Germline Colonization by Retroviruses:

A New Rodent Model to Understand Host-Virus Interactions at the Early Stages of Retroviral Endogenization

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by

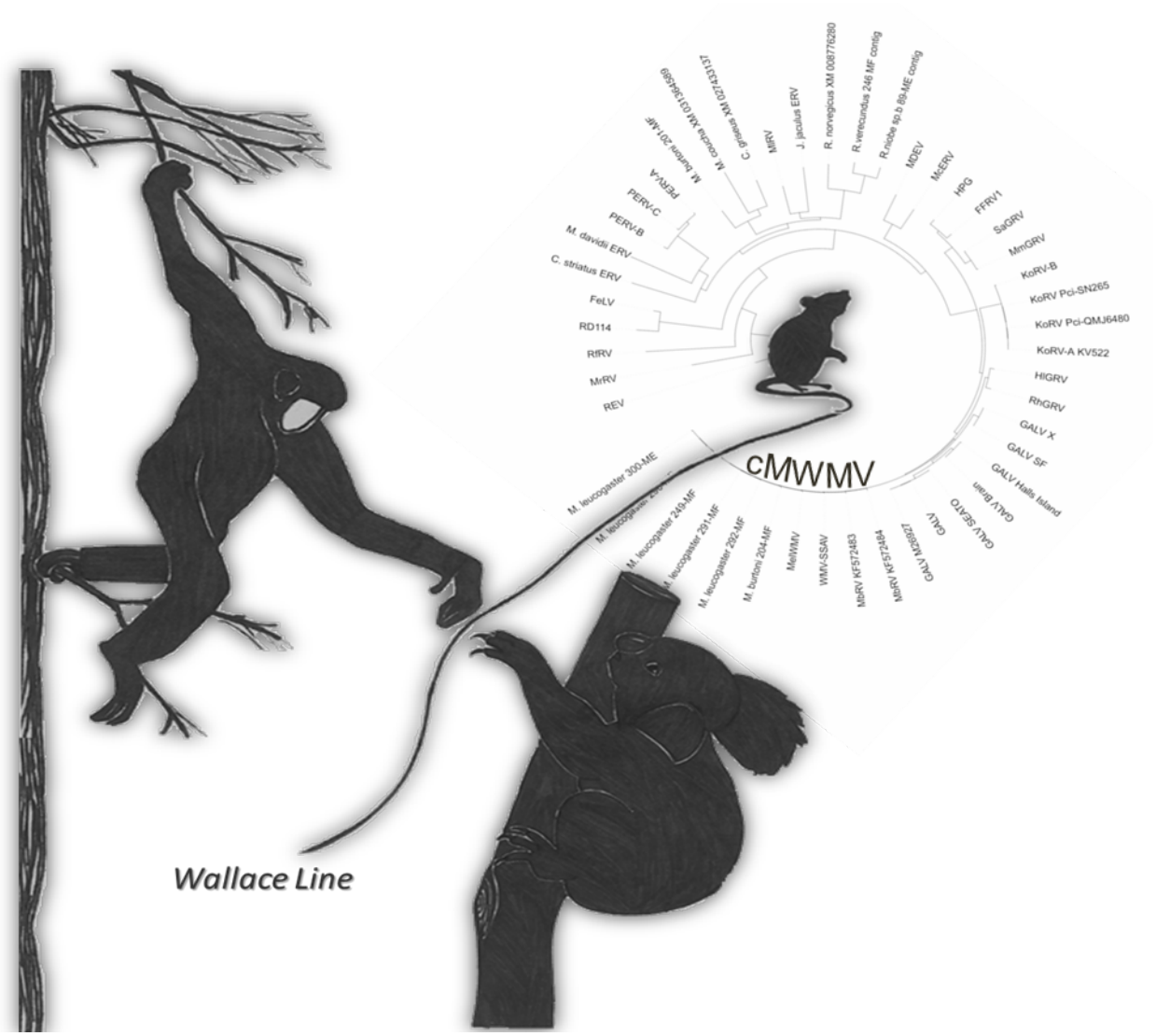
Saba Mottaghinia

from Tehran-Iran

Berlin, 2022

With the tools and the knowledge, I could turn a developing snail's egg into an elephant. It is not so much a matter of chemicals because snails and elephants do not differ that much; it is a matter of timing the action of genes.

— Barbara McClintock



Research presented in this dissertation was carried out under the supervision of Prof. Alex D. Greenwood at the Leibniz Institute for Zoo and Wildlife Research from 07.05.2018 until 06.02.2022 and it is submitted to the Department of Biology, Chemistry and Pharmacy of Freie Universität Berlin.

1st Reviewer: Prof. Alex D. Greenwood, PhD.

Department of Veterinary Medicine, Freie Universität Berlin Berlin, Germany

and Leibniz Institute for Zoo and Wildlife Research, Department of Wildlife Diseases

2nd Reviewer: Prof. Dr. Dino McMahon

Institute of Biology - Zoology, Freie Universität Berlin Berlin, Germany

Date of defense: 25.08.2022

Contents

Acknowledgements

The reverse transition from an ordinary life, quitting a well-paying job in industry to an unprecedented student life living off savings was not an easy path. I realize now how fortunate I was to stumble upon the growing field of evolutionary biology. But I could not have realized my passion if it wasn't for my master thesis supervisor, Prof. Bernhard Misof, who opened my eyes to the world of viral evolution and bioinformatics.

I could not have known that a significant amount of my genome is derived from viruses if my supervisor and mentor, Prof. Alex D. Greenwood, would not have given me this chance to work on this incredible project. I probably would not have had a chance to hug a koala. Thank you for your encouragement and sparking my interest into the extraordinary world of transposable elements, a field I'm determined to build my scientific career on.

My gratitude to the late Prof. Kenneth P. Aplin, whom I never met but his lifelong fieldwork efforts, especially in Australia, New Guinea and across Asia, made this project possible. I'm thankful to the most generous human I know, lady Rosie Nekzad, who kept my Persian side alive in Berlin. Cheers to Dr. Claudia Szentiks (unser pathologe) who taught me to spread flower seeds rather than seeds of disappointment and not to be ashamed of the amount of coffee one consumes! To Karin Hönig, my lab mom, and to Susanne Auls, a friend and the protector of students against German bureaucracy. To Dorina Meneghini, I give her the most patient teacher award. To Layla Mpinou (yes, don't go away) for her random act of kindness. To the coffee breakers and future doctors, Morgane Gicquel, Miguel Veiga, Alex Badry, Juan Li, John Galindo, Rohit Chakravarty and Seth Wong, thank you for your friendship and the extra 10 kg. To the cool PhD coordinators, Dr. Sarah Benhaiem and Dr. Gábor Czirják, and to the past PhD students who paved the way, Drs. Sofia Paraskevopoulou, Michal Hryciuk, Daniela Numberger, Sanatana Soilemetzidou and Peter Seeber; to old friends and the new; to all the members of department 3, especially Katja Pohle and Carin Hoffmann; to IZW and the good they do, especially Prof. Heribert Hofer to keep us in line; to my collaborators from all over the globe, I say thank you, Danke and Merci.

Lastly, to my family, whom I owe everything to. Thank you maman Mehri for passing on your sense of humor; thank you baba Jamshid for your endless support, no matter how old I get. Thank you Sina for bringing music and adventure to our lives. You guys sacrificed so much of your own comfort and happiness and always filled my life with your unconditional love. I hope I made you proud.

Declaration of Independence

Herewith, I certify that I have prepared and written my thesis independently and that I have not used any sources and aids other than those indicated by me.

Statement of Contributions

This study was supported by the Deutsche Forschungsgemeinschaft (DFG) grant GR3924-12-1.

The South Australian Museum provided access to the tissue samples that were used for this study.

The Australian Genome Research Facility (AGRF) performed the illumina high-throughput sequencing run for the second chapter.

Prof. Nikolas Nikolaidis and Dr. Kyriakos Tsangaras performed the protein modeling and constructed Figure 2.9 that was used in the second chapter.

Dr. Karin Müller at the Leibniz Institute for Zoo and Wildlife Research in Berlin, helped this work with acquisition of immunofluorescence microscopy.

GenScript synthesized our viral sequences and cloned viruses (cMWMV and KoRV-A) that were used in the second chapter.

Claudia Quedenau at the Max-Delbrück-Centrum für Molekulare Medizin (MDC) in Berlin, constructed the PacBio library and performed the sequencing run for the third chapter.

Saskia Stenzel at the Institute of Virology, Charité Universitätsmedizin Berlin, performed virus propagation, infection experiments and fixed slides for the electron microscopy (EM) that were used for the second chapter.

Dr. Michael Laue at the Laboratory for Diagnostic Electron Microscopy of Infectious Pathogens, Robert Koch-Institut, Berlin produced the virus electron microscopy images in the second chapter.

Dr. Gayle McEwen at the Leibniz Institute for Zoo and Wildlife Research in Berlin, helped this work by constructing the coding sequences of antiretroviral genes in chapter III, section 3.3.2.

Carin Hoffmann at the Leibniz Institute for Zoo and Wildlife Research in Berlin, translated the summary section from English to German.

List of Figures

List of Tables

List of Text Files

Zusammenfassung

Im Gegensatz zu anderen Viren replizieren sich Retroviren, indem sie ihr RNS-Genom in DNS kopieren. Daher werden sie nach der Infektion zu einem weitgehend untrennbaren Teil des Zellgenoms. Retroviren können durch Infektion horizontal und vertikal übertragen werden und haben oft einen breiten Zelltropismus. Eine exogene retrovirale Infektion (XRV) findet in somatischen Zellen statt. Erfolgt jedoch die Infektion in der Keimbahn, wird das resultierende Provirus als endogenes Retrovirus (ERV) bezeichnet. Die Anhäufung dieser retroviralen Sequenzen im Laufe der Evolution hat dazu geführt, dass sie ca. 8 % des menschlichen Genoms einnehmen und zusammen mit anderen transponierbaren Elementen (TEs) eine wichtige Determinante der DNS-Sequenzvielfalt sowie eine treibende Kraft für die Evolution der Arten darstellen. Jahrmillionen der Evolution haben den Verlauf von Mutation, Indel, Umlagerung und Verbreitung, die ERVs seit ihrer Integration erfahren haben, verschleiert. Die Art und Weise, wie sich ERVs in einem Wirtsgenom etablieren, ist entscheidend, um Rückschlüsse auf die adaptive Immunität von Wirbeltieren und das erzeugte Gedächtnis dieser Genom-Invasoren zu ziehen. Das Koala-Retrovirus (KoRV), das einzige bekannte Säugetier-Retrovirus, das derzeit eine Genomkolonisierung durchläuft, wird im Allgemeinen als Modellsystem für den Mechanismus der Endogenisierung verwendet. Die Vorläufer-Vektorspezies, die KoRV und das eng verwandte pathogene Gibbon-Affen-Leukämievirus (GALV) hervorgebracht hat, ist jedoch nach wie vor unbekannt. In einem Versuch, das Reservoir von GALV-KoRV zu identifizieren, haben wir ein neuartiges infektiöses GALV-Virus in einer bestimmten Population eines in Papua-Neuguinea heimischen Nagetiers, *Melomys leucogaster*, nachgewiesen. Das Virus wurde *complete Melomys Woolly Monkey Virus* (cMWMV) genannt. Mit Hilfe von Zellkulturmethoden, Fluoreszenz- und Elektronenmikroskopie haben wir dieses *Gammaretrovirus* charakterisiert. Die Besonderheit von cMWMV besteht darin, dass es, wie KoRV, derzeit in das Genom einer neuen Wirtsart eindringt. Da KoRV nur bei Koalas vorkommt, könnte cMWMV ein zusätzliches Nagetiermodell sein, um die evolutionären Prozesse zu untersuchen, die zur Keimbahninvasion und Anpassung an einen neuen Wirt beitragen.

