An endogenous gibbon ape leukemia virus (GALV) identified in a rodent (*Melomys* sp.) from Indonesia

Niccolo Alfano[1], Johan Michaux[2,3], Pierre-Henri Fabre[4,5], Serge Morand[2,3], Ken Alpin[5], Kyriakos Tsangaras[1*], Ulrike Löber[1], Yuli Fitriana[6], Gono Semiadi[6], Yasuko Ishida[7], Kristofer M. Helgen[5], Alfred L. Roca[7], Maribeth V. Eiden[8], Alex D. Greenwood[1,9#]

[1] Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany

[2] Conservation Genetics Unit, Institute of Botany, University of Liège, Liège, Belgium

[3] CIRAD, Campus international de Baillarguet, Montpellier Cedex, France

[4] Harvard Museum of Comparative Zoology, Cambridge, Massachusetts, USA

[5] National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

[6] Museum Zoologicum Bogoriense, Research Center For Biology, Indonesian Institute of Sciences (LIPI), Cibinong, Indonesia

[7] Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[8] Section on Directed Gene Transfer, Laboratory of Cellular and Molecular Regulation, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland, USA

[9] Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany

Running Title:

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

*Present address: Kyriakos Tsangaras, Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus.

Word counts:

Abstract:

Main text:

Number of figures:

Number of tables:

ABSTRACT

Gibbon ape leukemia virus (GALV) and koala retrovirus (KoRV) most likely originated from a cross-species transmission of an ancestral retrovirus into koalas and gibbons via one or more intermediate as yet unknown hosts. A highly similar virus to GALV has been identified in an Australian rodent (*Melomys burtoni*) after extensive screening of Australian wildlife. GALV-like viruses have also been discovered in several Southeast Asian species although screening has not been extensive and viruses discovered to date are only distantly related to GALV. We therefore screened 26 Southeast Asian rodent species for KoRV- and GALV-like sequences, using hybridization capture and high-throughput sequencing, in the attempt to identify potential GALV and KoRV hosts. Only one species, an undescribed species of *Melomys* from Indonesia, was positive yielding an endogenous provirus very closely related to a strain of GALV. The sequence of the critical receptor domain for GALV infection in the Indonesian *Melomys* sp. was consistent with the susceptibility of the species to GALV infection. The discovery of a GALV in a second *Melomys* species provides further evidence that *Melomys* may play a role in the spread of GALV-like viruses, especially since the genus is found in Indonesia, Papua New Guinea and Australia, connecting the home ranges of koalas and gibbons.

IMPORTANCE

The gibbon ape leukemia virus (GALV) and the koala retrovirus (KoRV) are very closely related, yet their hosts are neither closely related nor overlap geographically. Direct cross-species infection between koalas and gibbons is unlikely. Therfore, GALV and KoRV may have arisen via a cross-species transfer from an intermediate host that overlaps in range with both gibbons and koalas. Using hybridization capture and high-throughput sequencing, we have screened a wide range of rodent candidate hosts from Southeast Asia for KoRV- and GALV-like sequences. Only a *Melomys* species from Indonesia was positive for GALV. We report the genome sequence of this newly identified GALV, the critical domain for infection of its potential cellular receptor and its phylogenetic relationships with the other previously characterized GALVs. We hypothesize that the genus *Melomys* may have played a key role in cross-species transmission to other taxa.

INTRODUCTION

The evolutionary mechanisms involved in cross-species transmissions (CST) of viruses are complex and generally poorly understood. Viral evolution, host contact rates, biological similarity and host evolutionary relationships have been proposed as key factors in CST rates and outcomes (1). However, there are cases where the CSTs occur between hosts that are biogeographically separated, distantly related or both. For example, the koala retrovirus (KoRV) and the gibbon ape leukemia virus (GALV) are very closely related viruses (2) that infect hosts that are neither sympatric nor closely related. GALV is an exogenous gammaretrovirus that has been isolated from captive white-handed gibbons (*Hylobates lar*) held in or originally from Southeast Asia (3-6). Of the five GALV strains identified so far, four have been isolated in gibbons (3-6) and one – the woolly monkey virus (WMV), formerly referred to as SSAV (7, 8) – in a woolly monkey (*Lagothrix lagotricha*), probably as the result of an horizontal transmission of GALV from a gibbon. KoRV is a potentially infectious endogenous retrovirus (ERV) of wild koalas (*Phascolarctos cinereus*) in Australia and captive koalas worldwide (9-11). Both viruses are associated with lymphoid neoplasms in their hosts (12, 13). KoRV and GALV share high nucleotide sequence similarity (80%) and form a monophyletic clade within gammaretroviruses (2). In contrast, the species range of koalas is restricted to Australia and does not overlap with that of gibbons, which are endemic to Southeast Asia. The lack of host sympatry suggests that an intermediate host with a less restricted range is responsible for GALV and KoRV CST (9, 14-16).

Mobile species such as rodents, bats, or birds have been proposed as potential intermediate hosts of GALV and KoRV (9, 14). Bats can fly and disperse rapidly; they have been linked to the spread of several zoonotic diseases (17) and some Southeast Asian species harbor retroviruses related to GALV and KoRV (18). Rodents, however, are plausible intermediate hosts as they have migrated from Southeast Asia to Australia multiple times with several Southeast Asian species having established themselves in Australia (19). Furthermore, endogenous retroviruses related to GALV have been reported to be present in the genome of several Southeast Asian rodents such as *Mus caroli*, *Mus cervicolor* and *Vandeleuria oleracea* (20-22). However, these reports were based on DNA hybridization techniques and sequences were not reported. In 2008, the full genome sequence of an endogenous retrovirus found in the genome of *Mus caroli* (McERV) was reported (23). Despite the relatively high similarity to the genomic sequences of GALV and KoRV, McERV has a different host range and uses a different receptor, and therefore it is unlikely a progenitor of GALV and KoRV (23). McERV is most closely related to *Mus dunni* endogenous virus (MDEV) (24) and the *Mus musculus* endogenous retrovirus (MmERV) (25), which together form a sister clade to the KoRV/GALV clade (2). Recently Simmons et al. (16) discovered fragments belonging to a retrovirus closely related to GALV and KoRV in the Australian native rodent *Melomys burtoni* (MbRV). MbRV sequence share 93 and 83% nucleotide identity with GALV and KoRV respectively, and *Melomys burtoni* overlaps with the geographic distribution of koalas. However, *Melomys burtoni* is currently not present in Southeast Asia. Consequently it is unlikely that MbRV represents the ancestor virus of KoRV and GALV, and *Melomys burtoni* is unlikely the intermediate host of GALV or KoRV(16).

The aim of this work was to screen a wide range of rodent species from Southeast Asia for the presence of KoRV and GALV-like sequences and characterize polymorphisms in their viral receptor proteins in the attempt to identify the intermediate host(s) of KoRV and GALV using a non-PCR based approach called hybridization capture (26, 27). We focused on Southeast Asian rodent species since 42 Australian vertebrate species were previously screened, with MbRV the only virus identified (16), and most of the rodent species with GALV-like sequences identified are from Southeast Asia suggesting that GALVs and KoRVs may be circulating naturally in rodent populations residing there. Twenty-six rodent species were screened of which only one species (*Melomys* sp., a newly identified

species in the process of being described) was positive for a GALV sequence distinct from MbRV and none were positive for KoRV-like sequences. We report the complete nucleotide sequence of the identified GALV-like virus, which we term *Melomys* Woolly Monkey Virus (MelWMV), its genomic structure, and its phylogenetic relationships with other related gammaretroviruses. We also examine GALV receptor variation among permissive and restrictive hosts including species belonging to the genus *Melomys*.

