO
GALV Halls Island
GALV Brain
GALV X
GALV SF
RhGRV
HlGRV
MmGRV
SaGRV
HPG
FFRV1
M.burtoni 201 contig *
KoRV Pci-QMJ6480
KoRV KV522
KoRV Pci-SN265
KoRV-B
MDEV
McERV
MlRV
J.jaculus
C.griseus
R.niobe sp.B 89 contig *
R.verecundus 246 contig *
R.norvegicus
M.coucha
C.striatus
PERV-C
PERV-B
PERV-A
M.davidii
FeLV
RfRV
MrRV
REV

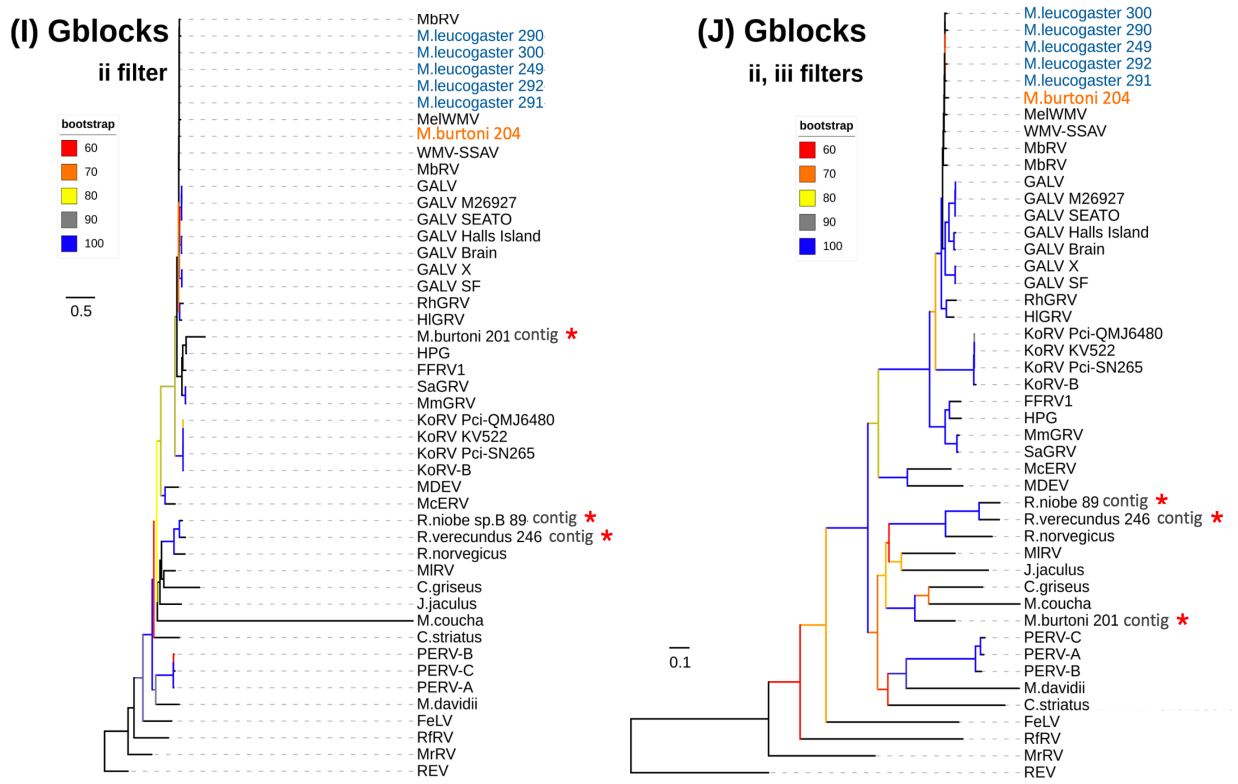
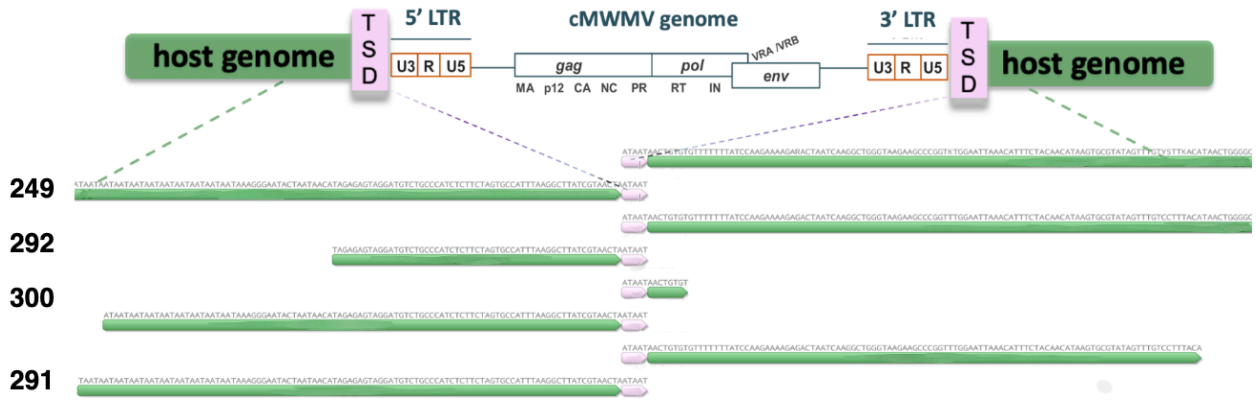


Fig. S2

The maximum likelihood phylogenetic relationship of cMWMV inferred from gammaretroviruses for **(A)** *gag*, **(B)** *pol*, **(C)** *env* genes and the proteins **(D-F)**. Combination of Gblock parameters were applied to the whole-genome alignments (refer to materials and methods) and the relationship of these viral sequences were constructed when three Gblock parameters **(G)**, first and second parameter **(H)**, second parameter **(I)**, second and third parameters **(J)** were applied. The viral trees were rooted using avian reticuloendotheliosis virus (REV) and bootstrap values represented by a color gradient as shown in the legends. The trees were visualized with the Interactive Tree Of Life (iTOL) v5 (44). The viral sequences identified in this study are marked with a red asterisk, *M. burtoni* 204 (MelWMV-NG) and the cMWMV clade are highlighted in orange and blue, respectively, showing the evolutionary relationship of these sequences remains largely consistent.

A.



B.

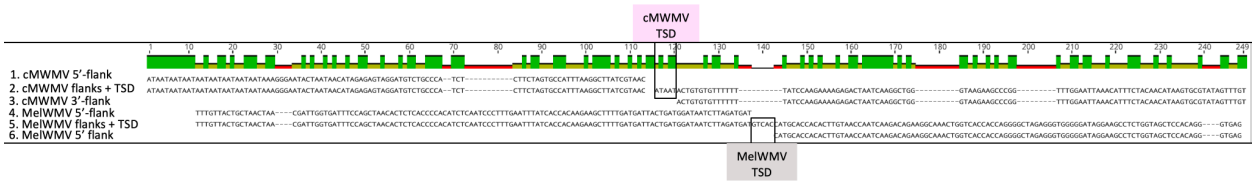
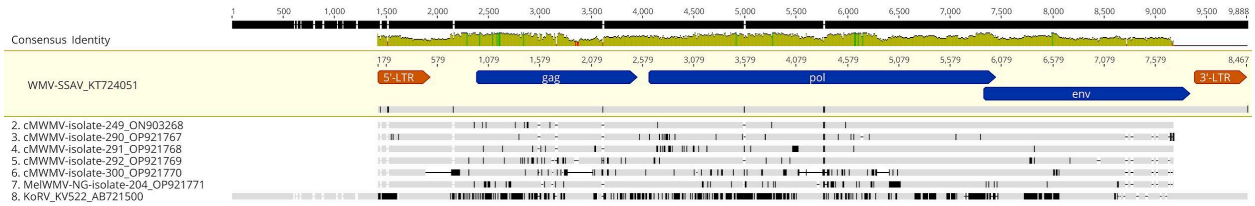


Fig. S3

We determined cMWMV is endogenizing in the genome of *Melomys* based on shared flanking sequences and an identical target site duplication (TSD). **(A)** We aligned the merged sequencing reads to WMV, using Geneious mapper. Except for *M. burtoni* (201 and 204), one of the *M. leucogaster* (290), we identified flanking LTR sequences for *M. leucogaster* (249, 291, 292 and 300) which were identical. TSD were identified by aligning the flanking 5' and 3' sequences. Using BLASTn, we confirm that the flanking sequences are host genomic sequences (green boxes). For each *M. leucogaster* harboring cMWMV, 5' and 3' sequences are shown for samples for which they could be retrieved given sequence coverage differences per sample. **(B)** For a better visualization, the identified TSD for cMWMV (pink box) and MelWMV (gray box) was concatenated to the respective 5' and 3' flanks (Text file S1). The flanking sequence for cMWMV does not align to the MelWMV flanking sequence found in *M. burtoni* from Indonesia and is therefore an independent germline integration event. The figure is displayed using Geneious Prime.