Diese jüngste retrovirale Invasion kann uns helfen, die allgemeinen Prinzipien der antiretroviralen Genevolution innerhalb von *Melomys* und zwischen Nagetierarten zu verdeutlichen, die bekanntermaßen der diversifizierenden Selektion der Primatenorthologen unterliegen. Mittels PacBio-Sequenzierung wurde das gesamte Genom von *Melomys* sequenziert. Ein geführter Sequenzabgleich wurde vorgenommen und die den relevanten Genen entsprechenden Exone extrahiert. Die kodierenden Sequenzen (CDS) wurden *de novo* assembliert und manuell kuratiert. Anschließend haben wir verschiedene Substitutionsmodelle angewandt, um den Selektionsdruck in diesen Immungenen zu quantifizieren. Unsere Daten deuten darauf hin, dass diese Gene, ähnlich wie bei den

Primaten, in den Abstammungslinien von *Mus musculus* und *Rattus norvegicus* an einigen Stellen (Codons) einer positiven Selektion unterlegen haben könnten. Der Überschuss an synonymen Stellen deutet jedoch auf einen langfristigen Trend der reinigenden Selektion hin. Ein schwaches, verstärkt diversifizierendes Selektionsmuster in der *Melomys*-Abstammungslinie des ZAP-Gens (Zink-Finger-CCCH-Typ antivirales Protein 1) könnte auf einen Versuch hindeuten, die virale mRNA-Translation des endogenisierenden cMWMV zu inhibieren.

Summary

Unlike other viruses, retroviruses replicate by copying their RNA genomes to DNA. They therefore become a largely inseparable part of the cell's genome upon infection. Retroviruses can be transmitted horizontally and vertically by infection and often have wide cellular tropism. Exogenous retroviral infection (XRV) occurs in somatic cells, but when infection is in the germline, the resulting provirus is known as an endogenous retrovirus (ERVs). Accumulation of these retroviral sequences over evolutionary time has granted them ~ 8% occupancy of the human genome and, along with other transposable elements (TEs), makes them a major determinant of DNA sequence diversity and driver of species evolution. Millions of years of evolution have obscured the history of mutation, indel, rearrangement and distribution events that ERVs have experienced since they integrated. Understanding how ERVs establish themselves in a host genome is crucial to infer vertebrate adaptive immunity and the generated memory of these genome invaders. Koala retrovirus (KoRV), as the only known mammalian retrovirus currently undergoing genome colonization, is generally used as a model system for mechanism of endogenization. However the precursor vector species that gave rise to KoRV and the closely related pathogenic Gibbon Ape Leukemia Virus (GALV) remains obscure. In an attempt to identify the reservoir of GALV-KoRV, we have identified a novel infectious GALV virus in a specific population of a native rodent of Papua New Guinea, *Melomys leucogaster*. We named this virus, complete melomys woolly monkey virus (cMWMV). Using cell culture methods, fluorescence, and electron microscopy, we have characterized this *gammaretrovirus*. The significance of cMWMV is that like KoRV, it is currently invading the genome of a new host species. As KoRV is restricted to koalas, cMWMV could provide an additional rodent model to further study the evolutionary processes that contribute to the germline invasion and adaptation to a new host.

This recent retroviral invasion can help us elucidate the general principles of antiretroviral gene evolution within *Melomys* and between rodent species that are known to be under diversifying selection in the primate orthologs. PacBio sequencing was used to sequence the whole genome of *Melomys*. Guided sequence alignment was performed and exons corresponding to genes of interest were extracted. Coding sequences (CDS) were *de novo* assembled and manually curated. We then used various substitution models to quantify the selection pressure in these immune genes. Our data suggest that, similar to primates, these genes may have experienced positive selection at some sites (codons) in *Mus musculus* and *Rattus norvegicus* lineages. However the excess of synonymous sites asserts a long-term trend of purifying selection. A weak intensified diversifying selection pattern in *Melomys* lineage of ZAP (zinc-finger CCCH-type antiviral protein 1) gene could indicate an effort to inhibit viral mRNA translation of the endogenizing cMWMV.

Chapter I - An Introduction to Retrovirology

1.1. The early evolution of life and the world of RNA

The RNA-first hypothesis, proposing that the first replicating molecules or precursors to life, was suggested in the 1960s when RNA molecules were found to act as messengers for genes (mRNA), adapters for proteins (tRNA), and primary components of ribosomes (rRNAs), which is essential for a more complex protein-based life [1]. Although debates continue over the RNA- or protein-world hypothesis, the fact that DNA is derived from ancestral RNA molecules is generally accepted [2,3]. The homologies between conserved domains and features of single-stranded RNA viruses with linear or circular DNA viruses strongly suggest that the latter evolved from viral RNAs [3–5]. A good example of this is the shared replication strategy, genome organization, and functionality between DNA hepadnaviruses and RNA retroviruses [6] or the remarkably conserved structural motifs found in viral capsids of both RNA and DNA viruses [7].

The principle for this transition to a more stable DNA life is the enzyme reverse transcriptase [8], a molecular machine for disrupting DNA fidelity, generating genetic variation, and the nature of retroviruses that have been around since the dawn of the DNA world, disproving the central dogma and promoting these DNAs with another ancient enzyme, the retroviral integrase [9].

If we assume that these elements are remnants of this prebiotic era, exploring the origin of RNA viruses, especially their host reservoirs, can offer a unique insight into the origin of life and their ongoing imprint on evolution.

1.2. The RNA tumor viruses

Although retroviruses have an immense impact on the evolution of their hosts, retrovirology is a relatively new field of research. Only in 1908, the first filterable agent that caused leukemia in chickens (later assigned as retrovirus avian leukemia virus-ALV) was discovered by Ellermann and Bang. In 1911, Rous presented the rous sarcoma virus (RSV) as the first tumor-inducing infectious agents [10]. However, it took another 60 years for Baltimore, Dulbecco, and Temin to discover reverse transcriptase in RNA tumor virus particles and demonstrate the carcinogenic nature of these viruses. Today, these RNA tumor viruses are called retroviruses and it is general knowledge that their integration into or near host genes can cause alteration in gene expression and permanent malignant transformation of their vertebrate

host cells [11]. Because of these unique properties that allow retroviruses to overcome natural cellular barriers, they are of considerable medical and veterinary importance, modified and routinely used as vectors for gene therapy and cancer research or as the basis for laboratory methods such as reverse transcription (RT)-qPCR [12].

1.3. Retrovirus taxonomy

Based on genome types and replication strategy, retroviruses belong to the VI group of the Baltimore scheme: the enveloped single-stranded positive sense (polarity) RNA (+ssRNA) viruses that have a double-stranded DNA (dsDNA) intermediate. The current taxonomic classification of the *Retroviridae* family is based on genetic and functional features which divides them into *orthoretrovirinae* and *spumaretrovirinae* (foamy viruses) subfamilies. The former consists of six genera; alpha-retroviruses (e.g. rous sarcoma virus), beta-retroviruses (e.g. mouse mammary tumor virus), delta-retroviruses (e.g. bovine leukemia virus), epsilon-retroviruses (e.g. walleye dermal sarcoma virus), gamma-retroviruses (e.g. gibbon ape leukemia virus) and lentiviruses (e.g. human immunodeficiency virus) (Figure 1.1).

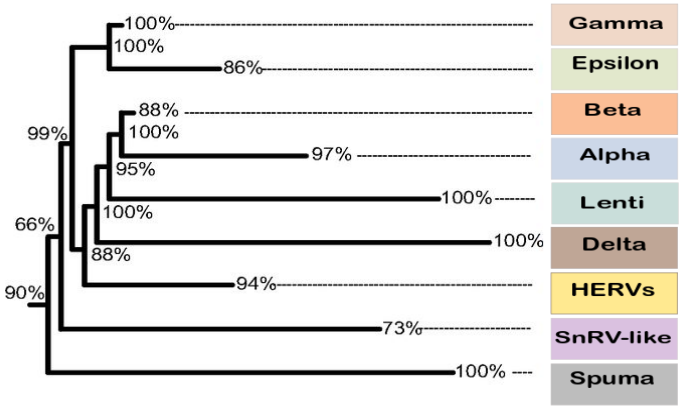


Figure 1.1. Schematic tree of the current retroviral genera is based on 65 vertebrate host genomes. Adopted from Hayward, et al. 2015 [13].

1.4. Retrovirus virion morphology and genome structure

Based on observed electron microscope morphologies, retroviral virions are grouped as types A to D [14]. Most oncoviruses such as gammaretroviruses and endogenous retroviruses have C-type viral particles that can be distinguished by a central electron-dense core enclosed in a spherical shaped envelope of 80-100 nm diameter [15].

The +ssRNA genome is diploid (two generally identical RNA molecules) and coated with nucleocapsid (NC), and, along with protease (PR), reverse transcriptase (RT) and integrase (IN) enzymes is enclosed by a protective shell known as capsid (CA). Alignment of matrix proteins (MA) outside CA and beneath the envelope membrane contributes to a virion's structure. Viral envelope is a lipid bi-layer derived from the host cell during the budding, where the inner transmembrane (TM) subunit acts as a bridge between MA and the outer surface unit (SU). Due to constant antagonistic host-virus evolutionary cycles, SU is subject to high substitution rates [16]. Receptor binding domain (RBD) in SU consists of variable regions A and B (VRA; VRB) which exhibit pathologically important motifs that determine the receptor specificity [17] (Figure 1.2).

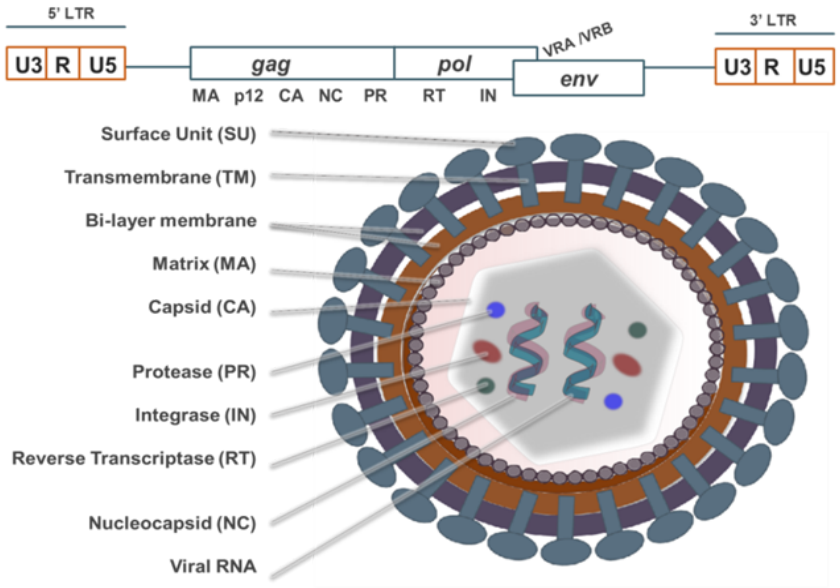


Figure 1.2. The schematic representation of a simple proviral genome structure; the coding region is flanked by identical 5' and 3' Long terminal repeats (LTR), comprising U3, repeats (R) and U5. Transcription is initiated at primer binding site (PBS) where *gag* gene encodes for matrix (MA), cleaved protein p12, capsid (CA) and nucleo-capsid (NC). The *gag-pol* reading frame, known as *pro-pol* codes for protease (PR). The remaining enzymes reverse transcriptase (RT) and integrase (IN) are encoded by *pol* gene. The *env* encodes structural polyprotein ENV that is a precursor of surface unit (SU) and Transmembrane (TM) proteins. The receptor binding domain (RBD) in SU is composed of variable regions A and B (VRA; B) that determine cell tropism.

Retroviruses are non-segmented with a relatively small genome of 7-10 Kb. These viruses have four open reading frames (ORFs) with a conserved 5'-LTR _ *gag* _ *pro*_ *pol* _ *env* _3'-LTR principle that codes for structural group-specific-antigen polyprotein GAG (MA, p12 protein, CA and NC subunits), envelope

polyprotein ENV (SU and TM subunits), and the functional polymerase POL (RT and IN enzymes). GAG-PRO-POL precursor polyprotein codes for PR that is responsible for gene product maturation as it cleaves the viral polyproteins into their prospective subunits. This coding sequence (CDS) is flanked at either side by identical non-coding regulatory long terminal repeat sequences (the 5′ and 3′ LTRs); each consists of untranslated-3′ (U3), repeat elements (R) and untranslated-5′ (U5) sequences, and contains transcriptional promoter and regulatory sequences that are applied in viral replication and gene expressions (Figure 1.2).