MATERIALS AND METHODS

Sample collection

For the screening for GALV and KoRV, tissue samples from Johan Michaux (details? Conserved in ethanol? date or period of collection? what kind of sample were sent to us - skin, muscle? who collected them?). All 49 samples belonging to the 26 species analyzed in the current study are listed in table 1. For the sequecing of the receptor of GALV, also a blood sample was collected from a male white-handed gibbon (*Hylobates lar*) from Nuremberg zoo, Germany, during a routine health check on 24th July 1996.

Ethics statement

All animal experiments were performed according to the directive 2010/63/EEC on the Protection of Animals Used for Experimental and Other Scientific Purposes. The animal work also complied with the French law (nu 2012–10 dated 05/01/2012 and 2013-118 dated 01/02/2013). The rodents were captured using Sherman traps and the study of the species used in this project did not require the approval of an ethics committee (European directives 86-609 CEE and 2010/63/EEC). The species used are not protected, and no experiment was performed on living animals. No permit approval was needed as the species were trapped outside any preserved areas (national parks or natural reserves). The rodents were euthanized by vertebrate dislocation immediately after capture at in agreement with the legislation and the ethical recommendations (2010/63/EEC annexe IV) (see also protocol available on http://www.ceropath.org/references/rodent_protocols_book). All experimental protocols involving animals were carried out by qualified personnel (accreditation number of the Center of Biology and Management of the Populations (CBGP) for wild and inbred animal manipulations: A34-1691). For the samples from Laos and Thailand, approval notices for trapping and investigation of rodents were provided by the Ministry of Health Council of Medical Sciences, National Ethics Committee for Health Research (NHCHR) Lao PDR, number 51/NECHR, and by the Ethical Committee of Mahidol University, Bangkok, Thailand, number 0517.1116/661. Oral agreements for trappings from obtained for local community leaders and land owners.

Cell lines, viruses and DNA extraction

GALV DNA for hybridization capture bait generation (26, 27) was obtained from the following productively infected cell lines: SEATO-88, GALV-SEATO infected Tb 1 Lu bat lung fibroblasts (ATCC CCL-88); GALV-4-88, GALV-Brain infected Tb 1 Lu bat lung fibroblasts (ATCC CCL-88); 71-AP-1, WMV infected marmoset fibroblasts; 6G1-PB, GALV-Hall´s Island infected lymphocytes; HOS (ATCC CRL-1543) GALV-SF infected human osteosarcoma cells. Genomic DNA extraction from the cell lines was performed using the Wizard Genomic DNA Purification Kit (Promega), following the manufacturer's protocol. Rodent tissue samples were first homogenized using a Precellys 24 (Bertin Technologies), with genomic DNA then extracted using the QIAamp DNA mini kit (QIAGEN) according to manufacturer's instructions. The genomic DNA of the white-handed gibbon was extracted following the method described in Sambrook and Russell (28). For all

DNA extracts, DNA concentration was determined using the dsDNA High Sensitivity Assay Kit on a Qubit 2.0 fluorometer (Invitrogen).

Illumina library preparation

All rodent sample DNA extracts were sheared using a Covaris S220 (Covaris) to an average size of 300 bp prior to building Illumina sequencing libraries. Libraries were generated as described in Meyer and Kircher (29) with the modifications described in Alfano et al. (30), except for using a variable starting amount of DNA extract according to each sample availability and using 1 µl Illumina adapter mix (20 µM) in the adapter ligation step. Each library contained a unique combination of index adapters, one at each end of the library molecule (double-indexing) (31), to allow for subsequent discrimination among samples after the sequencing of pooled libraries. Negative control extraction libraries were also prepared and indexed separately to monitor for experimental cross contamination. Each library was amplified in three replicate reactions to minimize amplification bias in individual PCRs. The amplifications of the libraries were performed using Herculase II Fusion DNA polymerase (Agilent Technologies) in 50 µl volume reactions, with the cycling conditions of 95°C for 5 min, followed by 7 cycles of 95°C for 30 s, 60°C for 30 s, 72°C for 40 s and finally 72°C for 7 min. After pooling the three replicate PCR products for each sample, amplified libraries were purified using the QIAquick PCR Purification Kit (QIAGEN) and quantified using a 2200 TapeStation (Agilent Technologies) on D1K ScreenTapes. Additional amplification cycles were performed for some of the libraries, when needed to balance library concentrations, using Herculase II Fusion DNA polymerase with P5 and P7 Illumina library outer primers with the same cycling conditions.

Hybridization capture baits

Two different approaches were used to amplify the genomes of GALV and KoRV for hybridization capture bait production (26, 27). The KoRV genome was amplified in thirty-eight 500 bp overlapping products as described in Tsangaras et al. (27) using the DNA of a northern Australian koala (PCI-SN248) from the San Diego Zoo. The thirty-eight amplicons were then pooled in equimolar ratios. By contrast, the genomes of the five isolated GALV strains (SEATO, SF, Brain, Hall's Island, WMV) were amplified in two ca. 4.3 kb-long overlapping PCR products using primers designed on an alignment of the recently published genomes of the GALV strains (accession numbers KT724047-51) (2). The amplicons were produced from five different GALV-infected cell lines. Primers U5 (5'- CAGGATATCTGTGGTCAT -3') and PolR1 (5'- GTCGAGTTCCAGTTTCTT -3') amplify the first 4.3 kb of the GALV genome (5' LTR, *gag* and part of *pol* gene) and primers PolF1 (5'- CTCATTACCAGAGCCTGCTG -3') and U3 (5'- GGATGCAAATAGCAAGAGGT -3') the second 4.3 kb (part of *pol* gene, *gag* gene and 3' LTR). Primer U3_SEATO (5'- GGATGCAATCAGCAAGAGGT -3') was used instead of primer U3 for the SEATO strain to account for two nucleotides difference existing in that region for GALV-SEATO. The GALV PCRs were performed in a volume of 23 µl using approximately 200 ng of DNA extract, 0.65 µM final concentration of each primer, 12.5 µl of 2× MyFi Mix (Bioline) and sterile distilled water. Thermal cycling conditions were: 95°C for 4 min; 35 cycles at 95°C for 30 s, 54-62°C (based on best PCR product yield per strain determined empirically) for 30 s, 72°C for 6 min; and 72°C for 10 min. An aliquot of each PCR product was visualized on 1.5% w/v agarose gels stained with Midori Green Direct (Nippon Genetics Europe). PCR products were purified using the MSB Spin PCRapace kit (STRATEC Molecular GmbH), quantified using a Qubit 2.0 fluorometer (Invitrogen) and Sanger-sequenced at LGC Genomics (Berlin, Germany) to verify that the correct target had been amplified. The PCR products from each GALV strain were then pooled in equimolar concentrations and sheared to obtain a fragment size of approximately 350 bp using a Covaris S220. The mixed sheared GALV amplicons were then pooled with the mixed KoRV amplicons at a 1:6 KoRV:GALV ratio to balance

the one KoRV amplicon set with the 5 GALV strains in the final bait pool. The GALV-KoRV mixed amplicons were then blunt ended using the Quick Blunting Kit (New England Biolabs), ligated to a biotin adaptor using the Quick Ligation Kit (New England Biolabs), and immobilized in separated individual tubes on streptavidin coated magnetic beads as described previously (26).

Hybridization capture

The 50 rodent indexed libraries were pooled in groups of 5 in order to reach a library input of 2 µg for each capture reaction. The negative controls for library preparation were also included in the capture reactions. Each indexed library pool was mixed with blocking oligos (200 µM) to prevent crosslinking of Illumina library adapters, Agilent 2× hybridization buffer, Agilent 10× blocking agent, and heated at 95°C for 3 min to separate the DNA strands (26). Each hybridization mixture was then combined with the biotinylated bait bound streptavidin beads. Samples were incubated in a mini rotating incubator (Labnet) for 48 hours at 65°C. After 48 hours the beads were washed to remove off-target DNA as described previously (26) and the hybridized libraries eluted by incubating at 95°C for 3 min. The DNA concentration for each captured sample was measured using the 2200 TapeStation on D1K ScreenTapes and further amplified accordingly using P5 and P7 Illumina outer primers (29). The enriched amplified libraries were then pooled in equimolar amounts to a final library concentration of 4.5 nM for paired-end sequencing (2×250) on an Illumina MiSeq platform with the v2 reagents kit at the Berlin Centre for Genomics in Biodiversity Research (BeGenDiv).