		249	290	291	292	300	204
	249		97	99	87	84	68
	290	97		96	84	82	66
M. leucogaster	291	99	96		86	83	68
	292	87	84	86		75	63
	300	84	82	83	75		60
M. burtoni	204	68	66	68	63	60	

Fig. S4

Nucleotide alignment of the assembled cMWMV sequences, KoRV (AB721500) and WMV (KT724051) showing positions of proviral genes *gag*, *pol*, *env* in blue arrows. The 5' and 3' long-terminal-repeats (LTRs) (orange arrows). Consensus sequence identity is based on a 75% threshold. The nucleotides that disagree with the consensus are indicated in black. cMWMV has retained the typical gammaretroviral structures and better aligns to WMV (98.9% n.t identity) than KoRV (81.3%). This figure was generated using Geneious R9.1. The bottom matrix shows a high percentage of nucleotide sequence identity amongst cMWMV isolates which is in confirmation of the short branch length in cMWMV clade. The Australian *M. burtoni* 204 (MelWMV-NG) is more divergent than the Papuan isolates.

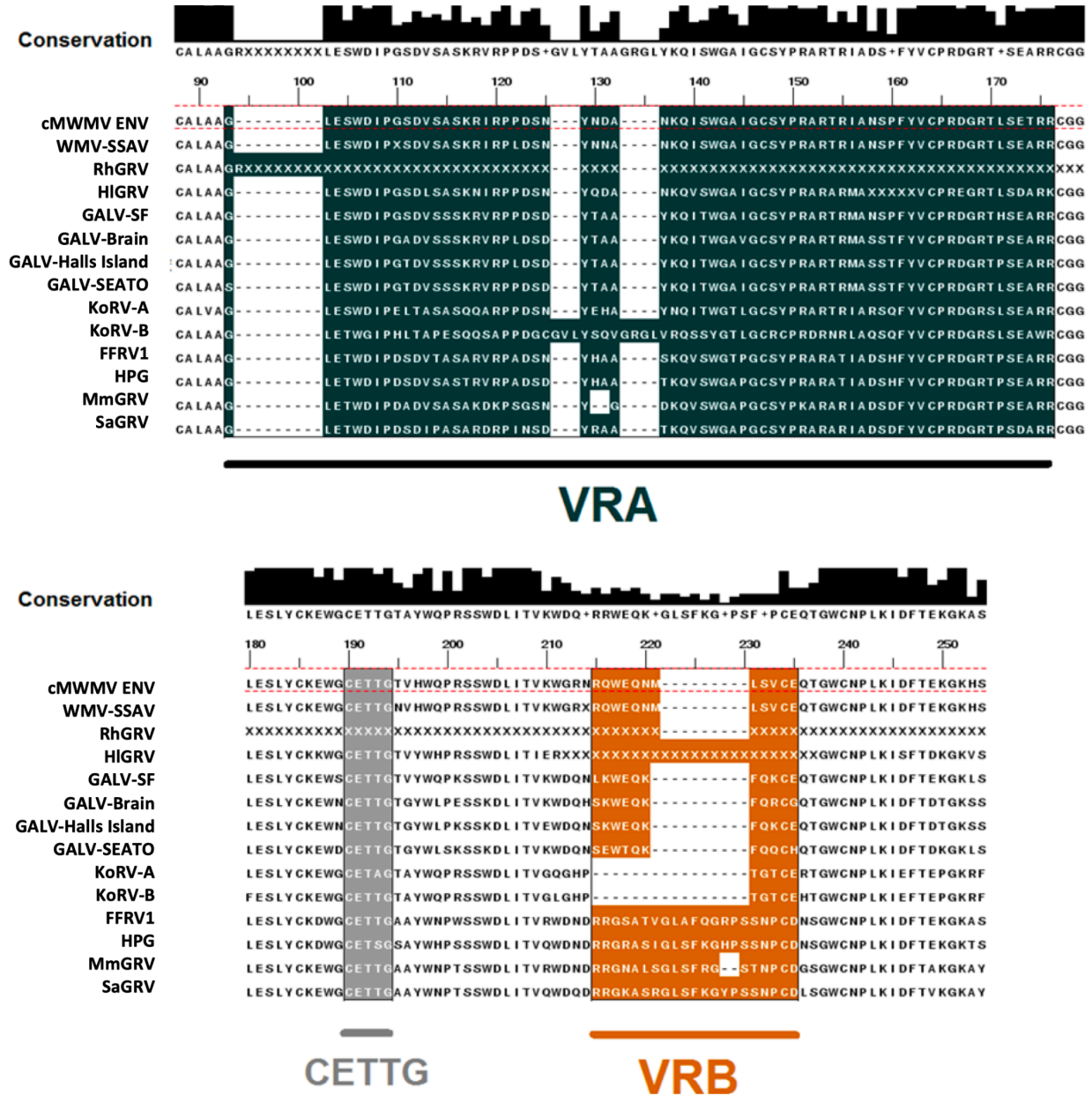


Fig. S5

Multiple sequence alignment of the *env* genes of cMWMV with other GALV-KoRV viruses. The alignment was generated using MAFFT (19) and visualized in Jalview v2.11.1.7 (20). Variable regions A and B (VRA; VRB) in the receptor binding domain (RBD) as determinants of cell tropism are marked. The CETTG motif upstream of VRB is present in all infectious GALVs and is different from CETAG of the endogenous KoRV-A (AB721500).

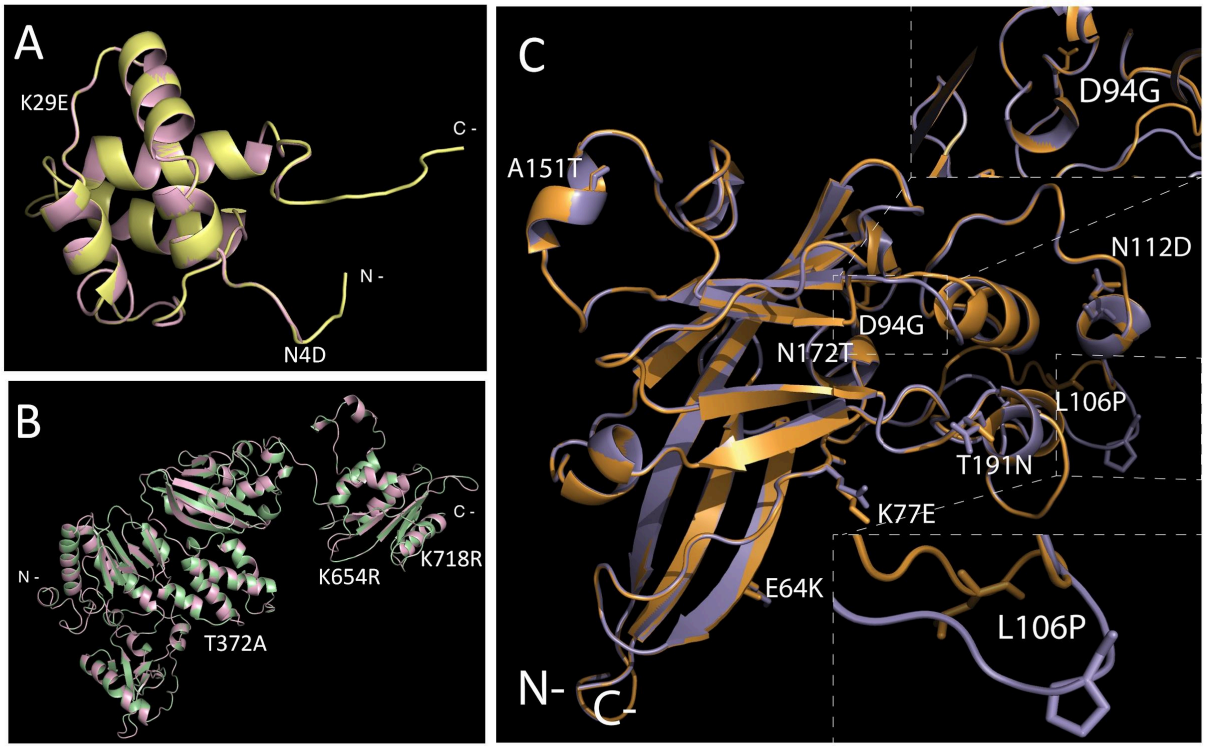


Fig. S6

Structural superimpositions of cMWMV and WMV proteins reveal similar 3D structures for: (A) cMWMV (yellow) and WMV (pink) GAG (residues modeled 1-101) protein structure, (B) cMWMV (pink) and WMV (green) POL (residues modeled 103-740) protein structure and (C) cMWMV (purple) and WMV (orange) ENV (modeled residues 44-255) protein structure. In all three panels, the identified mutated amino acids are indicated in white.