1.5. Life cycle of simple retroviral genomes

Based on genome complexity and mechanism of infection, retroviruses are categorized as simple (alpha and gammaretroviruses) or complex (beta-, delta-, epsilon-, spuma-retroviruses and lentiviruses). Retroviruses with simple arrangements contain three conserved retroviral genes and make a single spliced mRNA whereas complex retroviral genomes encode auxiliary proteins. Complex retrovirus genomes have multiple-spliced mRNAs allowing them to enter the nucleus of non-dividing cells (such as memory T-cells) via nuclear pores and are less dependent on the host cell function, which is an advantage for spreading into new cell types [18].

The life cycle of retroviruses with simple arrangement is divided into an early (entry, provirus production) and a late (budding) phases. The first phase defines the target cell tropism and host range of the virus as SU binds on a specific plasma membrane receptor of the host cell [19]. As a result of this interaction, TM will undergo conformational changes to mediate fusion with the cell membrane before the virus core is released into the cytoplasm [20]. The viral +ssRNA functions as mRNA. Reverse transcription is initiated at a primer binding site (PBS) from 5′-LTR where tRNA resides and synthesizes a -ssDNA and uses it as a template for a complementary positive strand. Transcription is terminated at 3′-LTR with a linear dsDNA copy flanked by identical viral RNA derived LTRs [21]. Transcription end product is a dsDNA intermediate still contained within the capsid, known as pre-integration complex (PIC). In simple retroviruses, PIC translocation into the nucleus depends on mitosis and breaking down of the nuclear membrane. Once inside the host cell chromosome, integrase will insert the viral DNA into the host chromosome, known as a provirus.

The provirus functions as a single expression unit that can be further transcribed into viral RNA from the promoter region located at the 5′-LTR. RNA genome and the transcribed viral RNA are transported to cytoplasm. Viral RNA acts as mRNA and uses cellular ribosomes for translation and production of viral proteins. The Env proteins are often separately spliced transcripts, translated from a spliced full length

mRNA, and are sent to the endoplasmic reticulum (ER) to form their glycoprotein complex. ENV is then cleaved via golgi apparatus and is transported to the plasma membrane. At the time of assembly with GAG and POL, ENV glycoproteins become incorporated into budding virions. Budding will cause virions to be enclosed within the host cell plasma membrane and are budded from the infected cell surface, completing one cycle [19].

1.6. Cross-Species Transmission (CST) and retroviral envelopes

The majority of emerging viral infections are the result of cross-species transmission (CST) [22]. The disturbance of wildlife habitats has inevitably led to more spillover events, frequently mediated by mutagenic RNA viruses. One of the most notorious examples initiated by viral CST is the human immunodeficiency viruses type-1 (HIV1) and type-2 (HIV-2) transmission to humans from non-human primates [23]. Most CSTs were presumed to involve host species from the same order because of the evolutionary constraints on divergent taxa [24]. However, retroviral CST events also frequently occur between species with different immunological responses and life-history traits such as the case of the koala retroviruses (KoRVs) in a marsupial and the closely related Gibbon Ape Leukemia Viruses (GALVs) in primates, rodents, and bats [25–27].

The main determinant of retroviral wide cellular tropism is the high mutation rate that allows SU of the envelope glycoprotein many opportunities to evolve and interact with different cellular surface receptors and cross host-specific barriers. In gammaretroviruses, this receptor binding domain (RBD) consists of VRA and VRB and is highly diverse which allows for horizontal and vertical viral transmission. GALV and KoRV-A have been shown to use the highly expressed mammalian sodium-dependent phosphate transporter membrane protein (SLC20A1, also known as PiT-1) to infect human cells [28,29]. Additionally, GALV and the more pathogenic KoRV strains employ PiT-2 and thiamine transport protein 1 (THTR1) respectively to infect a wide variety of mammalian hosts [30,31]. The envelope gene diversification in KoRV is proposed to be the result of genetic innovation to overcome hosts' superinfection blockage mechanism [32]. Ultimately the *env*-less vertically transmitted retroviruses can shift to retrotransposition and greater proliferation [33]. Identifying these variations in retroviral envelope's functional motifs is essential to understand the degree of pathogenicity. Such studies identified CETTG motif in GALV envelope to be highly conserved across horizontally transmitting gammaretroviruses while the vertically transmitted KoRV-A has a CETAG motif [34].

Knowing how viruses interact with host cell receptors is critical to understand how viruses invade host cells which can specify tissue tropism and elucidate the disease outcomes. Although a successful virus

establishment is determined by the biological compatibility of both host and pathogen, geographic overlap provides the opportunity for CST and viral genome expansions. Biogeographical barriers are a potent driving force for population isolation [35]. For example, the Wallace Line is a prominent biogeographical barrier to gene flow between animal fauna of Southeast Asia from Austrlo-Papuan and pathogen transmission [36,37].

1.7. Endogenous retroviruses (ERVs)

As a consequence of their life cycle, viral elements may have access to the host genome. If viral integration occurs in the germline or in the early stages of embryogenesis by chance, this could lead to provirus being inherited vertically in every host cell and across generations, though expression is not guaranteed. In other virus taxa, this is an anomalous occurrence, but, for retroviruses, this is the basis of their replication known as endogenous retroviruses (ERVs) [38–40]. The abundance and variety of ERVs in vertebrate genomes, reflects extensive prior colonization activities that have contributed to genomic evolution of their host species. The preliminary evidence of these relics of past infection was provided in the late 1960s by hybridization of viral DNA with host DNA and subsequent southern blotting, revealing multiple ERV insertions at chromosomal loci [41]. Phylogenetically, ERVs are categorized with their exogenous counterparts (XRV) and named after the host species which they were identified in but not necessarily limited to. Such is the case of GALV which was identified initially in gibbons but is frequently found in a wide variety of rodents and bats [42]. But, based on the mechanism of transposition, ERVs are classified as transposable elements (TEs) within retrotransposons (Class I: copy and paste mechanism) and along DNA transposons (class II: cut and paste mechanism) occupy a major component of eukaryotic genomes [43]. Previously, they were considered as "junk DNA" or "jumping genes". However, with increasing sequencing resolution and curated genomes across species, their profound impact on host evolution, particularly vertebrate genomes, is becoming apparent.

ERVs are selfish elements that, once having entered the host genome, seek to proliferate. ERV copy number will be determined by the number of CSTs and the subsequent proliferation [11]. There are three known routes for this mechanism: (i) reinfection, to produce intact viral particles that can re-integrate into the germline, (ii) *cis*-activating retrotransposition and (iii) *trans*-activating complementation [33,44,45]. These mechanisms create insertionally polyphormic ERV copies that would be subject to the strongest evolutionary forces: internally from the intrinsic immune system of the host and externally from topographical barriers such as Wallace Line [46]. These factors make the rate and outcome of endogenization unpredictable and different amongst host populations and taxa [47].

Retroviral invasion triggers a continuous arms race evolutionary response from the host population to evolve pathways to stop spread of infection. Generally, the deleterious insertions would be removed from the host population by purifying selection. Those which are beneficial would be positively selected for co-option and would be fixed in that population or species [48]. Such as the mouse viral derived Friend virus susceptibility1 (*Fv1*) restriction factor that blocks retroviral infection in mice [49] or the evolution of retroviral *env* to the vital *syncytin* in mammalian embryo implantation [50]. Selectively neutral insertions would decay through point mutation or transcriptionally silenced by methylation [51]. Occasionally, recombination between ERVs and or their exogenous counterparts may interfere with new infections (superinfection interference) [52] or permit re-mobilization to enhance novel infections [53].

1.8. Mechanism of germline invasion

Retroviral endogenization is not an essential step but rather a stochastic event in every vertebrate genome that has been screened to date. ERVs insertion site preferences include AT-rich regions with abundant microsatellites, mirror repeats, and repressive histone marks [54]. These mutational loads have been linked to a wide range of diseases such as cancer [55] where their fate is governed by population genetics processes. Most ERV insertions are therefore deleterious and are removed by the host population with purifying selection [56]. Regions favoring fixation are therefore those of evolutionarily conserved with low recombination rates and depleted of genes [54]. The remaining neutral ERVs, with an antisense orientation bias, are found in non-coding regions that are mildly deleterious, neutral, or confer a selective advantage [57]. This integration versus fixation preferences, result in an uneven distribution of ERVs along the genome whereas for XRVs this includes transcriptional units, gene dense regions, and regions associated with gene activity [54,57].

Millions of years of evolution result in the accumulation of disruptive mutations, leaving these ERVs degraded and obscuring the deep evolution of ancestral lineages. Retroviral repressive epigenetic mechanisms, such as DNA methylation in the time of germline reprogramming compared to somatic cells [58], could indicate alternative pathways for controlling ERVs. Cellular apoptosis, interferons, cytokines and the germline piwi-interacting RNAs (piRNAs) that can protect the genome from transposon activation in a pervasive adaptive manner may also help control ERVs [59]. Our knowledge for retroviral origins and their coevolution with vertebrate hosts is still fragmented. With few exceptions, endogenization has already occurred in the genomes of most vertebrates and the progenitor viral lineage has often gone extinct. This makes reconstructing the preliminary responses to viral germline

colonization that were critical to endogenization from the deep evolutionary timescale a notoriously challenging task.

Additionally, factors that contribute to different rates of germline reconciliation with a *de novo* retroviral insertion are not completely understood. For example, the activity of most human ERVs have decreased significantly whereas mouse ERV loci are still highly active. This striking pattern difference between primates and rodent ERV activity has been presumably linked 37% to body size and 68% to variance in rate of ERV integration [11]. However, the same pattern is not observed in other mammals. Replication-competent young ERVs are likely to be insertionally polymorphic and are present at low allele frequencies [60]. In early stages of retroviral endogenization, these autonomous retrotransposons may still transcribe and encode a RNA or protein that interferes with transposon silencing. Therefore an endogenizing young ERV can be used to explore early evolutionary pressure leading to endogenization and the functional consequences on the regulatory network of the host population.

1.9. Gammaretroviruses in spotlight

With few exceptions in birds and reptiles, gammaretroviruses have predominantly been identified in mammals. Unlike complex lentiviruses that are known to cause degenerative diseases, alpharetroviruses and gammaretroviruses are C-type oncogenic retroviruses that have been linked to proliferative diseases such as immunosuppressive disorders and malignancies [20]. Gammaretroviruses have attracted significant research interest due to some of its prominent members and their frequent occurrence in several vertebrates [24]. These include porcine endogenous retroviruses (PERVs) and the possibility of CST to humans via xenotransplantation [61]. Gammaretroviruses are incapable of infecting post-mitotic cells, but they employ a wide variety of cellular receptors to their advantage, such as the case of murine leukemia viruses (MLVs). Since its discovery, MLVs are used as a prototype to study leukemia and are often utilized as a starting material in vectors for gene therapy [62]. Gibbon Ape Leukemia Viruses (GALVs) and the closely related koala retrovirus (KoRV) group are used as a model system to study the early stages of retroviral endogenization and adaptation to a new host. The history of these two viruses, which are the focus of this thesis, is described below.

1.9.1. Gibbon Ape Leukemia Viruses (GALVs)

GALV is an exogenous *gammaretrovirus* with oncogenic potential [14,63]. The recognized strains of GALV includes the initial isolates from cases of lymphoid neoplasia in captive white-handed gibbons (*Hylobates lar*) at research facilities in Bangkok (GALV-SEATO) and in San Francisco (GALV-SF). GALVs were

subsequently detected at other locations in the USA (GALV-Br) and in Bermuda (GALV-Hall's Island), and in cultured cells (GALV-X). Woolly monkey virus (WMV-previously known as simian sarcoma associated virus-SSAV), which was isolated from a brown woolly monkey (*Legothrix lagothrica*) that had been housed with a GALV-infected gibbon, clusters phylogenetically with the five GALV strains [63–70]. Despite extensive screening, GALV infection (either virus or antibodies) has never been reported in wild gibbons. There has been no definitive evidence of GALV infection or GALV-induced diseases in captive gibbons for nearly 40 years [71,72], though a serological study in 2015 detected GALV antibodies in 21 out of 76 captive gibbons in North America [73]. It has been suggested that the GALV infections in captive gibbons in the 1970s stemmed from an initial horizontal transmission event, most likely at SEATO in the mid to late 1960s followed by transportation of gibbons from that region to research facilities in North America [71,72]. Although the nature of the transmission event remains uncertain, it was probably either iatrogenic inoculation of gibbons with material derived from humans and other species or direct contact between gibbons and rodents which were held in large collections at SEATO [71].