Genome sequence assembly

A total of 12,502,407 paired-end sequence reads 250-bp long were generated (average = 250,046.8 paired-end reads per sample, SD = 113,859.9) and sorted by their double indexes sequences. Cutadapt v1.2.1 (32) and Trimmomatic v0.27 (33) were used to remove adaptor sequences and low-quality reads using a quality cutoff of 20 and a minimal read length of 30 nt. After trimming, 97.6% of the sequences were retained. Thereafter reads were aligned to the NCBI nucleotide database using BLASTn (34) and the taxonomic profile of BLAST results were visualized using Krona (35) in order to assess the taxonomic content of the captured libraries. Reads were then mapped to the genome sequences of GALV strains (KT724047-51), KoRV (AF151794) and closely related gammaretroviruses (McERV - KC460271; MDEV - AF053745; MmERV - AC005743) using BWA v0.7.10 with default parameters (BWA-MEM algorithm)(36). The alignments were further processed using Samtools v1.2 (37) and Picard (http://broadinstitute.github.io/picard) for sorting and removal of potential duplicates, respectively. Mapping was used as a preliminary screen to identify samples potentially positive or negative for viral sequences. Only samples that produced reads mapping across the genome of a viral reference were considered positive and subjected to further analyses. Samples that exhibited reads mapping only to limited portions of the reference, likely due to random homology of part of the bait to host genomic regions, were not further considered. Reads from positive samples were mapped to the reference of interest and the resulting alignments visualized and manually curated using Geneious v7.1.7 (http://www.geneious.com; Biomatters, Inc.).

PCR amplifications

Two primer pairs based on the GALV consensus sequences generated from the hybridization capture data were designed to fill in gaps found in the bioinformatics assembly. Primers GagF1 (5'-TGAGTAGCGAGCAGACGTGTT-3') and GagR1 (5'-GGCAAAATCACAGTGGAGTCA-3') were used to amplify a region encompassing part of the *gag* gene and the interspace fragment between 5' LTR and *gag*, while primers EnvF1 (5'-CAGTTGACCATTCGCTTGGA-3') and EnvR1 (5'-CCGAGGGTGAGCAACAGAA-3') were used to amplify part of the *env* gene. The PCR reaction mix comprised 12.5 µl of 2× MyFi Mix (Bioline), 0.6 µM final concentration of forward primer, 0.6 µM

final concentration of reverse primer, approximately 100 ng of DNA template and sterile distilled water to a final volume of 22 µl. Thermal cycling conditions were: 95°C for 3 min; 40 cycles at 95°C for 15 s, 59°C for 20 s, 72°C for 30 s; and 72°C for 30 s. For EnvF1-EnvR1, the annealing temperature was set to 61°C instead of 59°C, and the extension time to 40 s instead of 30 s.

Five primer sets were designed based on the alignment of the phosphate transporter 1 (*PiT1* or *SLC20A1*) and the phosphate transporter 2 (*PiT2* or *SLC20A2*) sequences available in GenBank of *Mus musculus*, *Rattus norvegicus*, *Cricetulus griseus*, *Homo sapiens*, *Macaca mulatta*, *Nomascus leucogenys* to sequence the region A of *PiT1* and *PiT2* from *Hylobates lar*, *Melomys* sp., *Melomys paveli* and *Mus caroli*. Primers PiT1-F1long (5'-AGATCCTTACAGCCTGCTTTGG-3') and PiT1-R1 (5'-TCCTTCCCCATRGTCTGGAT-3') were designed to amplify a region approximately 600-bp long and encompassing the exons 7 and 8 of *PiT1* – which contains region A – compared to *M. musculus* sequence (800-bp long and targeting exons 8 and 9 compared to *H. sapiens* sequence). Primers PiT1-F1short (5'-CCTCTGGTTGCTTTGTATCTTGTT-3') for the rodent templates and PiT1-F1short_apes for the gibbon template (5'-GGCCTCTGGTTGCTTTATATTTG-3'), both in combination with the above mentioned PiT1-R1, were designed to amplify a 150-bp long fragment including region A. Two primer pairs – PiT2-F1 (5'-TGCTATTGGTCCCCTTGTGG-3') and PiT2-R1 (5'-CCCCAAACCCAGAGACCTGT-3') for the rodents, and PiT2-F1_apes (5'-CCTGGTAGCCTTGTGGCTGA-3') and PiT2-R1_apes (5'-TGATGGGAGTGAGGTCCTTC-3') for the gibbon – were designed to amplify a fragment approximately 150-bp long including PiT2 region A. The PCRs were performed using approximately 100 ng of DNA extract, 0.6 µM of final concentration of each primer, 12.5 µl of 2× MyFi Mix (Bioline) and sterile distilled water to a final volume of 22 µl. Cycling conditions were: 95°C for 3 min; 35 cycles at 95°C for 15 s, 57°C for 20 s, 72°C for 10 s; and 72°C for 10 s. For PiT1-F1long and PiT1-R1, the extension at 72°C was prolonged to 30 s.

An aliquot of each PCR product was visualized on 1.5% w/v agarose gels stained with Midori Green Direct (Nippon Genetics Europe). PCR products were purified using the MSB Spin PCRapace kit (STRATEC Molecular GmbH), quantified using a Qubit 2.0 fluorometer (Invitrogen) and Sanger-sequenced at LGC Genomics (Berlin, Germany). Sequences were then screened against the NCBI nucleotide database using the BLAST online search tool (https://blast.ncbi.nlm.nih.gov/).

Evolutionary analyses

To characterize the phylogenetic relationships among the identified viral consensus sequences, the known GALV strains, MbRV and other related gammaretroviruses, phylogenetic trees were inferred based on the viral nucleotide sequences. The following reference sequences were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/GenBank): GALV-SEATO (KT724048), GALV-SF (KT724047), GALV-Brain (KT724049), GALV-Hall's Island (KT724050), woolly monkey virus (WMV; KT724051), *Melomys burtoni* retrovirus (MbRV; KF572483-6). KoRV (AF151794) was used as an outgroup. Genomic sequences and individual gene (*env*, *gag*, and *pol*) sequences were aligned using MAFFT (38). Phylogenetic analysis was performed using the maximum-likelihood (ML) method available in RAxML v8 (39), including 500 bootstrap replicates to determine the node support. The general time-reversible substitution model (40) with among-site rate heterogeneity modeled by the Γ distribution and four rate categories (41) were used. Nucleotide sequences of *env*, *gag*, and *pol* were concatenated and analyzed in a partitioned framework, where each partition was allowed to evolve under its own substitution model. In order to infer the phylogenetic trees, the nucleotide sequences of *env*, *gag*, and *pol* were both analyzed separately and concatenated including noncoding LTRs and spacers and analyzed in a partitioned framework.

RESULTS

Screening for GALV and KoRV in rodents using hybridization capture

Twenty-six rodent species (1-6 individuals per species) were screened for the presence of KoRV- and GALV-like sequences (table 1). None of the 26 species yielded sequences mapping to KoRV. Only the six samples belonging to an Indonesian *Melomys* species that has not yet been described in the literature produced reads mapping uniformly across the genome of the woolly monkey virus (WMV), which is considered a strain of GALV. All of the tested species of *Mus* produced sequence reads mapping to one of the GALV-related murine retroviruses (MmERV, McERV, MDEV). These sequences were likely captured by GALV/KoRV baits based on the homology of these ERVs with GALV and KoRV. Specifically, we recovered portions of the genome of MmERV from the samples belonging to *Mus musculus*. *Mus nitidulus* and *Mus booduga* samples demonstrated the presence of a virus similar to MmERV. We also detected sequences similar to McERV in *Mus caroli*, *M. cervicolor*, *M. cookii*, *M. fragilicauda* and *M. lepidoides*.