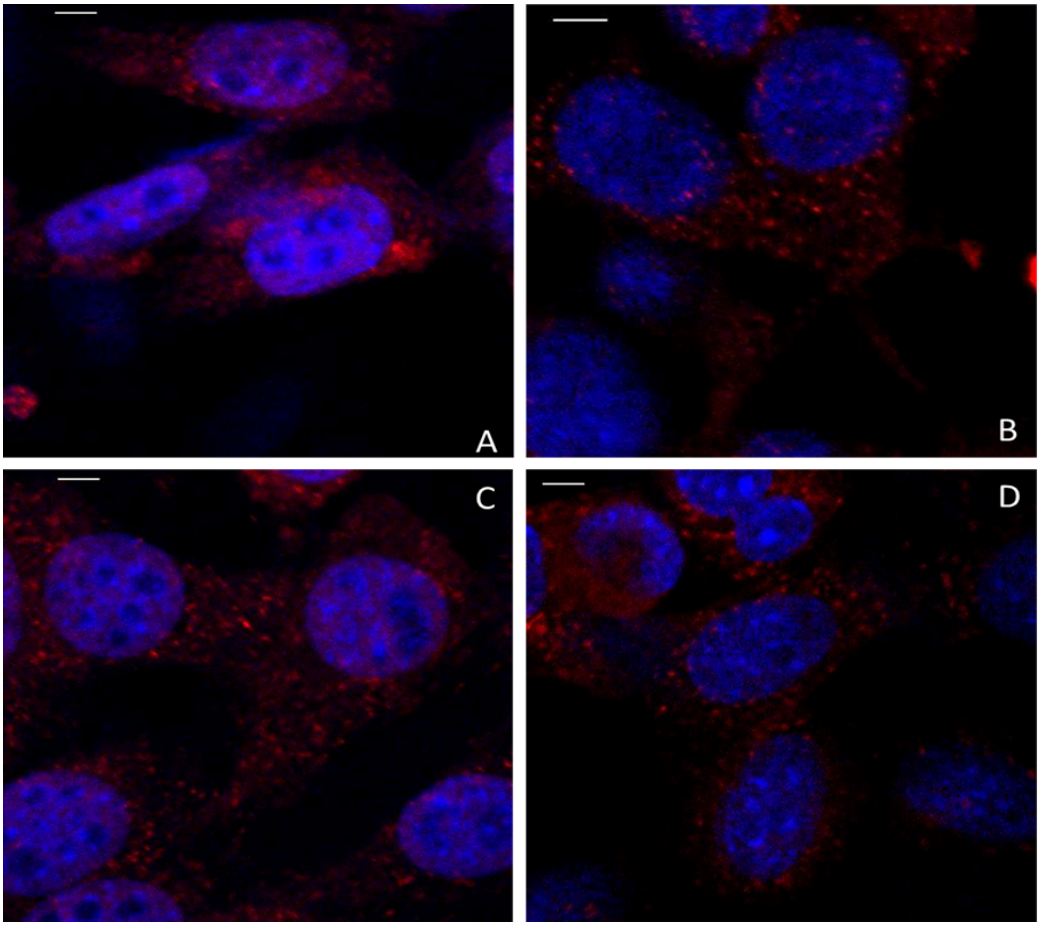


Fig. S7

Confocal immunofluorescence microscopy images of PiT-1 protein (red) in human HEK293T (A) and mouse NIH3T3 cells (C) as well as PiT-2 protein (red) in HEK293T (B) and NIH3T3 (D) cells. Both, PiT-1 and PiT-2 antibodies, were conjugated with AlexaFluor®546. The Hoechst nuclear counterstain is shown in blue and the bar represents 5 µm.

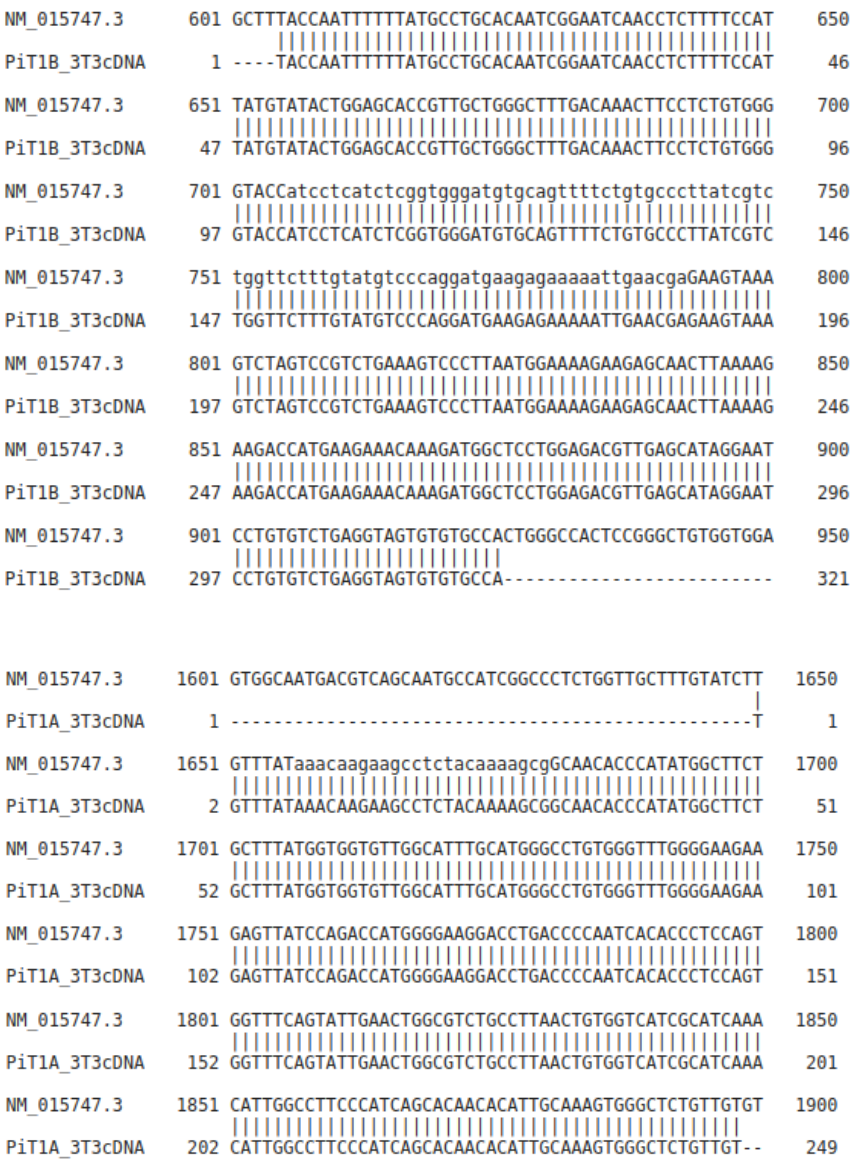


Fig. S8

EMBOSS Needle global pairwise alignment (45) of *Mus musculus* PiT-1 (NM_015747.3) with the NIH3T3 cells used in our experiments. The underscored residues on the top panel correspond to motif B and the bottom to motif A, showing that the KoRV-A infection observed here was not due to the receptor's evolution.