GALVs are closely related to the recently identified ERVs in *Melomys* rodents, endemic to Australia (MbRV) [74] and North Moluccas Islands of Indonesia (MelWMV) [26,75], forming a monophyletic clade with KoRV which thus far is restricted to koalas. The recently characterized gammaretroviruses FFRV1, HPG, MmGRV, and SaGRV from the Australian *Pteropus alecto*, *Macroglossus minimus*, and *Syconycteris australis* bats form a clade basal to GALV-KoRV while HlGRV and RhGRV isolates from the Chinese *Hipposideros larvatus* and *Rhinolophus hipposideros* bats are a GALV sister clade [27,76].

1.9.2. Koala retrovirus (KoRV)

Where most ERVs are replication defective due to mutations and deletions, KoRVs represent replication-competent gammaretroviruses that can be horizontally transmitted as XRVs. Thus far, KoRV is exclusively found in koalas (*Phascolarctos cinereus*), and exists in both endogenous (KoRV-A) and exogenous forms (KoRV-B to KoRV-I) [77]. KoRVs differ primarily in sequences encoding ENV, in particular the hypervariable region of the receptor binding domain (RBD) [77]. KoRV-A is a recently endogenized virus that retains its infectious properties, having a prevalence of 100% in the northern and 14.8% in southern koala populations [78]. Such a prevalence suggests the infection has entered the genome of koalas from north of Australia [14,25,78,79]. KoRV has been shown to be absent from potential insect vectors where the most likely source of infection is presumed to be a mammal [78]. As the natural hosts, koalas and the closely related GALV in gibbons are evolutionarily and geographically distant; the 78% nucleotide

similarity between these two viruses are considered to be a result of a CST event from an unknown vector to gibbons in Southeast Asia and koalas in Australia [29].

1.9.3. Wallace Line as a barrier to GALV-KoRV distribution

Although such CST events are common, the significance of GALV-KoRV is that infection and colonization occur in regions that are constrained by multiple biogeographical barriers such as Wallace Line. Only rodents and bats extend across this barrier that separates Southeast Asia from Australo-Papuan region (Australia and New Guinea) [46]. The number of identified GALV-KoRV related viruses in Australo-Papuan wildlife is greater than in Southeast Asia or China, with only two isolates from bat species that have historically spanned the Wallace line [80,81]. Therefore, this complex distribution suggests the GALV-KoRV clade derives from the Australo-Papuan side of the Wallace Line and that the related viruses found in Southeast Asia or China represent either human-mediated spillover or infection of taxa in the Australo-Papuan region that can cross the Wallace Line.

1.10. Objectives of this thesis

Chapter II

There is much research into KoRV dynamics with its koala host however, the shared common vector species with GALV remains obscure. The recentness of this genome invasion suggests the virus, or a close relative, may still be in a reservoir species and not extinct like the ancestors of most ERVs. The primary objective for this chapter was to screen rodent and bat species that have distributions spanning the Wallace Line. Using viral target enrichment and hybridization capture method followed by illumina sequencing, we have identified a novel GALV in multiple *Melomys leucogaster* rodents in Papua New Guinea. We named this virus, cMWMV (complete melomys woolly monkey virus). In order to characterize cMWMV, we have used various cell culture methods, fluorescence, and electron microscopy. Our results suggest that, like KoRV, cMWMVs are at the early stages of retroviral germline colonization.

Chapter III

To control the expression of the persistent ERVs in the genome, vertebrate hosts have heavily invested in restriction factors to keep pace with the rapidly evolving viruses. Bats are known carriers of different viral families resulting in many adaptations. This has recently led to efforts to elucidate the differences observed between the antiviral interactions of bats with other mammals, especially in primates. Such studies have identified that genes related to immunity in bats and primates are under a significant

degree of positive selection as a result of adaptation. When it comes to understanding the host-virus evolutionary conflicts, the focus is merely on primates and bats for their obvious biomedical importance. However it has been proposed that, while bats are highly capable recipients of retroviral CST events, rodents are more commonly the originator of these events [82] such as murine retrovirus transmission to PERV [83].

The *Melomys* rodents span the rich biodiversity of Australo-Papuan region. These rodents are classified in their own separate genus from the well studied *R. norvegicus* and *M. musculus*. This genus represents host to numerous ERVs [26,74], including our described cMWMV. However, little is known about their biology, community and environmental pressures experienced. The objective for this chapter was to measure selection pressures exerted on *Melomys* rodents by the endogenizing cMWMV. For this purpose, we used PacBio long read technology to sequence a *M. leucogaster* genome. We constructed coding sequences for multiple antiretroviral proteins and used various substitution models to measure the selection pressure on these proteins that are under positive selection in ortholog genes of primates and bats.

Chapter II: Germline Integration of Gibbon Ape Leukemia Virus in Australo-Papuan Rodents

2.1. Abstract

The Wallace Line is a biogeographical boundary separating much of Southeast Asia from the Australo-Papuan region (Australia and New Guinea). Faunal dispersion and natural pathogen transmission are restricted by this boundary. West of the Wallace Line, horizontally transmitted gibbon ape leukemia viruses (GALVs) have been isolated exclusively from captive gibbons in Thailand and two Yinpterochiroptera microbat species from China (Guangxi and Sichuan provinces) appear to have been infected naturally. East of the Wallace Line, the often vertically transmitted koala retrovirus (KoRV) and closely related gammaretroviruses such as woolly monkey virus (WMV, a basal GALV strain) have been detected in marsupials (koalas) and eutherians (rodents and bats) in the Australo-Papuan region. The number of detected GALV-like viruses in Australo-Papuan wildlife and evidence of germline invasion compared to sporadic findings of horizontal transmission in Southeast Asia and China, suggests the origin of the GALV-KoRV clade is in the former region. Using high throughput molecular methods, 280 samples of endemic bat and rodent species on both sides of the Wallace Line were screened, representing seven bat and one rodent families of this region. We identified multiple rodents (*Melomys*) from Australia and Papua New Guinea and no bat species harboring GALV-like retroviruses. Several were genomically complete, infectious in cell culture models and are at the earliest stages of integrating into the *Melomys* genome suggesting the Australo-Papuan region is a hotspot for mammalian retroviral germline colonization.

2.2. Introduction

Approximately 8-10% of vertebrate genomes are composed of ERVs [84–86]. Not all retroviral families colonize vertebrate genomes with equal frequency. Gammaretroviral ERVs, such as the murine leukemia virus (MLV) related viruses, frequently colonize the germline of various vertebrates compared to other retroviral groups [24,62]. Whereas most vertebrate retroviral colonization events completed millions of years ago, Gibbon ape leukemia virus (GALV) and the closely related koala retrovirus (KoRV) represent gammaretroviruses that have colonized or have recently begun to colonize the genomes of a variety of mammals in Southeast Asia, Australo-Papuan region (Australia, New Guinea), and Wallacea (herein

corresponds to 1924 original description; spanning Philippines apart from the Palawan) [87]. Besides the recentness of these germline integrations, the significance is that these regions represent historic biogeographical realms that have limited natural faunal dispersion by the Wallace (1863) and Huxley (1868), as western limits of Australasian fauna to Lydekker (1896-western limit of Australo-Papuan mainland fauna) Lines (Figure 2.1). To date, according to IUCN (The International Union for Conservation of Nature Red List of Threatened Species. v 2020-3., http://www.iucnredlist.org), a small number of the total species that inhabit this region including approximately eight bat families and four genera of rat (*Haeromys*, *Rattus*, *Maxomy*s and *Mus*) from the following tribes are known to have distributions that span the Wallace Line: (i) Hydromyini: *Haeromys* (Borneo and Sulawesi), *Chiropodomys* (Sunda Shelf), *Chrotomys* division (Philippines) and *Hydromyini* which includes *Melomys* (Australo-Papuan), (ii) Rattini: *Rattus* (from Asia to Philippines and Australia), *Maxomys* (from Asia to Sulawesi) and (iii) Murini: *Mus musculus*, [88].

Based on the recent phylogenetic analysis published for GALV-KoRV clade, both rodents and bats are host to retroviruses in basal and crown positions [89,90]. However, the number of identified GALV-KoRV related viruses in Australo-Papuan wildlife is greater than in Southeast Asia or China with only two isolates from bat species that have historically spanned the Wallace line [80,81]. Therefore, this complex distribution suggests the GALV-KoRV clade derives from the Australo-Papuan side of the Wallace Line, and related viruses found in Southeast Asia or China represent either human-mediated spillover or infection of taxa in the Australo-Papuan region that can cross the Wallace Line.

To determine potential reservoirs for GALV-KoRV viruses in the Australo-Papuan region and Wallacea, we screened bat and rodent species for GALV and KoRV-like viruses from both sides of the Wallace line, using pan-GALV-KoRV PCR, hybridization capture viral enrichment and a high throughput sequencing [26]. We identified genomically intact WMV, denoted complete melomys woolly monkey retrovirus (cMWMV) in two rodent species, *Melomys burtoni* and *Melomys leucogaster*, endemic to Australia and New Guinea. No bats, including members of the same species found harboring GALV relatives in Australia and China, were positive. Structural modeling of the few variable amino acids among the retrieved sequences with WMV and *in vitro* infection models suggest all cMWMVs identified are replication-competent. The cMWMVs conserved target site duplication in various *M. leugocaster* museum sample tissues indicates that cMWMV is an ERV in the midst of colonizing the genome of this rodent. No bats were positive, including members of the same species found harboring GALV relatives in Australia and China [27,76]. Our data suggest the Australo-Papuan region is a hotspot for

gammaretroviruses with the potential to infect the germlines of mammals and that cMWMV represents an additional model to KoRV for exploring the earliest stages of retroviral germline colonization.

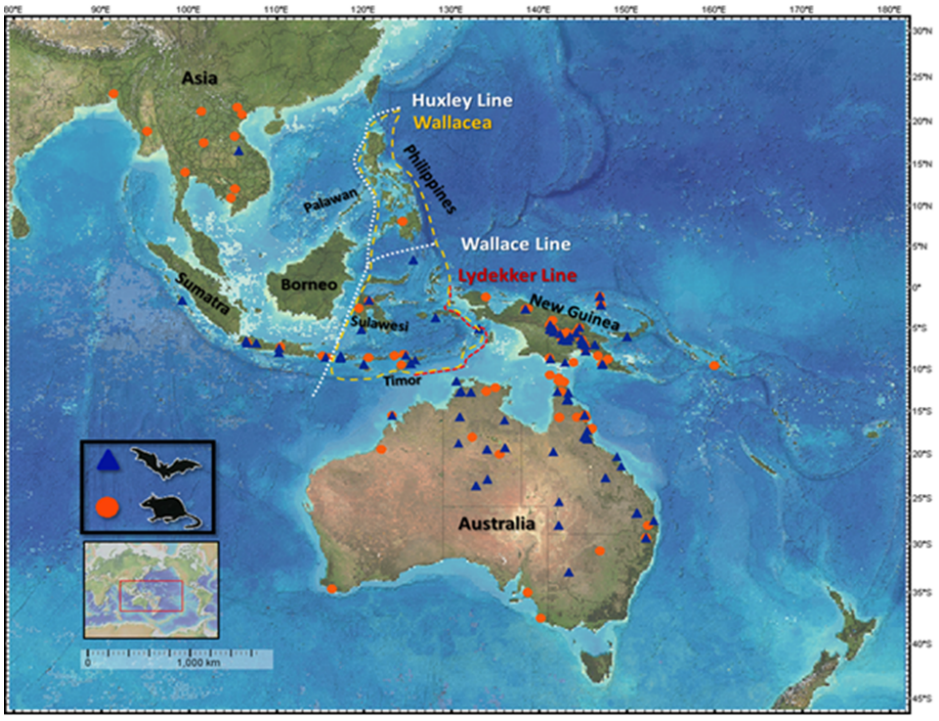


Figure 2.1. The approximate locality for bat (blue triangles) and rodent (orange circles) samples tested in this study, corresponding to sample details shown *Appendix*, **Table S2.1**. Base image generated using GeoMapApp (www.geomapapp.org) / CC BY [91]. The Lydekker (1896) (red dashes), Wallace (1863) and Huxley's extension to the Wallace Line (1868) (white dashes), and Wallacea region (1924) (yellow dashes) were drawn manually.