Melomys woolly monkey virus (MelWMV)

Seven *Melomys* spp. samples were screened, of which six were from a new species of *Melomys* from Indonesia which is in the process of being described (Fabre et al. unpublished data) (here referred to as *Melomys* sp.). In addition, a sample of *Melomys paveli* from Seram Island (Maluku Province, Indonesia) was included. Only *Melomys* sp. yielded GALV-like sequences, with reads mapping to the woolly monkey virus (WMV) detected in all six *Melomys* sp. samples. For most of the samples only few reads were found: from a minimum of 24 to a maximum of 1,008 mapping reads, but in each case distributed evenly across WMV genome. However, in sample WD279 almost full coverage of the viral genome was obtained with an average per-base coverage of 18×. The enrichment (proportion of on-target reads mapping to WMV) was low (below 1%) in all samples, similarly to our previous experiments (2). The negative control generated few sequence reads, none mapping to GALV.

Two primer sets (GagF1-GagR1 and EnvF1-EnvR1) based on the mapped reads were designed to fill gaps in the assembly to WMV. The generated PCR products were used both to complete the viral genomic sequence and to confirm the bioinformatics assembly of the sequences obtained by hybridization capture. Primers EnvF1-EnvR1 were specifically designed to cover a gap in the assembly in the *env* gene of the virus, but the resulting Sanger sequences confirmed that this portion of *env*, corresponding to positions 6,777 to 7,758 in the WMV sequence, is not present in the viral genome. A schematic representation of the genome assembly based on captured sequences and of the PCR products is shown in Fig. 1.

The primers were applied to the *Melomys paveli* sample as well and confirmed the absence of GALV-like sequences suggested by the hybridization capture experiment. Identical amplification products from each primer set were produced for all 6 *Melomys* sp. samples. Based on the Sanger sequences and the hybridization capture Illumina reads, we determined that the viral sequences were identical in the 6 *Melomys* sp. samples. The identified virus was characterized by the common genetic structure of simple type C mammalian retroviruses with a 5' LTR-*gag*-*pol*-*env*-3' LTR organization (Fig. 1). The 5' and 3' LTRs were identical. Nevertheless, the virus lacked approximately 60% of *pol*, with the whole reverse transcriptase domain missing, and almost half of the surface unit gp70 (SU) and most of the transmembrane subunit p15E (TM) of *env* (Fig. 1). The remaining protein domains of Pol – the protease (PR) and integrase (IN) – and all Gag protein domains – the matrix p15 (MA), p12,

capsid p30 (CA), and nucleocapsid p10 (NC) – were intact. However, the ORF of *gag* was truncated by a premature stop codon. Therefore, the Gag protein was 324 amino acids long, instead of the 521 residues expected for WMV. The same regulatory motifs found in WMV and in the other GALVs (2) were identified: a tRNAPro primer binding site, a CAAT box, a TATA box, a Cys-His box, a polypurine tract, and a polyadenylation signal (Fig. 1). Furthermore, no differences between MelWMV and WMV were observed in the domains known to affect GALV and KoRV differential infectivity: the CETTG motif (42) of the Env protein (residues 167 to 171) and the PRPPIY and PPPY motifs (42, 43) of the Gag protein (residues 123-128 and 140-143). In addition, MelWMV showed high levels of conservation compared to WMV in the variable regions A and B (VRA and VRB) of the Env protein (residues 86-153 and 192-203, respectively), which are known to influence receptor specificity (44): only 6 out of 80 residues were polymorphic between the two viruses.

The integration sites, which were captured for 4 out 6 *Melomys* sp. samples, were identical in each sample. Only a single 5' and 3' integration site was found. The genomic sequences of *Melomys* sp. flanking MelWMV 5' and 3' integration sites were queried by BLAST against the NCBI nucleotide database and returned a hit to BAC clone RP23-133I8 from chromosome 1 of *Mus musculus* (accession AC124760), the closest relative of *Melomys* sp. with genome sequence available in GenBank. 5' and 3' flanking sequences were found to match contiguous regions of the genome of *Mus musculus*, suggesting that the two flanks correspond to genomic sequence of *Melomys* sp. on either side of the integration site of MelWMV. Comparing the 5' and 3' host genomic flanks also allowed the identification on both sides of the provirus of the target site duplication, a segment of host DNA that is replicated during retroviral integration and that appears as an identical sequence immediately upstream and downstream of the integrated provirus. The duplicated sequence for MelWMV was "GTCAC" flanking both the 5' and 3' ends of the virus.

The newly identified virus shared 98% nucleotide identity with WMV. A phylogenetic analysis was performed including sequences from the genomes of the GALV strains, the *Melomys burtoni* retrovirus (MbRV) and KoRV as an outgroup. The new virus formed a sister taxon to WMV, which together formed a monophyletic group with MbRV (Fig. 2). These three viruses in turn constituted a sister clade to the other GALV strains. The evolutionary relationship between the new virus and WMV was well-supported (bootstrap 88 – 91%) using both concatenated partitioned nucleotide sequences (Fig. 2) and *gag* and *env* nucleotide sequences (Suppl. Fig. 1; Suppl. Fig. 3). Therefore the new virus can be considered a strain of GALV and is here designated *Melomys* woolly monkey virus (MelWMV). Lower support was found using *pol* nucleotide sequences (Suppl. Fig. 2), likely due to the large deletion of the gene in MelWMV, which reduced the number of phylogenetically informative sites. The support for the relationship among the WMV-MelWMV clade and MbRV was not very robust (bootstrap 61 – 75%) since only partial sequences of *pol* and *env* were recovered for MbRV (Fig. 2; Suppl. Fig. 2-3).

Sequencing of region A of PiT1 and PiT2

Residues present in the C-terminal region of the fourth extracellular domain of PiT1, the receptor used by GALV to infect host cells (45), have been identified as critical for receptor function and therefore GALV infection (46-49). This nine-residue region, designated region A, has been extensively analyzed by mutational analysis and by comparative alignment of PiT1 orthologs that function as GALV receptor to PiT1 orthologs that fail to support GALV entry. Substitution of region A residues of PiT1 for the corresponding residues of two proteins that do not support GALV entry, Pit2 (PiT1 paralog) (49) and the distantly related phosphate transporter Pho-4 from the filamentous fungus *Neurospora crassa* (48), renders these proteins functional as GALV receptors. Five primer sets were designed to sequence region A of *PiT1* and *PiT2* from *Hylobates lar*, *Melomys* sp., *Melomys*

paveli and *Mus caroli*. PiT2 was also sequenced since it is used by GALV to infect Chinese hamster and Japanese feral mouse cells (47, 50). An amplification product was obtained from each of the five primer sets. Sanger sequencing of the amplicons and the subsequent BLAST search confirmed the amplification of the region A of *PiT1* and *PiT2*. The sequences were then aligned with the reference sequences of *Mus musculus*, *Rattus norvegicus*, *Cricetulus griseus*, *Homo sapiens*, *Macaca mulatta* and *Nomascus leucogenys* available in GenBank and translated into amino acid sequences. The amino acid sequences were then aligned and compared with the amino acid sequences of region A of PiT1 and PiT2 of all the species known to be permissive (*Homo sapiens*, *Rattus norvegicus*, *Mus musculus molossinus*, *Cricetulus griseus*) or resistant (*Mus musculus musculus* and *Mus dunni*) to GALV infection according to the literature (table 2) (45, 47, 50-52).