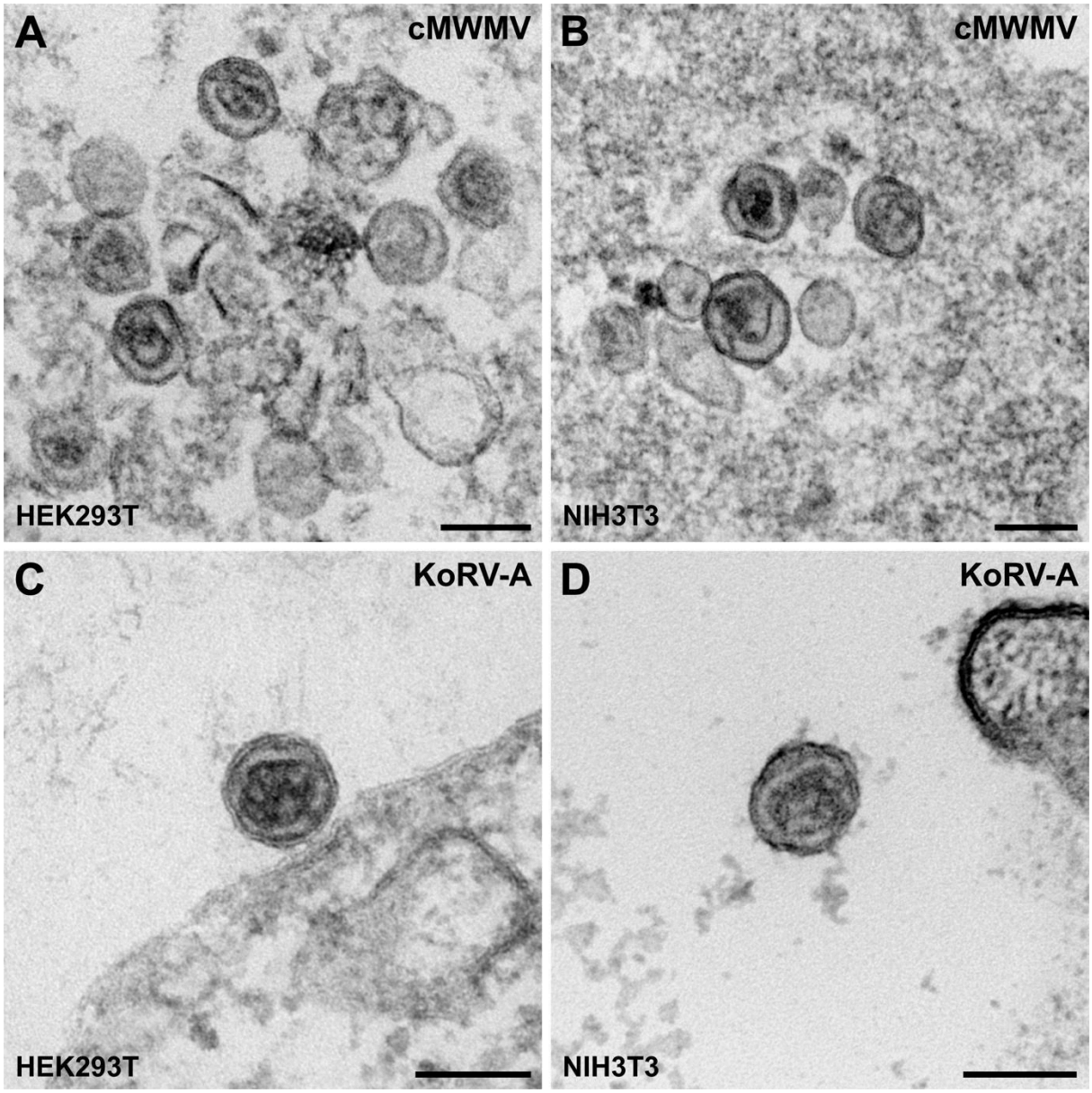


Fig. S9

Electron microscopy of thin sections through either cMWMNV-infected (**A, B**) or KoRV-A-infected (**C, D**) HEK293T (**A, C**) and NIH3T3 (**B, D**) cells. **A, B** cMWMV particles produced by the different cell types reveal the same morphology and structural variability, such as particle size, core density and shape. **C, D** KoRV-A particles produced by the cells demonstrate a similar morphology than the cMWMV particles produced in the same cell lines (compare also with Fig. 4A). Scale bars = 100 nm.

Supplementary Tables and Files

Table S1

The percentage of total reads mapping to the target genome was considered as an enrichment efficiency factor. We calculated this per amplified sequencing libraries from the VIP coverage output of gammaretroviruses.

Sample	Total reads	Gammaretroviral hits	VIP allocated RefSeq	Refseq length (bp)	Coverage RefSeq %	Enrichment efficiency
89	4,775,007	38,767	Baboon (NC_022517)	8,507	72.92	8.12E-03
201	207,628	3,090	KoRV (AB721500)	8,440	55.91	1.49E-02
204	962,301	14,545	MelWMV (KX059700)	5,488	94.56	1.51E-02
246	2,836,868	20,973	Baboon (NC_022517)	8,507	67.1	7.39E-03
249	7,264,635	142,342	WMV (KT724051)	8,467	100	1.96E-02
290	2,665,507	38,593	WMV (KT724051)	8,467	100	1.45E-02
291	1,290,687	22,424	WMV (KT724051)	8,467	100	1.74E-02
292	2,545,823	35,853	KoRV (AB721500)	8,440	98.34	1.41E-02
300	3,870,630	77,153	KoRV (AB721500)	8,440	100	1.99E-02

Table S2

GenBank nucleotide sequences used for alignment and phylogenetic trees. Genbank records in bold are from this study.