2.3. Materials and Methods

2.3.1. Samples and DNA extraction

A total of 280 rodent (*n* = 124) 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, four rodent genera (*Chiruromys*, *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 (Figure 2.1, *Appendix*, Table S2.1). 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.

2.3.2. 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 [92] was used to confirm that the sequences were related to GALV or KoRV.

2.3.3. 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 [93] with the modifications of Alfano et al. [94]. Each library was ligated to a unique combination of P5-P7 oligonucleotide index adapters [95] and amplified for 7 cycles with the same cycling conditions described in Alfano et al. [26]. 0.8× 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.

2.3.4. 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. [96], was used with further customization. This bait set was based on the retrieved viral oligonucleotides in the generation 5 of the ViroChip (Viro5) [97], 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-ME, 249-MF and 300-ME), medium (samples 246-MF, 290-MF, 291-MF and 292-MF) and low (samples 201-MF, 204-MF 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.

2.3.5. 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 [98], Trimmomatic v0.27 [99], Picard v1.4 (http://broadinstitute.github.io/picard), and BBMerge [100] respectively. Two pipelines were applied for identification and assembly of viral reads (Figure 2.2): Virus Integrated Pipeline (VIP) [101] in sense mode and Genome Detective [102], a web-based bioinformatics pipeline.

VIP uses Bowtie2 [103] 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 [104] 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-ME and 246-MF with 76.19% and 94.44% nucleotide identity, respectively. For 249-MF (74.53%), 290-MF (81.50%), and 291-MF (84.16%) was WMV (KT724051), 204-MF (77.71%) was MelWMV (KX059700) and for 201-MF (76%), 292-MF (86.69%) and 300-ME (74.27%) was KoRV (AB721500).

In the second approach, reads were assembled as contigs via Genome Detective. This workflow employs DIAMOND [105], 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 [106] 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-MF, 290-MF, 291-MF, 292-MF and 300-ME) were aligned by Geneious mapper (medium sensitivity/fast and iterate up to 5 times) to 249-MF consensus sequence as the most complete representative. The resulting alignments were manually curated towards the 3’ termini where mis-incorporations tend to cluster [107]. Based on majority consensus sequence, the yielded 8,459 bp was used for virus synthesis and subsequent infection assays (*Appendix*, Text file S2.1).

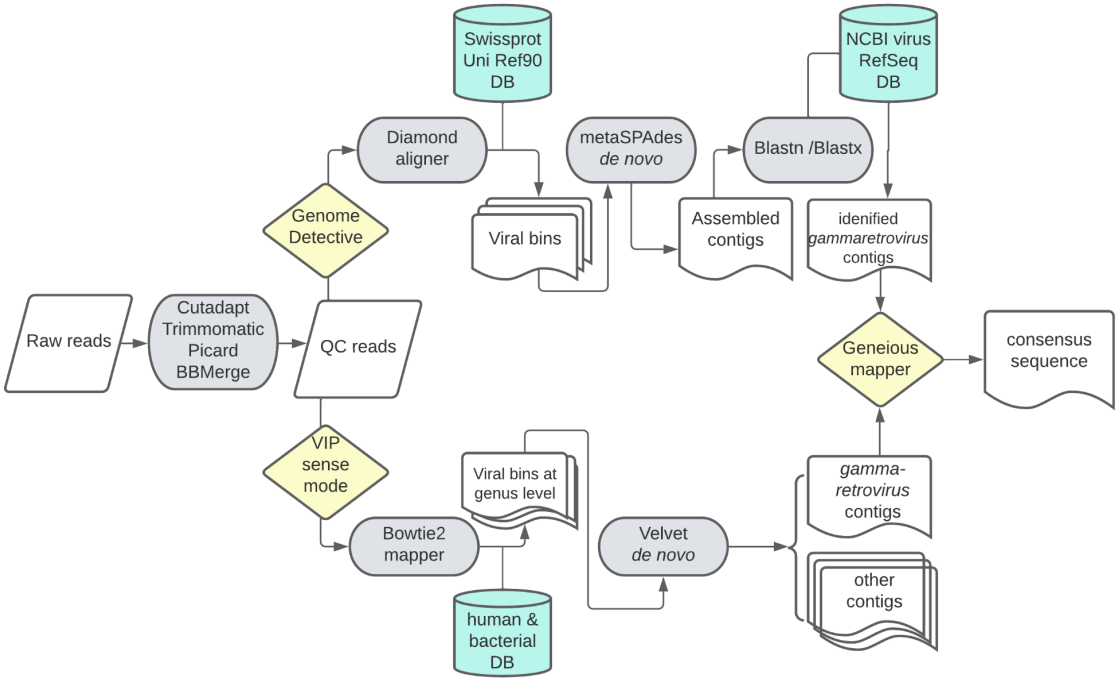


Figure 2.2. Overview of the workflow applied for identification and classification of viral reads and constructing the consensus sequences shows the curated reads were passed through Genome Detective and Virus Integrated Pipeline (VIP) in parallel. These pipelines use different databases and mappers 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 in Lucidchart, www.lucidchart.com

2.3.6. Phylogenetic analysis

Thirty eight genome sequences of gammaretroviruses (Table 2.1) from GenBank (NCBI-GenBank Flat File Release 240) were retrieved and manually curated to include the exogenous bat gammaretroviruses [27,108]. Multiple nucleotide alignments of the consensus sequences with the curated database were performed using default settings in MUSCLE [109]. Statistical selection of the best-fit model for the phylogenetic analysis performed using jModelTest [110]. Phylogenetic relationships were depicted based on the nucleotide alignments of 47 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 [111]. 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) [112]. 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 [113]. 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 [114] 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.

Table 2.1. Genbank sequences used in this study for alignment and phylogenetic trees.

	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	RD114 retrovirus (strain Sc3c)	RD114	cat	AB705392	BAM17305	BAM17305	BAM17306
34	Reticuloendotheliosis virus (strain SDAUR-S1)	REV	bird	MF185397	ASH96781	ASH96780	ASH96782
35	Rhinolophus ferrumequinum retrovirus (isolate RfRV)	RfRV	bat	JQ303225	AFA52558	AFA52559	AFA52560
36	Rhinolophus hipposideros gammaretrovirus	RhGRV	bat	MN413614	QJT93258	QJT93259	QJT93260
37	Syconycteris australis gammaretrovirus	SaGRV	bat	MN413612	QJT93252	QJT93253	QJT93254
38	Woolly monkey virus strain WMV SSAV	WMV	gibbon	KT724051	ALV83311	ALV83312	ALV83313

2.3.7. 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- 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. leugocaster* (249-MF, 290-MF, 291-MF, 292-MF and 300-ME). 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-ME, 201-MF, 204-MF and 246-MF.

2.3.8. Retroviral protein structure modeling

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 [115]. Sequence alignment visualized and annotated with Jalview v2.11.1.7 [116].

The structure characteristics of the cMWMV viral genome were examined in comparison to the WMV genome. SWISS-MODEL server [117] was used for the prediction of the three-dimensional (3D) structures of WMV and cMWMV (249-MF, a representative sequence with high sequence coverage). From the output structures predicted, only high quality protein models as defined by QMEAN4 [118] 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 [34,119].

To predict the functional effect of non-synonymous mutations, five major criteria were used [120–122]. 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) [123], SNAP (screening for non-acceptable polymorphisms) [124] and PROVEAN (protein variation effect analyzer) [125].

2.3.9. Immunofluorescence staining and microscopy of cells

The crucial region of PiT-1 receptor that allows for GALV infection in variety of mammals, including humans, gibbons, koalas and the flying fox, is quite divergent from the same region of the *M. musculus* and *M. dunni* proteins, granting NIH 3T3 and MDTF (*M. dunni* tail fibroblasts) cells, a natural resistance to GALVs infection [27,126,127]. To determine if tropism of cMWMV is comparable to GALVs, HEK293T and NIH3T3 cells were stained for PiT-1 and PiT-2 proteins. HEK293T and NIH3T3 Cells seeded at $9x10^4$ cells/ml density and grown on uncoated µ-slide 8-well high glass bottom slide (Ibidi, Germany). At ~ 70% confluency, 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 (PFA), 1 h blocking the non-specific binding 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) and anti-PiT-2 (sc-377326 AE546) primary antibodies conjugated with AlexaFluor®546 (1:50) and 10 min with Hoechst 33342 membrane accessible (Thermo Fisher Scientific, USA) for nuclear counterstain (1:40), followed by 3x 5 min DPBS wash. Slides were not mounted but instead kept moist

with 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.

2.3.10. Cell cultures

One to two million cells/ml were maintained in T75 flasks with 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 sustained at 37°C under 5% CO_2 and every two days were split at a 1:10 ratio. This method was done by aspirating the medium, followed by 1xPBS wash. After PBS removal, cells were incubated for 5 minutes with 1.5 ml Trypsin/ EDTA solution to facilitate cell detachment followed by 8.5 ml DMEM medium to inactivate trypsin. Lastly, 9 ml cell suspension was replaced with fresh DMEM medium. Cells transferred into a new flask once per week.

2.3.11. Consensus sequences and viral sequence synthesis

In Geneious R9.1, the consensus sequences of cMWMV were aligned to the near complete 249-MF 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 [107]. Based on majority consensus sequence, the yielded 8459 bp for cMWMV (*Appendix* Text file S2.1) 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 (NIH 3T3) and Human embryonic kidney (HEK 293T) cells at the Institute of Virology in Charité.

2.3.12. Transfection of cMWMV and KoRV-A

The cMWMV and KoRV-A were produced through transient transfection of HEK293T cells. Five million cells were seeded in 10 cm dishes. Next day, cells were transfected with 16 µg plasmid DNA 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 30,000 rpm at 4°C for 90 minutes. Virus-containing pellets were resuspended in medium and aliquots were stored at -80°C.

2.3.13. Infection

HEK293T cells were seeded at a cell density of 2 x 10^5 cells/ml and NIH3T3 cells at 1.6 x 10^5 cells/ml in a 48 well format. Infection of cells were performed with different volumes of virus-suspension (1 µl, 10 µl and 100 µl) overnight at 37°C with 5% CO_2. Upon 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.

2.3.14. 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´, F_env_R 5´-TAGTGGGGCTATTCCTTTTA-3´ and specific probe K_env_P 5´-TCCTCTTAAGTTGCGTGTTCGGCG-3´ (product = 95 bp). DNA concentrations were calculated using standards of known DNA concentrations, consisting of plasmid dilutions that contained a defined plasmid copy number.

2.3.15. Thin sectioning of virus-infected cells and 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 before. At 48 hpi, medium was discarded and virus-infected cells were fixed with 2.5% glutaraldehyde in 0.05 M Hepes buffer (pH:7.2) and incubated at room temperature for 2h. Afterwards, µ-Dish were filled with the fixative buffer. Thin section microscopy and image processing was performed at the Laboratory for Diagnostic Electron Microscopy of Infectious Pathogens the Robert Koch-Institute.

2.4. Results

2.4.1. WMVs in the Australo-Papuan region

The degenerate oligonucleotide primer set (KOGAWM-1) was designed to amplify the *gag* gene of any GALV, KoRV and WMV. The DNA of *R. norvegicus* was used as a negative control as these viral clades are absent from *Rattus*. Various tissue samples (*n* = 280) from seven bat families, *Rattus* (span the Wallace Line), *Hydromys* and *Melomys* (span the Lydekker's Line) rodent genera were PCR-screened for the presence of GALV and KoRV-like sequences (Figure 2.1 and *Appendix*, Table S2.1). No bat samples yielded an amplicon. However five *M. leucogaster* (*n* = 10), a *R. verecundus* (*n* = 5) and a *R. niobe_sp.B* (*n* = 5) collected in Western and Southern Highland Provinces and two *M. burtoni* (*n* = 7) from Queensland of Australia yielded an amplicon which had 89-100% sequence identity (Sanger sequencing) to WMV (*Appendix*, Table S2.1 marked *). *Melomys* is one of the biggest and most diverse genera (*Melomys* = 23 species, *Pseudomys* = 23 species, *Rattus* = 26 species) [128,129] in the Australo-Papuan region with ongoing taxonomic revisions [130–133]. Though currently confined to the east side of the Wallace line, this paraphyletic group descends from a mixture of Asian and Australo-Papuan "old endemic" rodents [130–133]. The Papuan white-bellied melomys (*M. leucogaster*) overlap with moss-forest rat (*R. niobe*) and slender rat (*R. verecundus*) in diet and space, especially at the Central Cordillera [134].