Region A of PiT1 and PiT2 is comprised of residues at positions 550-558 and 522-530 respectively. Positions 550 and 553 of PiT1, and 522 and 529 of PiT2 are crucial for receptor function (47-49). Functional GALV receptors have an acidic residue, either Asp(D) or Glu(E), at one or both of these positions. However, a Lys(K) at position 550 (522 in PiT2) is known to abrogate receptor function (47, 53). The PiT1 sequence of *M. caroli* had an Asp(D) at position 553 but also a Lys(K) at position 550, and overall it was identical to that of *M. dunni*, the cells of which are resistant to GALV infection (52). The sequence of PiT2 was identical to that of *Mus musculus molossinus* which serves as a functional GALV receptor (52): they both have a Gln(Q) at position 522, but a Glu(E) at position 529. The sequence of *H. lar* PiT1 region A had an Asp(D) at both positions 550 and 553, and was identical to the human sequence (45), whereas PiT2 displayed one amino acid difference – Thr(T) to Met(M) at position 527 – when compared to human (51). Both human cells and gibbons are permissive to GALV infection, but human PiT2, which has a Lys(K) at positions 522, like gibbon PiT2, does not function as a GALV receptor. The sequence of PiT1 region A of *Melomys* sp. was very similar to the sequence carried by susceptible species such as rats, humans, gibbons and *Mus musculus molossinus*. *Melomys sp.* had a Glu(E) at position 550 and an Asp(D) at position 553, identical to rat. The Thr(T), Val(V) and Lys(K) at positions 551, 554 and 557 respectively were invariant among *Melomys* sp. and the other permissive species, with the Lys(K)-557 shared with both resistant and permissive species. The residues at positions 555, 556 and 558 of PiT1 varied randomly among resistant and susceptible species, while residue 552 was missing in the resistant ones. The PiT2 sequence of *Melomys* sp. had a Glu(E) at position 522 and differed in only one residue – Met(M) to Thr(T) at position 527 – compared to *C. griseus* (54), which is also susceptible to GALV infection. The sequence was identical to *Mus musculus molossinus* PiT2, which is also considered a functional GALV receptor (52). The PiT1 and PiT2 region A sequences of *Melomys paveli* were almost identical to *Melomys* sp., but the PiT1 sequence of *Melomys paveli* lacked the residue – a Gly(G) in *Melomys* sp. – at position 552, like in the resistant species.

DISCUSSION

KoRV and GALV are closely related retroviruses (2). However, their respective hosts, koalas and gibbons, share neither a recent common ancestor nor overlapping geographic distributions. Thus, KoRV and GALV may have arisen from a cross-species transmission that involved an intermediate host (9, 14-16). In order to identify such a vector, Simmons et al. (16) screened 42 Australian vertebrate species (birds and mammals including rodents and bats) for KoRV and GALV-like sequences. An ERV closely related to GALV was found in an Australian murid species (*Melomys burtoni*), but, even if related to GALVs, particularly WMV, it does not represent an ancestor of GALV or KoRV because the distribution of the genus *Melomys* and gibbons do not currently overlap (16). Because GALV-like viruses have been identified in Southeast Asian rodents (20, 21, 55), we screened

rodent species from this geographic area in the attempt to identify potential intermediate hosts and retrieve ancestral viral strains of KoRV and GALV. Twenty-six rodent species were screened (Tab.1). Some of the species tested (*Bandicota savilei*, *Bandicota indica*, *Bandicota bengalensis*, *Berylmys berdmorei*, *Mus musculus*) had been reported as negative for GALV and KoRV by Simmons et al. (16), consistent with the absence of GALV and KoRV from the Southeast Asian samples from the same species in this study. None of the species tested in the current study or in Simmons et al. (16) was positive for KoRV-like sequences, while only *Melomys burtoni* from Australia and *Melomys* sp. from Indonesia were found positive for GALV-like sequences in Simmons et al. (16) and in the current study, respectively. Based on the homology (97%) and phylogenetic affinity, MelWMV is a subtype of WMV whereas MbRV is a sister taxon (Fig. 2; Suppl. Fig. 1-3).

Only one integration site was found for MelWMV. Therefore there may be only a single copy of MelWMV in the genome of *Melomys* sp., and this would explain the low hybridization capture coverage. Furthermore, MelWMV was detected in all 6 individuals of *Melomys* sp. tested and the integration site was identical in all 4 individuals for which they were identified by hybridization capture. This result, the premature stop codon in *gag* and the deletions in *pol* and *env* (Fig. 1) strongly indicate that MelWMV is an endogenous retrovirus. Furthermore, we hypothesize that MelWMV has recently integrated in the genome of *Melomys* sp., based on the identical 5' and 3' LTR sequences (56) and its absence from *M. burtoni* and *M. paveli* which diverged from a common ancestor X million years ago.

MelWMV along with WMV and MbRV represent the basal clade of the GALV phylogeny, so it can be argued that the WMV-like viruses are the most ancestral GALV strains currently known to be circulating and most likely the closest viruses to the progenitor of GALV and KoRV. The only species shown to have such close GALV relatives out of 68 total species tested in Australia (16) and SE Asia belong to the murine genus *Melomys*. Furthermore, more distantly related GALV-like ERVs are found in rodents belonging to the genus *Mus* (20, 55). Taken together, this suggest an overall rodent origin of the clade. However, since MelWMV is an ERV in *Melomys* sp. but *M. paveli* did not yield any GALV-like sequences, it is not clear whether *Melomys* is a reservoir or a susceptible host for GALVs. Thus, it is formally possible that GALV did not originate in *Melomys* and some of the *Melomys* species (*Melomys burtoni*, *Melomys* sp. from this study) were independently infected with GALV in Indonesia and Australia from an unknown reservoir species. As the vast majority of samples in the current study were from Southeast Asia and those of Simmons et al. (16) exclusively from Australia, Indonesia and Papua New Guinea remain largely unexplored. In addition, only three species of *Melomys* have been tested out of a total of 23 *Melomys* species, 20 of which are found in Indonesia and Papua New Guinea (IUCN 2015. *The IUCN Red List of Threatened Species. Version 2015-4*. http://www.iucnredlist.org), suggesting that many more GALVs, including potentially exogenous GALVs, and possibly KoRV-like sequences may be present. Of particular relevance to the current host range of GALV, *Melomys* species are found in both Australia and Southeast Asia which connects them to their accidental hosts, gibbons and koalas. However, the genus *Melomys* is (currently) not present in mainland Southeast Asia, where the gibbon isolates of GALV were identified. Therefore, it is still not clear how the virus moved from Australia and Indonesia to mainland Southeast Asia crossing the Wallace line. Gibbons in particular are surprising hosts as GALVs have only been isolated from captive and not wild gibbons suggesting they have had infrequent but regular contact with a GALV reservoir or host species but only in captive facilities.

GALV infects cells using a ubiquitous transmembrane protein that functions as a sodium-dependent phosphate transporter called PiT1 or SLC20A1 (45). GALV can alternatively infect cells using a related phosphate transporter, PiT2 or SLC20A2, originally recognized as the amphotropic murine leukemia virus (A-MuLV) and 10A1 MuLV receptor, to infect Chinese hamster and Japanese

feral mouse cells (47, 50, 51). This similarity of receptor usage is consistent with the phylogenetic relationship of GALVs and MuLVs, which belong to the same overall retroviral group (2).