	Virus	Abbreviation	Host	GenBank accession no.	GAG	POL	ENV
1	Predicted: Cricetulus griseus (LOC113837738)	C. griseus	rodent	XM_027433137		XP_027288938	
2	Predicted: Colius striatus endogenous retrovirus group K (LOC104555315)	C. striatus	bird	XM_010198675		XP_010196977	
3	Feline endogenous retrovirus ERV-DC7	FeLV	cat	AB807599			
4	Flying-fox retrovirus (isolate FFRV1)	FFRV1	bat	MK040728	QDA02049	QDA02050	QDA02051
5	Gibbon ape leukemia virus strain Brain	GALV Brain	gibbon	KT724049	ALV83305	ALV83306	ALV83307
6	Gibbon ape leukemia virus strain Hall's Island	GALV Hall's Island	gibbon	KT724050	ALV83308	ALV83309	ALV83310
7	Gibbon ape leukemia virus strain SEATO	GALV SEATO	gibbon	KT724048	ALV83302	ALV83303	ALV83304
8	Gibbon ape leukemia virus strain SEATO	GALV M26927	gibbon	M26927	AAA46809	AAA46810	AAA46811
9	Gibbon ape leukemia virus strain SEATO	GALV NC_001885	gibbon	NC_001885	NP_056789	NP_056790	NP_056791
10	Gibbon ape leukemia virus strain San Francisco	GALV SF	gibbon	KT724047	ALV83299	ALV83300	ALV83301
11	Gibbon ape leukemia virus strain X	GALV-X	gibbon	U60065	AAC80263	AAC80264	AAC80265
12	Hipposideros larvatus gammaretrovirus	HlGRV	bat	MN413613	QJT93255	QJT93256	QJT93257
13	Hervey pteropid gammaretrovirus	HPG	bat	MN413610	QJT93246	QJT93247	QJT93248
14	Predicted: Jaculus jaculus endogenous retrovirus group K (LOC105945030)	J. jaculus	rodent	XM_012952149		XP_012807603	
15	Koala retrovirus - variant A (clone KV522)	KoRV-A (AB721500)	koala	AB721500	BAM67146	BAM67146	BAM67147
16	Koala retrovirus - variant A (isolate Pci-QMJ6480)	KoRV-A (KF786284)	koala	KF786284	AHY24811	AHY24812	AHY24813
17	Koala retrovirus - variant A (isolate Pci-SN265)	KoRV-A (KF786285)	koala	KF786285	AHY24814	AHY24815	AHY24816
18	Koala retrovirus - variant B (isolate Br2-1CETTG)	KoRV-B	koala	KC779547	AGO86849	AGO86849	AGO86848
19	Predicted: Mastomys coucha (LOC116086244)	M. coucha	rodent	XM_031364589		XP_031220449	
20	Predicted: Myotis davidii endogenous retrovirus group K (LOC107184980)	M. davidii	bat	XM_015571801		XP_015427287	
21	Melomys burtoni retrovirus (isolate BRME001)	MbRV	rodent	KF572483		AIK23433	
22	Melomys burtoni retrovirus (isolate BRME002)	MbRV	rodent	KF572484		AIK23434	
23	Mus caroli endogenous virus	McERV	rodent	KC460271	AGP25479	AGP25480	AGP25481
24	Mus dunni endogenous virus	MDEV	rodent	AF053745	AAC31803	AAC31805	AAC31806
25	Melomys woolly monkey virus (isolate WD279)	MelWMV	rodent	KX059700			
26	Megaderma lyra retrovirus (isolate MlRV)	MlRV	bat	JQ951956		AFM52260	
27	Macroglossus minimus gammaretrovirus	MmGRV	bat	MN413611	QJT93249	QJT93250	QJT93251
28	Myotis ricketti retrovirus	MrRV	bat	JQ292912		AFF57737	
29	Porcine endogenous retrovirus A (clone 907F8)	PERV-A	pig	HQ540591	ASU50141	ASU50141	ASU50142
30	Porcine endogenous retrovirus B (clone 742H1)	PERV-B	pig	HQ540594	AAM29194	AAM29194	AAM29193
31	Porcine endogenous retrovirus C	PERV-C	pig	HM159246	ADK35877	ADK35878	ADK35879
32	Predicted: Rattus norvegicus (LOC102557044)	R. norvegicus	rodent	XM_008776280		XP_008774502	
33	Reticuloendotheliosis virus (strain SDAUR-S1)	REV	bird	MF185397	ASH96781	ASH96780	ASH96782
34	Rhinolophus ferrumequinum retrovirus (isolate RfRV)	RfRV	bat	JQ303225	AFA52558	AFA52559	AFA52560
35	Rhinolophus hipposideros gammaretrovirus	RhGRV	bat	MN413614	QJT93258	QJT93259	QJT93260
36	Syconycteris australis gammaretrovirus	SaGRV	bat	MN413612	QJT93252	QJT93253	QJT93254
37	Woolly monkey virus strain WMV SSAV	WMV	gibbon	KT724051	ALV83311	ALV83312	ALV83313
38	**complete Melomys Woolly monkey virus isolate 249**	**cMWMV-249**	**rodent**	**ON903268**			
39	**complete Melomys Woolly monkey virus isolate 290**	**cMWMV-290**	**rodent**	**OP921767**			
40	**complete Melomys Woolly monkey virus isolate 291**	**cMWMV-291**	**rodent**	**OP921768**			
41	**complete Melomys Woolly monkey virus isolate 292**	**cMWMV-292**	**rodent**	**OP921769**			
42	**complete Melomys Woolly monkey virus isolate 300**	**cMWMV-300**	**rodent**	**OP921770**			
43	**Melomys woolly monkey virus variant New-Guinea (isolate 204)**	**MelWMV-NG**	**rodent**	**OP921771**			

Table S3

Properties of 46 amino acid substitution detected in cMWMV with reference to WMV. This result indicates none of the identified mutations alter the protein functionality as none fulfilled more than three of the set criteria (SI Methods) in our computational strategy.

Replacement Mutations	Protein	BLOSUM 62*	BLOSUM 80*	Mismatches (%)	Motif	Conserved Domain Consensus Amino Acid	3D model	SIFT prediction	PROVEAN prediction	Major Structural change
N4D	GAG	1	1	95	No	N/A	Yes	Affect (low confidence)	neutral	No
K29E	GAG	1	1	95	No	N/A	Yes	tolerated	neutral	No
A100V	GAG	0	0	60.4	No	N/A	Yes	tolerated	neutral	No
S115A	GAG	1	1	78.2	No	N/A	Yes	tolerated	neutral	No
P154Q	GAG	-1	-2	63.4	No	N/A	Yes	tolerated	neutral	No
D371E	GAG	2	1	4	No	N/A	Yes	Affect (low confidence)	deleterious	No
E374S	GAG	0	0	29.7	No	N/A	Yes	tolerated	neutral	No
R445K	GAG	2	1	26.7	No	N/A	Yes	tolerated	neutral	No
Q507R	GAG	1	1	94.1	No	N/A	Yes	tolerated	neutral	No
A518S	GAG	1	1	42.6	No	N/A	No	tolerated	neutral	N/A
I30V	POL	3	3	53.5	No	N/A	No	Affect (low confidence)	neutral	N/A
P372A	POL	-1	-1	96	No	N/A	Yes	tolerated	neutral	No
K654R	POL	2	2	17.8	Yes	C654R	Yes	tolerated	neutral	No
K718R	POL	2	2	23.8	Yes	A718R	Yes	tolerated	neutral	No
A734T	POL	0	0	2	Yes	A734T	Yes	tolerated	deleterious	No
K755E	POL	1	1	63.4	No	N/A	No	tolerated	neutral	No
V870I	POL	3	3	4	Yes	I870I	Yes	tolerated	neutral	No
I1041M	POL	1	1	60.4	No	N/A	Yes	tolerated	neutral	No
P1058S	POL	-1	-1	85.1	Yes	Q1058S	Yes	tolerated	neutral	No
G1059S	POL	0	-1	13.9	Yes	K1059S	Yes	tolerated	neutral	No
V1088I	POL	3	3	56.4	Yes	V1088I	Yes	tolerated	neutral	No
E64K	ENV	1	1	95	N/A	N/A	Yes	tolerated	neutral	No
K77E	ENV	1	1	85.1	N/A	N/A	Yes	tolerated	neutral	No
D94G	ENV	-1	-2	72.3	N/A	N/A	Yes	tolerated	neutral	No
L106P	ENV	-3	-3	43.6	N/A	N/A	Yes	tolerated	neutral	Yes
N112D	ENV	1	1	65.3	N/A	N/A	Yes	tolerated	neutral	No
A151T	ENV	0	0	32.7	N/A	N/A	Yes	Affect protein function (low confidence)	neutral	No
N172T	ENV	0	0	94.1	N/A	N/A	Yes	tolerated	neutral	No
T191N	ENV	0	0	62.4	N/A	N/A	Yes	tolerated	neutral	No

Replacement Mutations	Protein	BLOSUM 62*	BLOSUM 80*	Mismatches (%)	Motif	Conserved Domain Consensus Amino Acid	3D model	SIFT prediction	PROVEAN prediction	Major Structural change
R269G	ENV	-2	-3	83.2	N/A	N/A	No	tolerated	neutral	N/A
P277L	ENV	-3	-3	42.6	N/A	N/A	No	tolerated	neutral	N/A
S279P	ENV	-1	-1	77.2	N/A	N/A	No	tolerated	neutral	N/A
R368H	ENV	0	0	68.3	N/A	N/A	No	tolerated	neutral	N/A
T371A	ENV	0	0	87.1	N/A	N/A	No	tolerated	neutral	N/A
Y418H	ENV	2	2	94.1	N/A	N/A	No	tolerated	neutral	N/A
A422S	ENV	1	1	17.8	N/A	N/A	No	Affect protein function	deleterious	N/A
Q445H	ENV	0	1	61.4	N/A	N/A	No	Affect protein function	neutral	N/A
S467P	ENV	-1	-1	66.3	N/A	N/A	No	tolerated	neutral	N/A
A490T	ENV	0	0	5	N/A	N/A	No	Affect protein function	neutral	N/A
T497A	ENV	0	0	50.5	N/A	N/A	No	tolerated	neutral	N/A
I515T	ENV	-1	-1	53.5	Heptad repeat 1 Domain	H515T	Yes	tolerated	neutral	No
A519T	ENV	0	0	82.2	Heptad repeat 1 Domain	A519T	Yes	tolerated	neutral	No
I524L	ENV	2	1	94.1	Heptad repeat 1 Domain	L524L (Same AA in consensus)	Yes	tolerated	neutral	No
N533D	ENV	1	1	94.1	Heptad repeat 1 Domain	K5533D	Yes	tolerated	neutral	Yes
A542V	ENV	0	0	95	Heptad repeat 1 Domain	V542V (Same AA in consensus)	Yes	tolerated	neutral	Yes
I631T	ENV	-1	-1	88.1	N/A	N/A	No	tolerated	neutral	N/A

Text file S1

The flanking 5’ and 3’ LTR sequences of cMWMV and MelWMV (6).