2.4.2. Hybridization capture viral enrichment

Samples (89-ME, 201-MF, 204-MF, 246-MF, 249-MF, 290-MF, 291-MF, 292-MF and 300-ME) that yield an amplicon were used for Illumina library preparation and subsequent hybridization capture enrichment. The target enrichment factor for each sample was calculated from the VIP coverage information of gammaretroviruses output plot (Table 2.2), based on the following formula:

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

Table 2.2. The percentage of total reads mapping to the target genome was considered as an enrichment efficiency factor that is calculated 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 ME	4,775,007	38,767	Baboon (NC_022517)	8,507	72.92	8.12E-03
201 MF	207,628	3,090	KoRV (AB721500)	8,440	55.91	1.49E-02
204 MF	962,301	14,545	MelWMV (KX059700)	5,488	94.56	1.51E-02
246 MF	2,836,868	20,973	Baboon (NC_022517)	8,507	67.1	7.39E-03
249 MF	7,264,635	142,342	WMV (KT724051)	8,467	100	1.96E-02
290 MF	2,665,507	38,593	WMV (KT724051)	8,467	100	1.45E-02
291 MF	1,290,687	22,424	WMV (KT724051)	8,467	100	1.74E-02
292 MF	2,545,823	35,853	KoRV (AB721500)	8,440	98.34	1.41E-02
300 MF	3,870,630	77,153	KoRV (AB721500)	8,440	100	1.99E-02

The novel retroviral sequences identified here were assembled into contiguous sequences (contigs) (refer to materials and methods) and aligned to all the KoRVs, GALVs, Australasian GALV-like bat sequences and the related gammaretroviruses. These alignments were used to perform phylogenetic analysis to infer the evolutionary relationships among the viral sequences. Contigs from *R. niobe_Sp.B* (89-ME) and *R. verecundus* (246-MF) formed a clade with *R. norvegicus* (LOC102557044), while partial sequence retrieved from one of the *M. burtoni* (201-MF, contig_4 ~ 1,300 bp) grouped with viral outgroup sequences from the rodents *C. griseus* and *M. coucha*. The remaining *Melomys* consensus contigs formed a clade with WMV, while the Asian HlGRV and RhGRV group resided within GALV-KoRV where KoRV has a basal position (Figure 2.3).

As described in Hayward et al. [27], the Gblock was applied to the full genome alignments, eliminating the divergent and poorly aligned regions and was further compared to the initial alignment. Tree topology was largely congruent for the whole-genome alignment, individual genes, and the amino acid sequences (*Appendix* Figure S2.1). Thus, we conclude that the poor node support (ranging from 28 to 74) for the WMV, HlGRV, RhGRV clade is due to the low sequence diversity of these sequences rather than phylogenetic approaches employed.

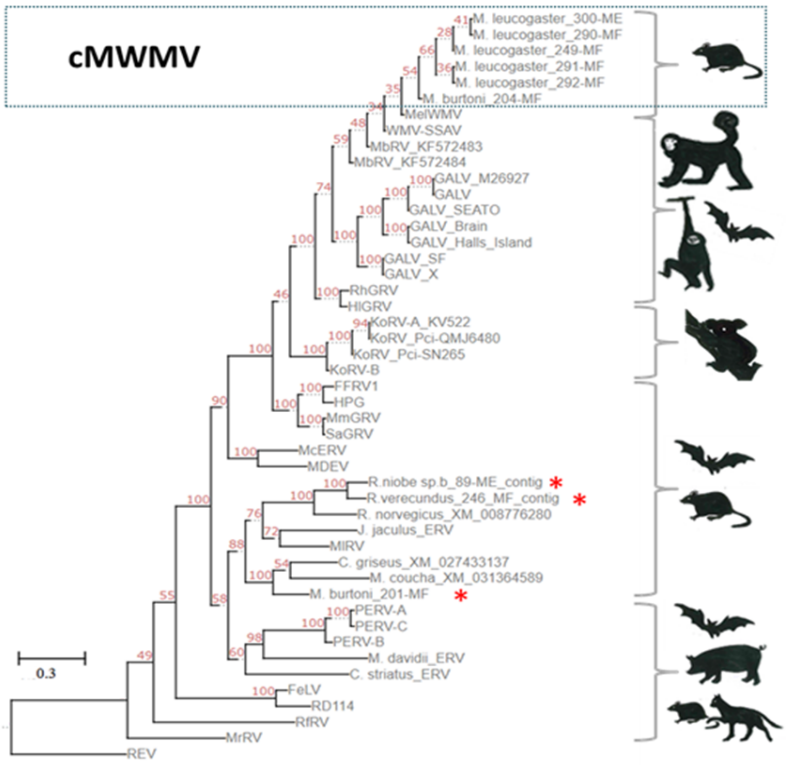


Figure 2.3. The maximum likelihood phylogenetic relationship of cMWMV inferred from complete genomic nucleotide sequences of 47 gammaretroviruses. Node support was assessed by 1000 rapid bootstrap pseudoreplicates and is indicated at each node. The newick file is visualized with the Environment for Tree Exploration (ETEv3) toolkit [135]. Branch length is with an average of 0.3 nucleotide substitutions per site. The avian reticuloendotheliosis virus (REV) was used as an outgroup and the sequences used for alignments and phylogenetic analysis are listed in **Table 2.1**. Silhouettes represent the host species. The viral contigs identified in this study are marked with a red asterix and the sequences for cMWMV clade is marked.

2.4.3. Characterization of viral integration flanking sites

The extended viral sequences for samples 249-MF, 290-MF, 291-MF, 292-MF and 300-ME were identified as the host genome integration site. Identical host flanking sequences were found each with the same target duplication site (Figure 2.4). These are different tissue samples of different *M. leucogaster* that were collected in 1985, 1987 and 2014 from four different collection sites and two different provinces in PNG. Identical integration sites in multiple tissues in multiple individuals can most parsimoniously be explained by vertical transmission indicating these WMV sequences are endogenous retroviruses (ERVs). Flanking sequences could not be extended for *M. burtoni* and the other two rodent species (89-ME, 201-MF, 204-MF and 246-MF) and therefore it could not be determined if these WMV sequences are ERVs or XRVs.

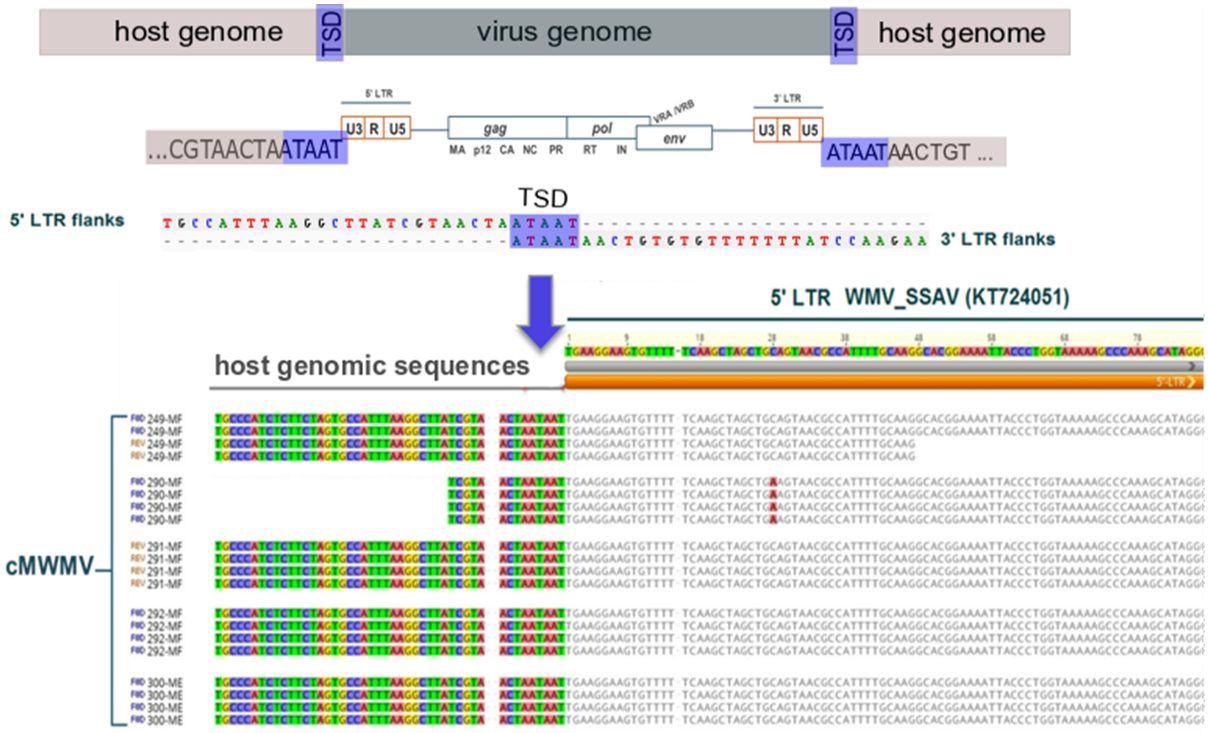


Figure 2.4. The merged sequencing reads for all samples were aligned to WMV, using Geneious mapper to retain flanking reads. Only cMWMV (no flanking sequence was identified for 204-MF) had sequences flanking both sides of the proviral integration (top). The presence of TAATA overlap between the two LTR flanks corresponds to the target-site duplication (TSD), a characteristic of retroviral integrations. The bottom figure shows identical host reads among different individuals extending 5'-LTR (3'-LTR not shown) in the representative sequences of cMWMV that along with the conserved TSDs (red boxes), is an indication that cMWMV is an ERV.

2.4.4. Structural characteristics of cMWMV

The cMWMV has retained the typical gammaretroviral structure with a genome of ~ 8.5 Kb, and unlike MelWMV has a functional *env* gene. The coding region is flanked by 5´ and 3´ untranslated LTRs (Figure 2.5). The conserved CETTG motif often found in highly infectious gammaretroviruses was identified in cMWMV (Figure 2.8). Further, cMWMV exhibited intact structural polyproteins (GAG, ENV) and functional POL with 97.3% pairwise identity to WMV and 57.6% to KoRV. The latter result indicates a closer phylogenetic relationship of cMWMV to WMV. This outcome is consistent with the results from the BLAST search, multiple sequence alignments and phylogenetic results compared to KoRV-A.

Receptor binding, mediated by the ENV is vital to the viral cellular entry process and the determinant factor in viral tropism. To predict the functional effects of the identified mutations a computational strategy with five major criteria (refer to materials and methods) was used. The results of these analyses,

which are summarized in *Appendix*, Table S2.2, strongly suggest that the identified mutations most probably do not functionally alter the proteins, as none fulfilled more than three of the criteria. Specifically, out of the 46 differences identified, only 10 are predicted to be radical substitutions according to both BLOSUM 65 and 80 scores. This finding suggests that only a few mutations are physico-chemically different enough to suggest functional changes. Furthermore, only four mutations were found in a highly conserved amino acid position based on conservation level analysis of each position. This finding also suggests that the positions of these mutations are highly variable, thus, might have a lower probability of being functionally important. Additionally, the two different prediction algorithms (SIFT and PROVEAN) that were used, found that only eight substitutions are predicted to alter protein function but only two of them predicted with high confidence by both algorithms.