Mutagenesis studies have shown that region A of PiT1, a stretch of nine residues corresponding to position 550-558 of human PiT1, which is highly polymorphic among species, is crucial for GALV entry into cells (46, 47). Because of its highly polymorphic nature, it is not clear which of the residues of region A are essential for GALV infection. Schneiderman et al. (47) had suggested that the functional GALV receptors have an acidic residue at either position 550 or 553 of PiT1 (522 or 529 of PiT2) or both, but lysine at position 550 (522 in PiT2) abrogates GALV receptor function, even when an acidic residue is present at position 553 (529 in PiT2). A subsequent study (53) demonstrated that PiT1 and PiT2 can serve as receptors for GALV when lysine is absent from the first position, regardless of the presence of acidic residues at the above mentioned positions. We have sequenced PiT1 and PiT2 region A from species resulted both positive (*Melomys* sp.) and negative (*Melomys paveli* and *Mus caroli*) to our GALV screening, and also from *Hylobates lar*, another natural host of GALV. When comparing with the previously reported sequences of species both permissive (human *Homo sapiens*, rat *Rattus norvegicus*, Japanese feral mouse *Mus musculus molossinus*, Chinese hamster *Cricetulus griseus*) and resistant (*Mus musculus*, *Mus dunni*) to GALV infection (table 2), the sequences generated here were consistent with the findings of previous functional studies (46, 47, 53). Positions 551-2 and 554-8 of PiT1 are not critical determinants of receptor function. All permissive species have a Thr(T) and a Val(V) at positions 551 and 554, whereas resistant species have a Gln(Q) and Ala(A) respectively. However, these positions in PiT1 may not be crucial as PiT2 of both resistant and permissive species have a Gln(Q) and a Val(V) at positions 523 and 526 respectively, which correspond to residues 551 and 554 of PiT1. Positions 555, 556 and 558 of PiT1, which varied randomly among resistant and susceptible species, and the Lys(K) at position 557, which was present in all species, are unlikely to be determinants of GALV susceptibility.

In contrast, positions 550 and 553 of PiT1 may play a key role, as previously proposed by Schneiderman et al. (47). All permissive species have an acidic residue – Asp(D) or Glu(E) – at either position 550 or 553 of PiT1. In PiT2 an acidic residue is found at either position 522 or 529 among permissive species. A Lys(K) is present at the first position, 550 of PiT1 or 522 of PiT2, in all resistant species and therefore it is likely to be the residue which determines the resistance to GALV infection. Therefore, the *Mus caroli* PiT1 sequenced in this study, which has a Lys(K) at position 550 and is identical to *Mus dunni* in region A, is unlikely to serve as a GALV receptor. This is consistent with the absence of any GALV-like sequence in this species. McERV sequences were detected but this virus uses a different receptor than GALV (23). However, GALV could potentially infect *Mus caroli* using PiT2, since *Mus caroli* PiT2 sequence is identical to that of *Mus musculus molossinus* PiT2 that is a functional GALV receptor. Regions A of human and gibbon PiT1 are identical, and both humans and gibbons have a Lys(K) at the first position of PiT2 region A. Human PiT1 functions as GALV receptor, while PiT2 does not. Given the similarity between human and gibbon PiT receptors captive gibbons were likely infected via PiT1.

Both PiT1 and PiT2 of *Melomys* sp. are potentially functional GALV receptors, consistent with our discovery of MelWMV in this species. However, MelWMV and WMV are highly similar in the VRA and VRB domains of the envelope, and WMV is known to be unable to use the PiT2 receptor to infect hamster cells due to a block mediated by WMV envelope, specifically VRA and VRB (44). Therefore, it is likely that *Melomys* sp. was infected by WMV via the PiT1 receptor. *Melomys paveli* is also potentially susceptible to GALV infection, since its PiT1 and PiT2 region A are identical to *Melomys* sp., with the exception that residue 552 is missing in PiT1, as observed in resistant species (*Mus musculus musculus*, *Mus dunni*). Since the lack of this residue was never taken into account as a

determinant of resistance to GALV in former functional studies, we cannot draw conclusions on the effect of this deletion on receptor functionality. However, we only detected GALV in *Melomys* sp.. As only one *Melomys paveli* sample was analysed we cannot rule out that GALVs may be circulating at low abundance in this species. Furthermore, it is also possible that *M. paveli* never came into contact with a GALV, since its distribution is restricted to Seram Island. Therefore, the absence of GALV may be biogeographically determined rather than driven by a receptor restriction for this species.

In conclusion, our screen of Southeast Asian rodents identified MelWMV in a *Melomys* species from Indonesia. MelWMV represents the most closely related retrovirus to GALV identified from rodents to date and the second record of a GALV relative identified from the *Melomys* genus, suggesting that either *Melomys* is a host of GALVs or several species within the genus are sympatric with the reservoir. The PiT1 and PiT2 region A sequences of the *Melomys* species tested in the current study are consistent with the general susceptibility of these species to GALV infection. Further screening of GALV and KoRV in *Melomys* across the range of this genus would be promising for identifying additional GALV sequences.

ACKNOWLEDGEMENTS

The authors wish to thank Karin Hönig (IZW) and Susan Mbedi (BeGenDiv) for sequencing support. Y.I., A.L.R., K.M.H. and A.D.G. were supported by Grant Number R01GM092706 from the National Institute of General Medical Sciences (NIGMS). M.V.E. contribution was supported by National Institute of Mental Health Intramural Research Program Project 1ZIAMH002592.The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the National Institutes of Health. N.A. was supported by the International Max Planck Research School for Infectious Diseases and Immunology (IMPRS-IDI) at the Interdisciplinary Center of Infection Biology and Immunity (ZIBI) of the Humboldt University Berlin (HU). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Data accession

The complete sequence and annotations of MelWMV genome was deposited in GenBank under accession number XXX. Illumina reads mapping to WMV for each (or only the sample with most reads?) *Melomys* sp. sample were deposited in the NCBI Sequence Read Archive as BioProject PRJNAXXX.

FIGURE LEGENDS

Figure 1. MelWMV genomic assembly and structure. Alignment of WMV and MelWMV consensus sequence generated from hybridization capture data combined with the PCR products that were produced to fill in the gaps in the bioinformatics assembly, shown as continuous black bars. Nucleotide positions identical among the strains are indicated in light grey, while mismatches are shown in black. Gaps are shown as dashes. The green bar above the alignment indicates the percent identity among the sequences (green: highest identity, red: lowest identity). The positions of proviral genes (*gag*, *pol* and *env*) and protein domains of WMV are indicated in yellow and sky blue respectively, and are used as reference also for MelWMV. The truncated ORF of MelWMV *gag* is indicated as an orange thin bar. The following structural regions are shown: the 5' and 3' long terminal repeats (LTRs) with the typical U3-R-U5 structure (in light blue), the CAAT box and TATA box (in red), the polyadenylation (polyA) signal (in violet), the primer binding site (PBS) (in green), the Cys-His box (in grey) and the polypurine tract (PPT) (in pink). Protein domain abbreviations: MA, matrix; CA, capsid; NC, nucleocapsid; Pro, protease; RT, reverse transcriptase; IN, integrase; SU, surface unit; TM, transmembrane subunit.

Figure 2. GALVs maximum likelihood phylogenetic tree inferred using concatenated partitioned full genome nucleotide sequences. Coding sequences, non-coding LTRs and inter-gene spacers were included in the analysis. The sequences obtained from GenBank with corresponding accession codes are: GALV-SEATO (KT724048); GALV-SF (KT724047); GALV-Brain (KT724049); GALV-Hall's Island (KT724050); woolly monkey virus (WMV; KT724051) and *Melomys burtoni* retrovirus (MbRV; KF572483-KF572486). KoRV (AF151794) was used as the outgroup. Node support was assessed with 500 rapid bootstrap pseudoreplicates and is indicated at each node. The scale bar indicates 0.05 nucleotide substitutions per site. The tree is midpoint-rooted for purposes of clarity.