>cMWMV_5prime_flank_LTR
GTTACGATAAGCCTTAAATGGCACTAGAAGAGATGGGCAGACATCCTACTCTCTATGTTATTAGTATTCCCTTTATTATTATTATTATTA
TTATTATTAT

>cMWMV_3prime_flank_LTR
ACAAACTATACGCACTTATGTTGTAGAAATGTTTAATTCCAAACCGGGCTTCTTACCCAGCCTTGATTAGTCTCTTTTCTTGGATAAA
AAAACACACAGT

>MelWMV_5prime_flank_LTR
TTTGTTACTGCTAACTAACGATTGGTGATTTCCAGCTAACACTCTCACCCCACATCTCAATCCCTTTGAATTTATCACCACAAGAAG
CTTTTGATGATTACTGATGGATAATCTTAGATGATGTCAC

>MelWMV_3prime_flank_LTR
GTCACCATGCACCACACTTGTAACCAATCAAGACAGAAGGCAAACTGGTCACCACCAGGGGCTAGAGGGTGGGGGATAGGAAG
CCTCTGGTAGCTCCACAGGGTGAG

Text file S2

The curated consensus sequence of cMWMV that was cloned and used for functional studies.

>cMWMV_consensus|complete_genome|8459bp|

TGTTTTTCAAGCTAGCTGCAGTAACGCCATTTTGCAAGGCACGGAAAATTACCCTGGTAAAAAGCCCAAAGCATAGGGAAAGTAC
AGCTAAAGGTCAAGTCGAGAAAAACAAGGAGAACAGGGCCAAACAGGATATCTGTGGTCATGCACCTGGGCCCCGGCCCAGGG
CCAAAGACAGATGGTTCCCAGAAATAGATGAGTCAACAGCAGTTTCCAGGGTGCCCCTCAACTGTTTCAAGAACTCCCACATGA
CCGGAGCTCACCCCTGGCCTTATTTGAACTGACCAATTACCTTGCTTCTCGCTTCTGTACCCGCGCTTTTTGCTATAAAATGAGCT
CAGAAACTCCACTCGGCGCGCCAGTCCTCCGAGAGACTGAGTCGCCCGGGTACTCGTGTGTTCAATAAAACCTCTTGCTATTTG
CATCCGAAGCCGTGGTCTCGTTGTTCCTTGGGAGGGTCCCTCCTAACTGATTGACTGCCCACCTCGGGGGTCTTTCATTTGGGG
GCTCGTCCGGGATCGGAGACCCCCACCCAGGGACCACCGACCCACCAACGGGAGGTAAGCTGGCCAGCGATCGCTCTGTGTC
TCGTTTCTGTGTCTAACTCCGTAACTCTGACTGTCCCTCTGAGTGCGCGCATTTTGGTTTCAGTTTGTTCCGGGCTGATCGCTCT
GTGAGCGACGTGTGAGTAGCGAGCAGACGTGTTCGGGGGCTCACCGCCCGATAATCCTGGGAGACGTCCCAGGATCAGGGGA
GGACCAGGGACGCCTGGTGGACCCCTCGGCAGAGGATCATTGTGTTCTGATCCCACTGCCGCGTCTAGAGAGGCGCGCTCTG
CCATCTGACTCTTTCTTTTGCCTCTACGCTACTCGATCTCGCCGCCGTTTCTGGTTTCTTTTTGTTTTGTTTCTGATAAGCCTCTGT
GTCGTAGTCCCTCTCTTCGAAATCTTCGAAATGGGACAAGATAACTCTACCCCTCTCTCCCTTACTCTAGATCACTGGAAAGATGT
GAGAACAAGGGCTCACAATCTGTCCGTGGAAATCAGAAAGGGAAAATGGCAGACTTTCTGTTCCTCCGAGTGGCCCACATTCGG
CGTGGGGTGGCCGCCGGAGGGAACTTTTAATCTCTCTGTCATTTTTGCAGTTAAAAGGATTGTCTTTCAGGAAACCGGAGGACA
CCCGGACCAAGTTCCATACATCGTGGTGTGGCAGGACCTCGCCCAGAGTCCCCCACCATGGGTGCCGCCCTCCGTCAAGATCG
CTGTTGTCTCTAGTCCAGAGAACACTCGAGGACCAGCTGCGGGGAGGCCATCCGCTCCTCCCCGACCCCCCATCTACCCGGCA
ACAGACGACTTGCTCCTTCTCTCTGAGCCCCCGCCCTATCCGGCGGCCCTGCCACCTCCTCTGGCCCCTCAGGCGGTCGGAC
CGGCGCCGGGCCAGGCGCCCGATAGTTCCGATCCTGAGGGACCAGCCGCTGGGACCAGGAGTCGCCGTGCCCGCAGTCCGG
CAGACGACTCGGGTCCTGACTCCACTGTGATTTTGCCCCTCCGAGCCATAGGACCCCCAGCCGAGCCCAACGGCCTGGTCCCT
CTACAATATTGGCCTTTTTCCTCAGCAGATCTTTATAATTGGAAATCTAACCATCCTTCTTTTTCTGAAAATCCAGCAGGACTCACG
GGGCTCCTTGAGTCTCTTATGTTTTCTCATCAGCCCACTTGGGACGATTGCCAACAGCTCCTACAGATTCTCTTCACCACTGAGG
AGCGGGAAAGGATTCTCCTGGAGGCCCGCAAGAATGTCCTTGGGGACAACGGGGCCCCTACTCAACTCGAGAACCTCATTAAT
GAGGCCTTCCCCCTCAATCGACCTCAGTGGGATTACAACACGGCCGCAGGTAGGGAGCGTCTCCTGGTCTACCGCCGGACTCT
AGTGGCAGGTCTCAAAGGGGCAGCTCGGCGCCCCACCAATTTGGCTAAGGTAAGAGAGGTCTTGCAGGGACCGGCAGAACCC
CCTTCGGTTTTCTTAGAACGCCTAATGGAGGCCTATAGGAGATACACTCCGTTTGAACCCTCTTCTGAGGGGCAGCAGGCTGCG
GTTGCCATGGCCTTCATCGGACAGTCAGCCCCAGATATCAAGAAAAAGTTACAGAGGCTAGAGGGGCTCCAAGATTATTCCTTAC
AAGATTTAGTGAGAGAGGCAGAGAAGGTGTACCACAAGAGAGAGACAGAAGAAGAAAGACAAGAAAGAGAAAAGAAAGAGGCA
GAAGAGAGAGAGAGGCGGCGCGATAAGCGCCAAGAGAAAAACTTGACTAGGATTTTGGCCGCAGTGGTAAGTGAAAGAGGGTC
TAGAGATAGGCAGACAGGGAACCTGAGCAACCGGGCAAGGAAGACACCTAGGGATGGAAGACCTCCTCTAGACAAAGACCAGT
GCGCGTACTGTAAAGAGAAGGGTCACTGGGCAAGAGAATGTCCCCGAAAGAAGAACGTCAGAGAAGCCAAGGTTCTGTCCCTA
GATGACTAGGGGAGTCGGGGTTCGGACCCCCTCCCCGAACCTAGGGTAACACTGACTGTGGAGGGGACCCCCATTGAGTTCCT
GGTCGATACCGGGGCTGAACATTCGGTATTGACCCAACCCATGGGAAAGGTAGGGTCCAGACGGACAGTCGTGGAAGGAGCGA
CAGGAAGCAAAGTCTACCCCTGGACCACAAAGAGACTTTTAAAAGTTGGACATAAACAAGTGACCCACTCCTTTCTGGTCATACC
CGAGTGCCCCGCTCCTCTGTTGGGCAGGGACCTCCTAACCAAACTAAAGGCCCAGATCCAGTTTTCTGCTGAGGGCCCACAGG
TAACATGGGAAGACCGCCCTACTATGTGCCTGGTCCTAAACCTAGAAGAAGAATACCGGCTACATGAAAAGCCAGTTCCCTCCTC