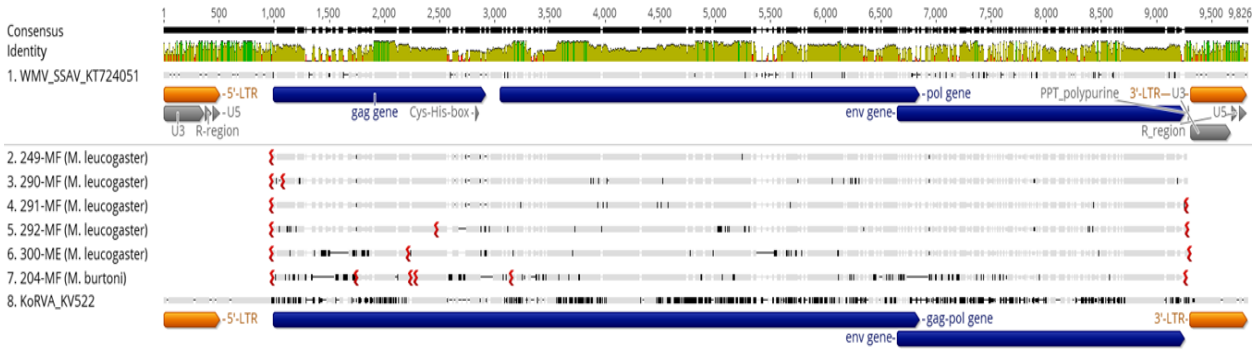


Figure 2.5. Nucleotide alignment of the assembled cMWMV sequences, KoRV (AB721500) and WMV (KT724051) showing positions of proviral genes *gag*, *pol*, *env* in blue boxes. The 5' and 3' long-terminal-repeats (LTRs) (orange boxes) with the typical U3-R-U5 structure, Cys-His-box and PPT polypurine are shown with gray boxes. Red character states indicate deletions. Consensus sequence identity is based on a 75% threshold and where character states not matching the reference sequence (WMV) 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%).

Analysis to determine whether any of the mutations change an amino acid of known function, suggested that only 12 mutations were found in known functional motifs, but none was predicted to alter the amino acid properties of the motif. Homology modeling of parts of the proteins were also performed, and tested whether a particular mutation alters the local conformation of the proteins. The latter analysis revealed that only four of the identified mutations might alter the local topography (Figure 2.9 and *Appendix* Table S2.2). Although a few of the identified mutations were predicted to alter some characteristics of the protein, none of them fulfilled more than three of the above mentioned criteria,

supporting the notion that these mutations most probably do not affect protein function. Our ENV modeling suggests like WMV, cMWMV, most likely employs PiT-1 cellular receptor.

To confirm the results of the modeling, we performed cell culture experiments employing viral vectors for cMWMV and KoRV-A as a control by infecting NIH3T3 and HEK293 cells and culturing over five days. KoRV-A viral kinetics as measured by qPCR demonstrated weak increase in viral titre in both cell lines peaking at 72 hours similar to previous results (Figure 2.7) [136]. cMWMV demonstrated a similar profile but reached higher titre in NIH3T3 cells (Figure 2.7). Staining of NIH3T3 cells with antibodies against PiT-1 and PiT-2 cellular receptors that are resistant to WMV infections suggest both receptors are present (Figure 2.10). We cannot therefore determine whether cMWMV exclusively binds to PiT-1 or is able to use PiT-2 cellular receptor as well to infect mouse cells (Figure 2.10). Electron micrographs of cMWMV in both human (not shown) and mouse cells, revealed a central electron-dense core enclosed in a spherical shaped envelope, a typical morphology of C-type viral particles that is also consistent with KoRV-A (not shown). Evidence of viral budding (Figure 2.6), suggest like other GALVs, cMWMV is capable of completing a retrovirus life cycle and forming infectious virions.

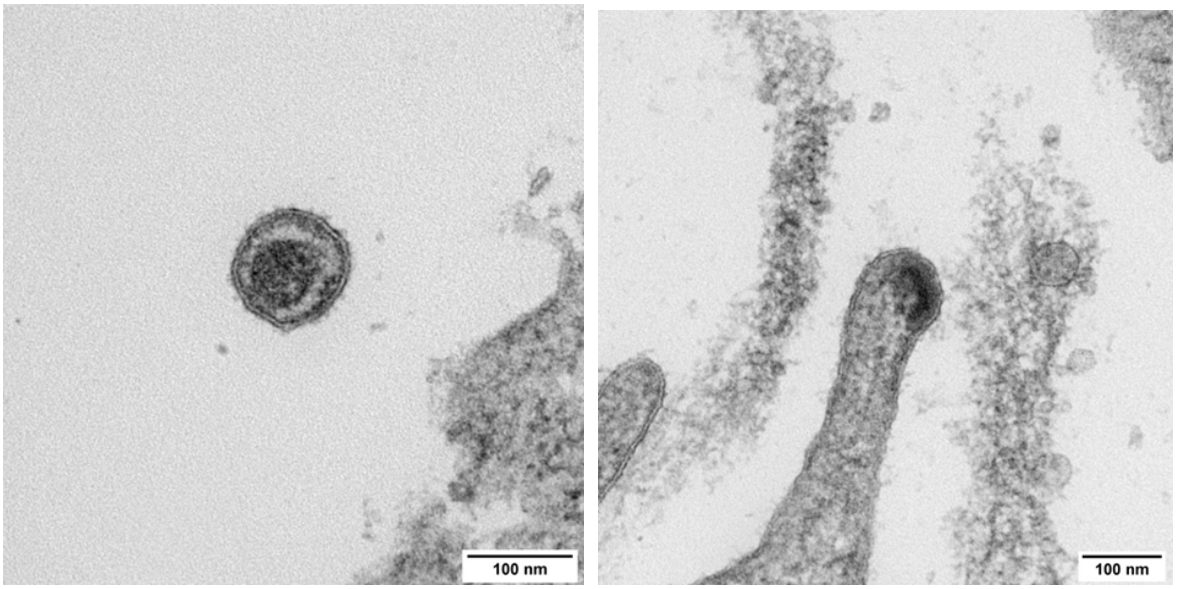


Figure 2.6. Electron micrographs of cMWMV in NIH3T3 cells at 48 hpi (cMWMV in HEK293 cells are not shown), displays a central dense core enclosed in a roughly spherical envelope with < 100 nm diameter. (right) Evidence of cMWMV budding from the plasma membrane.

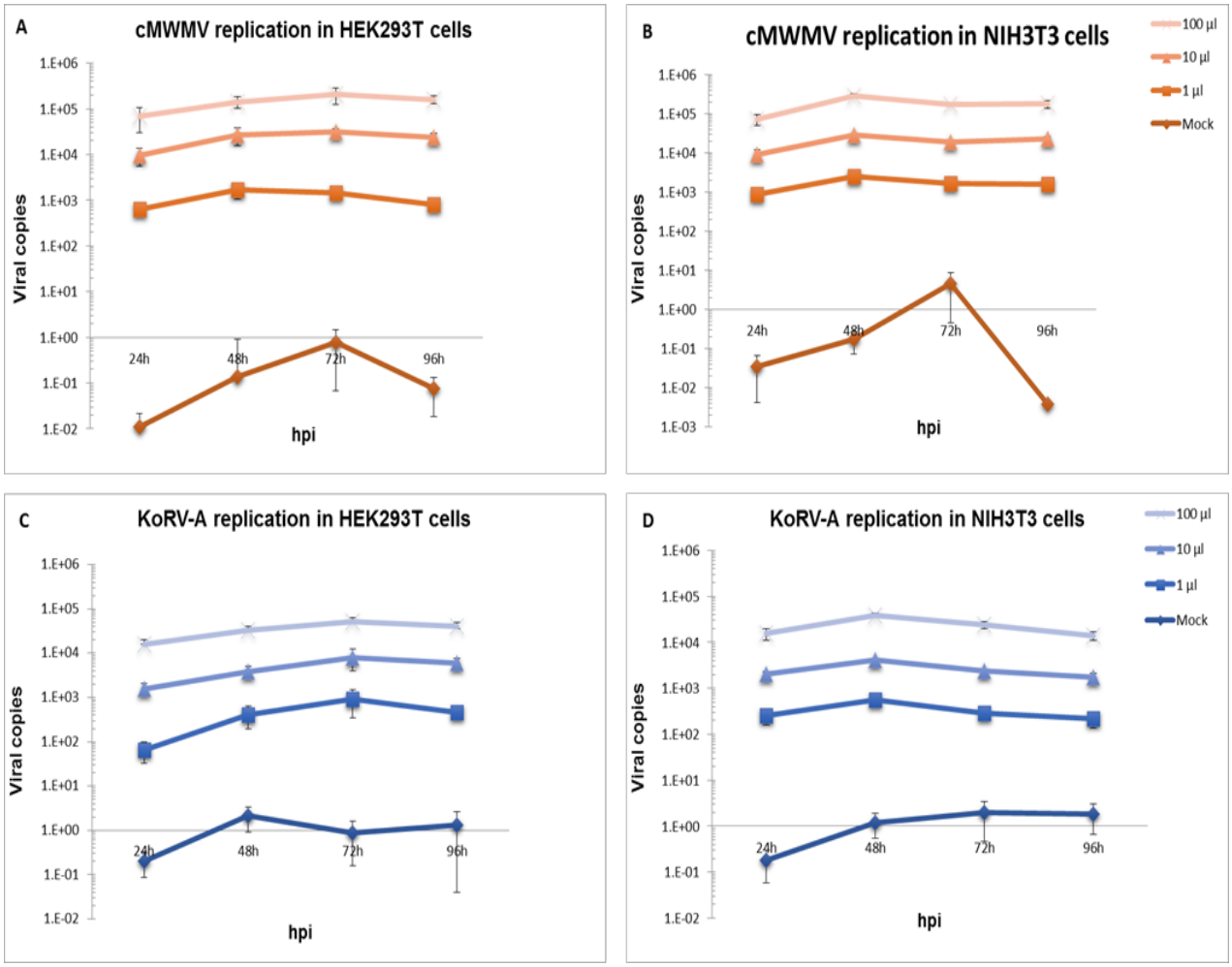


Figure 2.7. Replication kinetics of cMWMW (orange shade-top panel) and KoRV-A (blue shades-bottom panel) in human (A, C) and mouse (B, D) cell lines. Cells were infected with increasing volumes of cMWMV and KoRV-A virus-suspension (1 µl, 10 µl and 100 µl) as indicated in the legend. Supernatant was harvested at the indicated hours post infection (hpi) for cDNA synthesis and subsequent qPCR. *De novo* synthesis of viral DNA was measured and concentrations of the samples were calculated using standards of known DNA concentrations. Both cMWMV and KoRV-A show a low but stable titre, peaking at 72 hpi in both cell lines. CMWMV shows a similar profile but is able to replicate slightly better in NIH3T3 cells. Error bars represent ± SEM (HEK293T, n = 4 and NIH3T3, n =3). These data are based on two technical replicates.