603 **TABLES**

604 **Table 1. Rodent species screened using hybridization capture for the presence of KoRV-like and**
605 **GALV-like sequences.**

Species n°	Species	Country	Code
1	*Bandicota bengalensis*	Bangladesh	2
2	*Bandicota indica*	Cambodia	10
3	*Bandicota savilei*	Myanmar	13
	Bandicota savilei	Myanmar	14
4	*Berylmys berdmorei*	Laos	19
	Berylmys berdmorei	Laos	20
	Berylmys berdmorei	Laos	22
5	*Berylmys bowersi*	Laos	27
	Berylmys bowersi	Laos	28
6	*Berylmys mackenzii*	India	31
7	*Chiromyscus chiropus*		32
	Chiromyscus chiropus	Laos	35
8	*Laonastes aenigmus*	Laos	37
	Laonastes aenigmus	Laos	41
9	*Leopoldamys edwardsae*	Laos	42
10	*Maxomys moi*	Laos	54
11	*Maxomys surifer*	Laos	55
12	*Mus booduga*	Bangladesh	60
	Mus booduga		61
13	*Mus caroli*	Laos	96
	Mus caroli	Cambodia	99
14	*Mus cervicolor*	Laos	103
	Mus cervicolor	Laos	104
	Mus cervicolor	Laos	106
	Mus cervicolor	Laos	108
15	*Mus cookii*		115
	Mus cookii	Laos	116
16	*Mus fragilicauda*	Cambodia	118
17	*Mus lepidoides*	Myanmar	121
	Mus lepidoides	Myanmar	123
18	*Mus musculus*	Bangladesh	124
	Mus musculus	Bangladesh	126
	Mus musculus	Bangladesh	128
	Mus musculus	Bangladesh	129
19	*Mus nitidulus*	Myanmar	133
	Mus nitidulus	Myanmar	134
20	*Mus terricolor*	Bangladesh	135
21	*Niviventer confucianus*	Laos	140
	Niviventer confucianus	Laos	141
22	*Niviventer fulvescens*	Laos	143
23	*Niviventer langbianis*	Laos	150
24	*Vandeleuria oleracea*	Myanmar	196
25	*Melomys* sp.	Indonesia	WD309
	Melomys sp.	Indonesia	WD282
	Melomys sp.	Indonesia	WD283
	Melomys sp.	Indonesia	WD310
	Melomys sp.	Indonesia	WD144
	Melomys sp.	Indonesia	WD279
26	*Melomys paveli*	Indonesia	YS284

606

607

Table 2. Residues of PiT1 and PiT2 region A of species permissive and resistant to GALV infection.

Receptor	Positions of region A residues									GALV
PiT1	550	551	552	553	554	555	556	557	558	recognition
Homo sapiens	*D*$^-$	T	G	*D*$^-$	V	S	S	K	V	+
Hylobates lar	*D*$^-$	T	G	*D*$^-$	V	S	S	K	V	+
Nomascus leucogenys	*D*$^-$	T	G	*D*$^-$	V	S	S	K	V	+
Rattus norvegicus	*E*$^-$	T	R	*D*$^-$	V	T	T	K	E	+
Mus musculus molossinus	I	T	G	*D*$^-$	V	S	S	K	M	+
Melomys sp.	*E*$^-$	T	G	*D*$^-$	V	S	T	K	A	+
Melomys paveli	*E*$^-$	T	-	*D*$^-$	V	S	T	K	A	?
Mus musculus musculus	**K**	Q	-	*E*$^-$	A	S	T	K	A	-
Mus dunni	**K**	Q	-	*D*$^-$	A	S	T	K	A	-
Mus caroli	**K**	Q	-	*D*$^-$	A	S	T	K	A	-
PiT2	522	523	524	525	526	527	528	529	530	
Cricetulus griseus	*E*$^-$	Q	G	G	V	M	Q	*E*$^-$	A	+
Melomys sp.	*E*$^-$	Q	G	G	V	T	Q	*E*$^-$	A	+
Melomys paveli	*E*$^-$	Q	G	G	V	T	Q	*E*$^-$	A	?
Mus musculus molossinus	Q	Q	G	G	V	T	Q	*E*$^-$	A	+
Mus caroli	Q	Q	G	G	V	T	Q	*E*$^-$	A	?
Homo sapiens	**K**	Q	G	G	V	T	Q	*E*$^-$	A	-
Rattus norvegicus	**K**	Q	G	G	V	T	Q	*E*$^-$	A	-
Hylobates lar	**K**	Q	G	G	V	M	Q	*E*$^-$	A	?

NOTE: Lys (K) is bold when present at the first position of PiT1 or PiT2 region A, which prevent GALV infection. Asp (D) and Glu (E), which are acidic and negatively charged residues, are italicized with a minus sign (-). A question mark (?) is used for those species which were never found infected with GALV or never experimentally tested for susceptibility to GALV infection.

REFERENCES

1. **Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE.** 2010. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science (New York, N.Y.) **329:**676-679.
2. **Alfano N, Kolokotronis SO, Tsangaras K, Roca AL, Xu W, Eiden MV, Greenwood AD.** 2015. Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses. Journal of virology **90:**1757-1772.
3. **Kawakami TG, Huff SD, Buckley PM, Dungworth DL, Synder SP, Gilden RV.** 1972. C-type virus associated with gibbon lymphosarcoma. Nature: New biology **235:**170-171.
4. **DePaoli A, Johnsen DO, Noll MD.** 1973. Granulocytic leukemia in white handed gibbons. J Am Vet Med Assoc **163:**624-628.
5. **Reitz MS, Jr., wong-Staal F, Haseltine WA, Kleid DG, Trainor CD, Gallagher RE, Gallo RC.** 1979. Gibbon ape leukemia virus-Hall's Island: new strain of gibbon ape leukemia virus. Journal of virology **29:**395-400.
6. **Todaro GJ, Lieber MM, Benveniste RE, Sherr CJ.** 1975. Infectious primate type C viruses: Three isolates belonging to a new subgroup from the brains of normal gibbons. Virology **67:**335-343.
7. **Theilen GH, Gould D, Fowler M, Dungworth DL.** 1971. C-type virus in tumor tissue of a woolly monkey (Lagothrix spp.) with fibrosarcoma. Journal of the National Cancer Institute **47:**881-889.
8. **Wolfe LG, Smith RK, Deinhardt F.** 1972. Simian sarcoma virus, type 1 (Lagothrix): focus assay and demonstration of nontransforming associated virus. Journal of the National Cancer Institute **48:**1905-1908.
9. **Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF.** 2000. The nucleotide sequence of koala (Phascolarctos cinereus) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus. Journal of virology **74:**4264-4272.
10. **Shojima T, Yoshikawa R, Hoshino S, Shimode S, Nakagawa S, Ohata T, Nakaoka R, Miyazawa T.** 2013. Identification of a novel subgroup of Koala retrovirus from Koalas in Japanese zoos. Journal of virology **87:**9943-9948.
11. **Xu W, Stadler CK, Gorman K, Jensen N, Kim D, Zheng H, Tang S, Switzer WM, Pye GW, Eiden MV.** 2013. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proceedings of the National Academy of Sciences of the United States of America **110:**11547-11552.
12. **Kawakami TG, Kollias GV, Jr., Holmberg C.** 1980. Oncogenicity of gibbon type-C myelogenous leukemia virus. International journal of cancer. Journal international du cancer **25:**641-646.
13. **Tarlinton R, Meers J, Hanger J, Young P.** 2005. Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. The Journal of general virology **86:**783-787.
14. **Tarlinton R, Meers J, Young P.** 2008. Biology and evolution of the endogenous koala retrovirus. Cellular and molecular life sciences : CMLS **65:**3413-3421.
15. **Fiebig U, Hartmann MG, Bannert N, Kurth R, Denner J.** 2006. Transspecies transmission of the endogenous koala retrovirus. Journal of virology **80:**5651-5654.
16. **Simmons G, Clarke D, McKee J, Young P, Meers J.** 2014. Discovery of a novel retrovirus sequence in an Australian native rodent (Melomys burtoni): a putative link between gibbon ape leukemia virus and koala retrovirus. PloS one **9:**e106954.
17. **Wong S, Lau S, Woo P, Yuen KY.** 2007. Bats as a continuing source of emerging infections in humans. Reviews in medical virology **17:**67-91.
18. **Cui J, Tachedjian G, Tachedjian M, Holmes EC, Zhang S, Wang LF.** 2012. Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats. The Journal of general virology **93:**2037-2045.