TATCGACCCATCCTGGCTCCAGCTTTTTCCCACTGTATGGGCAGAGAGGGCAGGCATGGGACTGGCCAATCAAGTCCCACCAGT
GGTAGTAGAACTGAGATCAGGTGCCTCACCGGTGGCTGTTCGACAATATCCAATGAGCAAAGAAGCCCGGGAAGGTATCAGACC
CCACATCCAAAGGTTCTTAGACCTAGGGGTCTTGGTGCCCTGTCAGTCGCCCTGGAATACCCCTCTACTACCTGTAAAGAAGCCA
GGGACCAATGACTATCGGCCAGTCCAAGACCTGAGAGAAATTAATAAGAGGGTACAGGATATTCATCCCACAGTCCCAAACCCTT
ACAACCTTCTGAGTTCCCTTCCGCCTAGCCACACTTGGTACTCAGTCTTAGATCTCAAGGATGCCTTTTTCTGCCTCAAGCTACAT
CCCAACAGCCAGCCGCTGTTCGCGTTCGAGTGGAGAGACCCAGAAAAAGGTAACACAGGTCAGCTGACCTGGACACGGCTAC
CACAAGGGTTCAAGAACTCTCCCACTCTCTTCGACGAGGCCCTCCACCGAGATTTGGCTCCCTTTAGGGCCCTCAACCCCCAG
GTAGTGTTACTCCAATATGTAGACGACCTCCTGGTGGCGGCCCCCACATATAGAGACTGCAAAGAAGGGACACAGAAGCTCCTA
CAGGAATTGAGTAAGTTGGGGTACCGGGTATCGGCTAAGAAGGCCCAGCTCTGCCAGAAAGAGGTCACCTATCTGGGGTACTTG
CTCAAGGAAGGGAAAAGATGGCTGACCCCGGCCCGAAAGGCTACTGTTATGAAGATCCCCGCTCCCACGACCCCCAGACAGGT
CCGTGAATTTCTGGGCACTGCTGGATTCTGCAGGCTCTGGATCCCTGGGTTTGCTTCCCTGGCTGCACCCTTGTACCCCCTAAC
AAAGGAAAGCATCCCTTTTATCTGGACTGAGGAACATCAGAAGGCTTTTGACCGCATAAAAGAAGCCTTGCTGTCAGCCCCCGCT
TTGGCCCTCCCAGACCTCACCAAACCATTTACTCTATACGTAGATGAGAGGGCCGGCGTGGCCCGGGGAGTGCTTACTCAGACT
TTAGGACCCTGGCGGCGGCCGGTAGCTTATCTATCGAAGAAACTGGATCCGGTGGCCAGCGGGTGGCCAACCTGCCTGAAAGC
GGTTGCAGCAGTGGCACTCCTTCTCAAGGACGCTGATAAGTTAACCTTGGGACAAAATGTGACTGTGATTGCTTCCCATAGCCTC
GAAAGCATCGTGCGGCAGCCCCCCGACCGGTGGATGACCAATGCCAGGATGACTCATTACCAGAGCCTGCTGCTAAATGAAAG
GGTATCGTTCGCGCCCCCTGCCGTCCTGAACCCAGCTACCCTACTACCAGTCGAGTCGGAAGCCACCCCAGTGCACAGGTGCT
CAGAAATCCTCGCCGAAGAAACTGGAACTCGACGAGACCTGAAGGACCAGCCATTGCCCGGGGTGCCAGCCTGGTATACGGAC
GGTAGCAGTTTCATCGCGGAAGGTAAACGGAGAGCAGGGGCCGCCATCGTAGATGGCAAGCGGACGGTGTGGGCAAGCAGCC
TGCCAGAAGGTACGTCAGCCCAGAAGGCCGAACTAGTGGCCTTGACGCAGGCATTACGCCTGGCCGAAGGAAGAGACATCAAC
ATCTACACAGACAGCAGGTATGCTTTTGCCACTGCTCATATTCATGGGGCAATATATAAACAGAGGGGGCTGCTCACTTCTGCTGG
AAAAGACATTAAAAACAAAGAAGAAATTTTGGCCCTGCTAGAGGCCATTCACCTTCCTAAGCGGGTCGCCATTATCCACTGCCCC
GGCCACCAGAGGGGAAATGACCCTGTGGCCACCGGGAACCGGAGGGCCGACGAGGCTACAAAACAAGCCGCCCTGTCGACCC
AGAGTGCTGGCAGAAACTACAAAACCTCAAGAGCTAATCGAACCCGCTCAGGTTAAGACCAGGCCAGGAGAGCTCACCCCTGA
CCGGGGGAAGGAATTCATTCAGCGGTTACATCAGTTAACACACTTAGGACCAGAGAAGCTTCTCCAACTAGTAAACCGCACCAG
CCTCCTCATCCCGAACCTCCAGTCTGCAGTTCGCGAAGTCACCAGTCAGTGTCAGGCTTGTGCCATGACTAATGCGGTCACAAC
CTACCGAGAGACCGGAAAGAGGCAACGAGGAGATCGACCCGGCGTGTACTGGGAGGTAGACTTCACAGAGGTGAAGCCTGGC
CGGTATGGAAACAGGTATCTGCTGGTATTCATAGATACTTTTTCCGGATGGATAGAAGCTTTTCCTACCAAAACTGAAACGGCCCT
GACCGTCTGCAAGAAGATATTAGAAGAAATTCTACCCCGCTTCGGGATCCCTAAGGTACTCGGGTCAGACAATGGCCCAGCCTTT
GTTGCTCAGGTAAGTCAGGGACTGGCCACTCAACTGGGGATAAATTGGAAGTTACATTGTGCGTATAGACCCCAGAGCTCAGGT
CAGGTAGAGAGAATGAACAGGACAATCAAAGAGACCTTGACCAAATTAGCCTTAGAGACCGGTGGGAAAGACTGGGTGGCCCT
CCTTCCCTTAGCGCTGCTCAGAGCCAGGAATACCCCTGGCCGGTTTGGTCTAACTCCTTATGAAATTCTCTATGGGGGACCGCCC
CCCATACTTGAGTCTGGAGGGACATTGGGTCCCGATGATAATTTTCTCCCTGTCTTATTTACTCATTTAAAGGCTTTAGAAGTTGTG
AGGACCCAGATCTGGGACCAGATCAAGGAGGTGTACAAGCCCGGTACCGTGGCAATGCCCCACCCGTTCCAGGTCGGGGACC
AAGTGCTTGTCAGACGCCATCGATCCAGCAGCCTTGAGCCTCGGTGGAAAGGCCCGTACCTGGTGTTGCTGACCACCCCGACC
GCGGTAAAAGTCGACGGTATCGCTGCCTGGATCCATGCTTCTCACCTCAAGCCTGCACCACCCTCGGCACCAGATGAGTCCTGG
GAGCTGGAAAAGACTGATCATCCTCTTAAGCTGCGTATTCGGCGGCGGCGGAACGAGTCTGCAAAATAAGAACCCCCACCAGCC
TATGACCCTCACCTGGCAGGTACTGTCCCAAACTGGAGACGTTGTCTGGGTCACAAAGGCAGTCCAGCCCCTTTGGACTTGGTG
GCCCTCTCTTGAACCTGATGTATGTGCCCTGGCGGCCGGTCTTGAGTCCTGGGATATCCCGGGATCCGATGTATCGGCCTCTAA
AAGAATCAGACCCCCTGACTCAAACTATAATGACGCTAATAAGCAGATCAGCTGGGGAGCCATAGGATGCAGCTACCCTCGGGCT
AGGACCAGGATTGCAAATTCCCCCTTCTACGTGTGTCCCCGGGATGGCCGGACCCTTTCAGAAACTAGAAGGTGCGGGGGGCT
AGAATCCCTGTACTGTAAAGAATGGGGTTGTGAGACCACGGGGACCGTTCATTGGCAACCTAGGTCCTCATGGGACCTCATAACT
GTAAAATGGGGCCGAAATCGCCAATGGGAGCAAAATATGCTGTCAGTCTGTGAACAAACCGGCTGGTGTAACCCCCTCAAGATA
GATTTCACAGAAAAAGGGAAACACTCCAGGGATTGGATAAAGGGGAGAACCTGGGGATTGAGATTCAATGTGGCTGGACATCCA
GGCGTACAGTTGACCATTCGCTTGAAGGTCACCAGCATGCCAGCTGTGGCAGTGGGCCCCGACCCCGTCCTTGCGGAACAAG
GACCTCCTAGCAAGCCCCTCCCTCTCCCCCCGAGGGAAGCGCCGCCCACCTCTCTACCCCCGGCGGCTAGTGGGCAAGCCCC
CACGGTGCGGGAGAGAACTGTTACCCTAAGCACTCCGCCTCCCACCACGGGCGACAGACTCTTTGGCCTTGTGCAGGGGGCC
TTCCTAGCCTTGAATGCTACCAACCCAGGGGCCACAGAGTCTTGCTGGCTCTGTTTGGCCATGGGCCCCCCTTATTATGAAGGA
ATAGCCTCTTTAGGAGAGGTCGCTTATACCTCCGACCATACCCGGTGCCACTGGGGGGCCCAAGGAAAGCTTACCCTCACTGAG
GTCTCAGGACACGGGTTGTGCATAGGAAAGGTGCCCTCTACCCATCAGCATCTTTGCAATCAGACCCTACCCATCAATTCCTCCA
AGGACCATCAGTACCTGCTCCCCTTCAACCATAGCTGGTGGTCCTGCAGCACTGGCCTCACCCCCTGCCTCTCCACCTCAGTTT
TTAATCAGTCTAGAGATTTCTGTATCCATATCCAGCTGATCCCTCGCATCTATTACCATCCTGAAGGAACTTTGTTGCAGGCCTATG
ACAATCCTCACCCCAGGCTTAAAAGAGAGCCTGTCTCACTTACCCTAGCTGTTTTACTGGGGTTGGGGATCGCGACAGGTATAGG
TACTGGCTCAGCCGCCCTAATTAAAGGGCCCATGGACCTCCAGCAAGGCCTGACCAGCCTCCAGACCGCCATGGATACTGACCT
CCGGGCCCTCCAGGACTCAATCAGCAAGCTGGAGGACTCGCTGACTTCCCTATCTGAGGTAGTGCTCCAAAATAGGAGAGGCC
TTGACTTACTGTTTCTAAAAGAAGGAGGCCTCTGCGCGGCCCTAAAAGAAGAGTGCTGTTTTTATGTGGACCACTCAGGTGCAGT
ACGAGACTCCATGAGAAAGCTCAAAGAAAGACTAGATAAGAGACAGTTAGAGCGCCAGAAGAACCAAAACTGGTATGAAGGGTG
GTTCAATAGCTCCCCTTGGTTCACTACCCTACTATCAACCATCGCCGGGCCCCTATTACTCCTCCTTCTGTTGCTCACCCTCGGG
CCCTGCATCATCAATAGGTTAGTCCAATTCATCAATGATAGGGTAAGTGCAGTTAAAATTCTGGTCCTTAGACAGAAATATCAGACC
CTAGATAACGAAGATAACCTTTGATTCCGCTCTAAGATTAGAGCTATCCACAAGAGAAATGGGGGAATGAAGGAAGTGTTTTTCAA
GCTAGCTGCAGTAACGCCATTTTGCAAGGCACGGAAAATTACCCTGGTAAAAAGCCCAAAGCATAGGGAAAGTACAGCTAAAGGT
CAAGTCGAGAAAAACAAGGAGAACAGGGCCAAACAGGATATCTGTGGTCATGCACCTGGGCCCCGGCCCAGGGCCAAAGACA
GATGGTTCCCAGAAATAGATGAGTCAACAGCAGTTTCCAGGGTGCCCCTCAACTGTTTCAAGAACTCCCACATGACCGGAGCTC
ACCCCTGGCCTTATTTGAACTGACCAATTACCTTGCTTCTCGCTTCTGTACCCGCGCTTTTTGCTATAAAATGAGCTCAGAAACTC
CACTCGGCGCGCCAGTCCTCCGAGAGACTGAGTCGCCCGGGTACTCGTGTGTTCAATAAAACCTCTTGCTATTTGCATCCGAAG
CCGTGGTCTCGTTGTTCCTTGGGAGGGTCCCTCCTAACTGATTGACTGCCCACCTCGGGGGTCTTTC