19. **Martin J, Herniou E, Cook J, O'Neill RW, Tristem M.** 1999. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. Journal of virology **73:**2442-2449.
20. **Lieber MM, Sherr CJ, Todaro GJ, Benveniste RE, Callahan R, Coon HG.** 1975. Isolation from the asian mouse Mus caroli of an endogenous type C virus related to infectious primate type C viruses. Proceedings of the National Academy of Sciences of the United States of America **72:**2315-2319.
21. **Callahan R, Meade C, Todaro GJ.** 1979. Isolation of an endogenous type C virus related to the infectious primate type C viruses from the Asian rodent Vandeleuria oleracea. Journal of virology **30:**124-131.
22. **Benveniste RE, Callahan R, Sherr CJ, Chapman V, Todaro GJ.** 1977. Two distinct endogenous type C viruses isolated from the asian rodent Mus cervicolor: conservation of virogene sequences in related rodent species. Journal of virology **21:**849-862.
23. **Miller AD, Bergholz U, Ziegler M, Stocking C.** 2008. Identification of the myelin protein plasmolipin as the cell entry receptor for Mus caroli endogenous retrovirus. Journal of virology **82:**6862-6868.
24. **Wolgamot G, Bonham L, Miller AD.** 1998. Sequence analysis of Mus dunni endogenous virus reveals a hybrid VL30/gibbon ape leukemia virus-like structure and a distinct envelope. Journal of virology **72:**7459-7466.
25. **Bromham L, Clark F, McKee JJ.** 2001. Discovery of a novel murine type C retrovirus by data mining. Journal of virology **75:**3053-3057.
26. **Maricic T, Whitten M, Paabo S.** 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PloS one **5:**e14004.
27. **Tsangaras K, Siracusa MC, Nikolaidis N, Ishida Y, Cui P, Vielgrader H, Helgen KM, Roca AL, Greenwood AD.** 2014. Hybridization capture reveals evolution and conservation across the entire Koala retrovirus genome. PloS one **9:**e95633.
28. **Sambrook J, Russell DW.** 2006. Purification of nucleic acids by extraction with phenol:chloroform. CSH protocols **2006**.
29. **Meyer M, Kircher M.** 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor protocols **2010:**pdb.prot5448.
30. **Alfano N, Courtiol A, Vielgrader H, Timms P, Roca AL, Greenwood AD.** 2015. Variation in koala microbiomes within and between individuals: effect of body region and captivity status. Scientific reports **5:**10189.
31. **Kircher M, Sawyer S, Meyer M.** 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic acids research **40:**e3.
32. **Martin M.** 2012. Cutadapt removes adapter sequences from high-throughput sequencing reads. . Bioinformatics in Action **17:**10-12.
33. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) **30:**2114-2120.
34. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research **25:**3389-3402.
35. **Ondov BD, Bergman NH, Phillippy AM.** 2011. Interactive metagenomic visualization in a Web browser. BMC bioinformatics **12:**385.
36. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv **1303**.
37. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) **25:**2078-2079.
38. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution **30:**772-780.
39. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) **30:**1312-1313.

40. **Lanave C, Preparata G, Saccone C, Serio G.** 1984. A new method for calculating evolutionary substitution rates. Journal of molecular evolution **20:**86-93.
41. **Yang Z.** 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of molecular evolution **39:**306-314.
42. **Oliveira NM, Satija H, Kouwenhoven IA, Eiden MV.** 2007. Changes in viral protein function that accompany retroviral endogenization. Proceedings of the National Academy of Sciences of the United States of America **104:**17506-17511.
43. **Shojima T, Hoshino S, Abe M, Yasuda J, Shogen H, Kobayashi T, Miyazawa T.** 2013. Construction and characterization of an infectious molecular clone of Koala retrovirus. Journal of virology **87:**5081-5088.
44. **Ting YT, Wilson CA, Farrell KB, Chaudry GJ, Eiden MV.** 1998. Simian sarcoma-associated virus fails to infect Chinese hamster cells despite the presence of functional gibbon ape leukemia virus receptors. Journal of virology **72:**9453-9458.
45. **O'Hara B, Johann SV, Klinger HP, Blair DG, Rubinson H, Dunn KJ, Sass P, Vitek SM, Robins T.** 1990. Characterization of a human gene conferring sensitivity to infection by gibbon ape leukemia virus. Cell growth & differentiation : the molecular biology journal of the American Association for Cancer Research **1:**119-127.
46. **Johann SV, van Zeijl M, Cekleniak J, O'Hara B.** 1993. Definition of a domain of GLVR1 which is necessary for infection by gibbon ape leukemia virus and which is highly polymorphic between species. Journal of virology **67:**6733-6736.
47. **Schneiderman RD, Farrell KB, Wilson CA, Eiden MV.** 1996. The Japanese feral mouse Pit1 and Pit2 homologs lack an acidic residue at position 550 but still function as gibbon ape leukemia virus receptors: implications for virus binding motif. Journal of virology **70:**6982-6986.
48. **Pedersen L, Johann SV, van Zeijl M, Pedersen FS, O'Hara B.** 1995. Chimeras of receptors for gibbon ape leukemia virus/feline leukemia virus B and amphotropic murine leukemia virus reveal different modes of receptor recognition by retrovirus. Journal of virology **69:**2401-2405.
49. **Pedersen L, van Zeijl M, Johann SV, O'Hara B.** 1997. Fungal phosphate transporter serves as a receptor backbone for gibbon ape leukemia virus. Journal of virology **71:**7619-7622.
50. **Wilson CA, Farrell KB, Eiden MV.** 1994. Properties of a unique form of the murine amphotropic leukemia virus receptor expressed on hamster cells. Journal of virology **68:**7697-7703.
51. **van Zeijl M, Johann SV, Closs E, Cunningham J, Eddy R, Shows TB, O'Hara B.** 1994. A human amphotropic retrovirus receptor is a second member of the gibbon ape leukemia virus receptor family. Proceedings of the National Academy of Sciences of the United States of America **91:**1168-1172.
52. **Wilson CA, Farrell KB, Eiden MV.** 1994. Comparison of cDNAs encoding the gibbon ape leukaemia virus receptor from susceptible and non-susceptible murine cells. The Journal of general virology **75 (Pt 8):**1901-1908.
53. **Chaudry GJ, Eiden MV.** 1997. Mutational analysis of the proposed gibbon ape leukemia virus binding site in Pit1 suggests that other regions are important for infection. Journal of virology **71:**8078-8081.
54. **Eiden MV, Farrell KB, Wilson CA.** 1996. Substitution of a single amino acid residue is sufficient to allow the human amphotropic murine leukemia virus receptor to also function as a gibbon ape leukemia virus receptor. Journal of virology **70:**1080-1085.
55. **Callahan R, Benveniste RE, Sherr CJ, Schidlovsky G, Todaro GJ.** 1976. A new class of genetically transmitted retravirus isolated from Mus cervicolor. Proceedings of the National Academy of Sciences of the United States of America **73:**3579-3583.
56. **Ishida Y, Zhao K, Greenwood AD, Roca AL.** 2015. Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. Molecular biology and evolution **32:**109-120.