Supplementary References

1. L. R. Heaney, D. S. Balete, E. A. Rickart, *The Mammals of Luzon Island: Biogeography and Natural History of a Philippine Fauna* (JHU Press, 2016).

2. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

3. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).

4. N. Alfano, *et al.*, Variation in koala microbiomes within and between individuals: effect of body region and captivity status. *Sci. Rep.* **5**, 10189 (2015).

5. M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).

6. N. Alfano, *et al.*, Endogenous Gibbon Ape Leukemia Virus Identified in a Rodent (Melomys burtoni subsp.) from Wallacea (Indonesia). *Journal of Virology* **90**, 8169–8180 (2016).

7. N. Alfano, *et al.*, Non-invasive surveys of mammalian viruses using environmental DNA https:/doi.org/10.1101/2020.03.26.009993.

8. N. L. Yozwiak, *et al.*, Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl. Trop. Dis.* **6**, e1485 (2012).

9. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

10. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

11. B. Bushnell, J. Rood, E. Singer, BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).

12. Y. Li, *et al.*, VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Scientific Reports* **6** (2016).

13. M. Vilsker, *et al.*, Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics* **35**, 871–873 (2018).

14. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

15. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

16. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

17. A. Bankevich, *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

18. L. Orlando, M. T. P. Gilbert, E. Willerslev, Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet.* **16**, 395–408 (2015).

19. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in

Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

20. J. B. Procter, *et al.*, Correction to: Alignment of Biological Sequences with Jalview. *Methods Mol. Biol.* **2231**, C1 (2021).

21. L. McMichael, *et al.*, A novel Australian flying-fox retrovirus shares an evolutionary ancestor with Koala, Gibbon and Melomys gamma-retroviruses. *Virus Genes* **55**, 421–424 (2019).

22. J. A. Hayward, *et al.*, Infectious KoRV-related retroviruses circulating in Australian bats. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9529–9536 (2020).

23. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

24. D. Posada, jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* **25**, 1253–1256 (2008).

25. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

26. E. M. Humphries, K. Winker, Working through polytomies: auklets revisited. *Mol. Phylogenet. Evol.* **54**, 88–96 (2010).

27. G. Talavera, J. Castresana, Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology* **56**, 564–577 (2007).

28. A. Waterhouse, *et al.*, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

29. P. Benkert, M. Biasini, T. Schwede, Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27**, 343–350 (2011).

30. K. Tsangaras, *et al.*, Hybridization capture reveals evolution and conservation across the entire Koala retrovirus genome. *PLoS One* **9**, e95633 (2014).

31. W. L. DeLano, The PyMOL Molecular Graphics System. Schrödinger LLC www.pymol.org Version 2, http://www.pymol.org (2002).

32. K. Hess, *et al.*, Concurrent action of purifying selection and gene conversion results in extreme conservation of the major stress-inducible Hsp70 genes in mammals. *Sci. Rep.* **8**, 5082 (2018).

33. R. Oliverio, *et al.*, Functional characterization of natural variants found on the major stress inducible 70-kDa heat shock gene, HSPA1A, in humans. *Biochem. Biophys. Res. Commun.* **506**, 799–804 (2018).

34. P. Nguyen, *et al.*, Origin and Evolution of the Human Bcl2-Associated Athanogene-1 (BAG-1). *Int. J. Mol. Sci.* **21** (2020).

35. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

36. Y. Bromberg, B. Rost, SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823–3835 (2007).

37. Y. Choi, A. P. Chan, PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).

38. C. Wilson, M. S. Reitz, H. Okayama, M. V. Eiden, Formation of infectious hybrid virions with gibbon ape leukemia virus and human T-cell leukemia virus retroviral envelope glycoproteins and the gag

and pol proteins of Moloney murine leukemia virus. *J. Virol.* **63**, 2374–2378 (1989).

39. C. A. Wilson, K. B. Farrell, M. V. Eiden, Comparison of cDNAs encoding the gibbon ape leukaemia virus receptor from susceptible and non-susceptible murine cells. *J. Gen. Virol.* **75 (Pt 8)**, 1901–1908 (1994).

40. M. Laue, Electron microscopy of viruses. *Methods Cell Biol.* **96**, 1–20 (2010).

41. U. Löber, *et al.*, Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 8609–8614 (2018).

42. D. E. Alquezar-Planas, *et al.*, DNA sonication inverse PCR for genome scale analysis of uncharacterized flanking sequences. *Methods Ecol. Evol.* **12**, 182–195 (2021).

43. G. K. McEwen, *et al.*, Retroviral integrations contribute to elevated host cancer rates during germline invasion. *Nat. Commun.* **12**, 1316 (2021).

44. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

45. F. Madeira, *et al.*, Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).