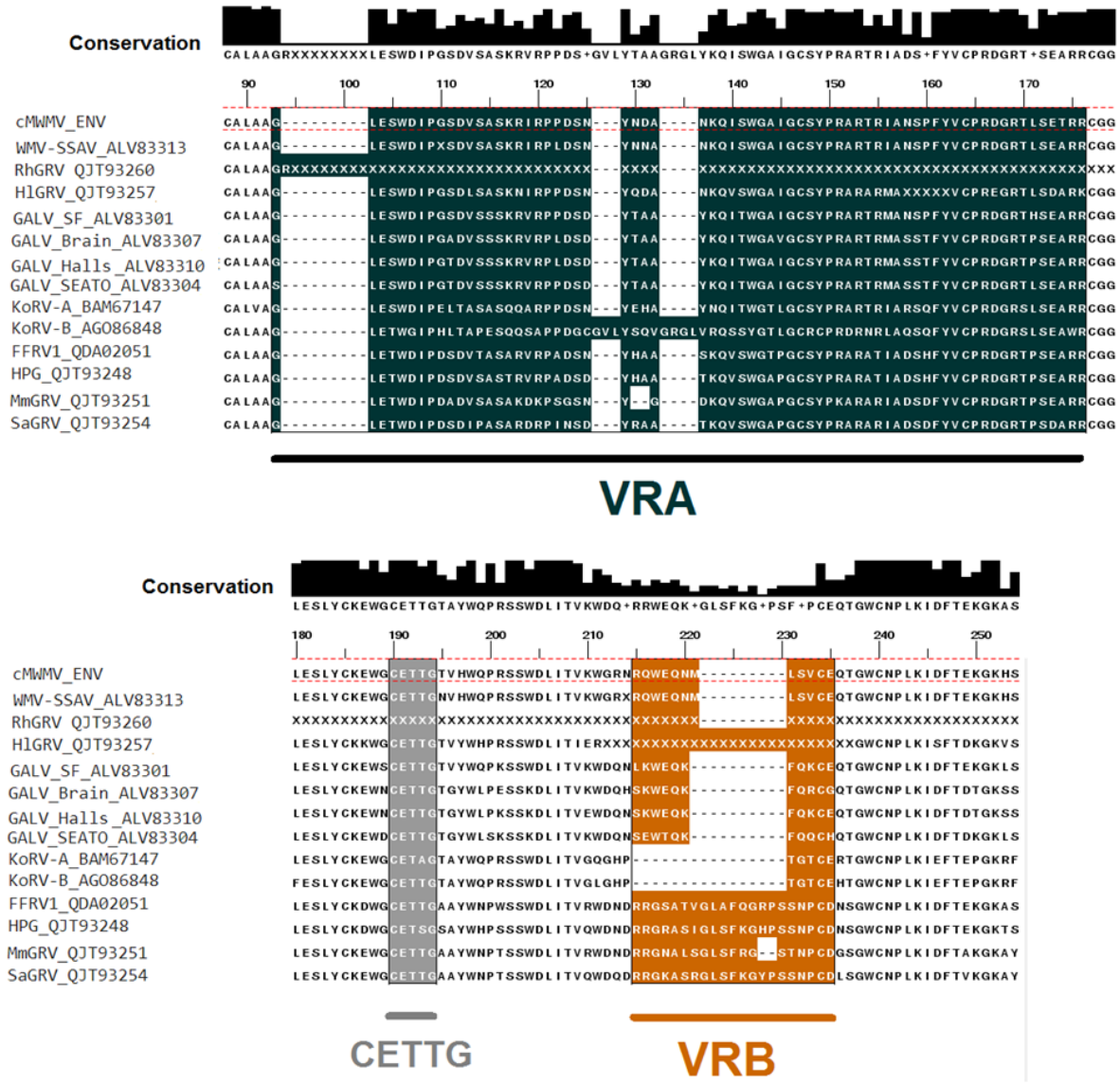


Figure 2.8. Multiple sequence alignment of the *env* genes of cMWMV with other GALV-KoRV viruses. The alignment was generated using MAFFT [115] and visualized in Jalview v2.11.1.7 [137]. Variable regions A and B (VRA; VRB) in the receptor binding domain (RBD) as determinants of cell tropism are marked. CETTG motif downstream of VRB is present in all infectious GALVs.

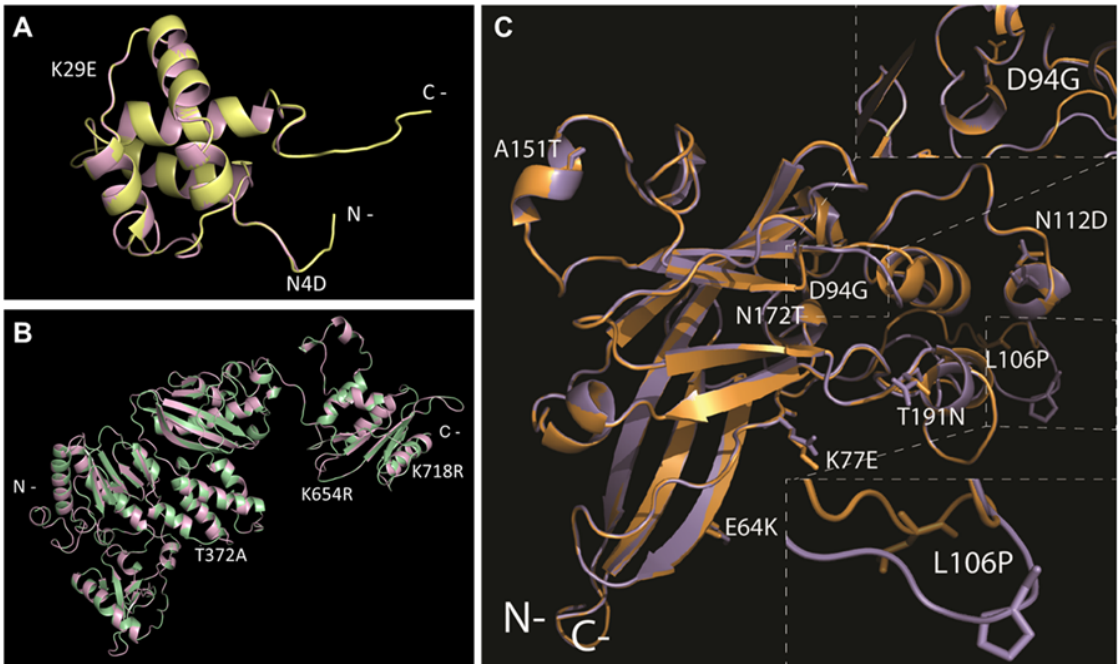


Figure 2.9. Structural superimpositions of cMWMV and WMV proteins reveal similar 3D structures for: (A) cMWMV (yellow) and WMV (pink) GAG (residues modeled 1-101) protein structure, (B) cMWMV (pink) and WMV (green) POL (residues modeled 103-740) protein structure and (C) cMWMV (purple) and WMV (orange) ENV (modeled residues 44-255) protein structure. In all three panels, the identified mutated amino acids are indicated in white.

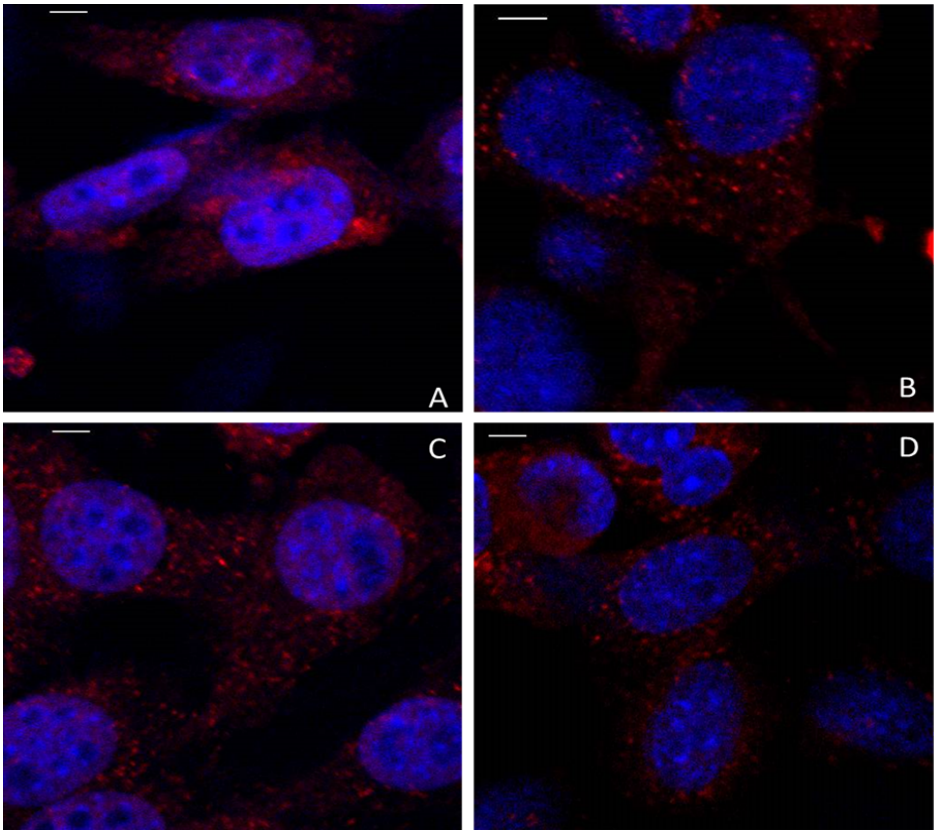


Figure 2.10. 40x confocal immunofluorescence microscopy images of PiT-1 protein expression (red) in human (A) and mouse (C) cells and PiT-2 (red) in human (B) and mouse (D). The initial anti-PiT-1 staining applied was conjugated with the green AlexaFluor®488 (images not shown), but with PiT-1 expression observed, staining repeated with AlexaFluor®546 (red). PiT-2 antibody was targeted with a primary antibody conjugated with AlexaFluor®546. Hoechst nuclear counterstain is shown in blue and bar represents 5 µm.

2.5. Discussion

Broad-scale phylogenomics indicates that 53% of all gammaretroviral-derived ERVs come from rodents [24], suggesting that rodents have transmitted XRVs and their integrated counterparts among mammals for millions of years. In laboratory mice, ongoing colonization continues to contribute to endogenization, whereas in most other species this process completed millions of years ago [62]. Given the frequency of CST and that rodent-derived ERVs are not monophyletic suggests rodents may still be a source of novel endogenizing retroviruses both among rodents and non-rodent mammals. GALV is thought to have been iatrogenically transmitted to captive gibbons as a result of experimental contamination with human material from New Guinea [66,138]. Most viruses ancestral to GALV-KoRV derive from murines such as

Asian *Mus caroli* ERV (McERV) [139], and the frequency of GALV-KoRV ERVs and XRVs in Australo-Papuan rodents suggests an overall rodent origin for this viral group.

In contrast, several bat families such as Hipposideridae [80], Pteropodidae, and Rhinolophidae [81] are documented to have crossed the Wallace Line, but only few GALV-like retroviral sequences have been reported in these species. In Queensland, Australia, two *P. alecto* [27,108] were found to harbor a GALV-like retrovirus though the same species was negative for GALV and KoRV relatives in PNG. The seven Australian bat species (*M. minimus*, *P. alecto*, *P. conspicillatus*, *P. macrotis*, *P. poliocephalus*, *P. scapulatus* and *P. vampyrus*) tested by Simmons et al. (2014) did not yield any GALV or KoRV-like gammaretroviruses. No evidence was also found in this study for any GALV-like sequences in bats from Australia, Indonesia, Laos, PNG and Timor-Leste (*n* = 156). While rodents in the Australo-Papuan region regularly have detectable GALV relatives, bats rarely carry such sequences and likely as a result of independent CST, such as *R. ferrumequinum* retrovirus (RfRV) which may have a treeshrew origin [82]. HlGRV and RhGRV were isolated from the pooled fecal and pharyngeal samples from Chinese *H. larvatus* and *R. hipposideros* bats such that the prevalence of the viruses is unclear [27]. Of 156 bats tested in the current study, 14 and 8 different species, from the Australo-Papuan *Hipposideros* and *Rhinolophus* respectively, tested negative for GALV and KoRV related viruses. Thus, WMVs in *Melomys* rodents are common but in bats exclusively exogenous and are sporadically detected, suggesting recent transmission.

Based on the phylogenetic analysis, bats and rodents are host to viruses in both basal and crown positions within the GALV-KoRV clade. Several bat sequences were successive sister lineages to the GALV-KoRV clade while the ancestral sequences of the GALV-KoRV clade are associated with rodent hosts. This interpretation is consistent with the evolutionary history of the entire gammaretroviral group, which shows a transition of rodent to non-rodent lineages [24]. While viruses identified in bats are GALV related and those previously identified in *M. burtoni* are ERVs, cMWMV is a derived, completely intact WMV with 98.8% nucleotide identity to WMV (79.8% to KoRV-A and 78% to KoRV-B). The phylogeny indicates that all the GALVs either represent derived WMVs or a clade that recently split from WMV. KoRV represents an older lineage and its exogenous counterpart may no longer exist as KoRV began colonizing the koala genome at least 50,000 years ago [140]. Why no other taxa within the region carry sequences with higher KoRV identity is unclear. This outcome may have to do either with chance that led to lack of germline invasion in the original host, or some unknown aspect of WMV biology that allows for a more opportunistic endogenization in rodents than other members of the clade. It should be noted that applying various alignment and phylogenetic approaches failed to adequately resolve the

relationships between RhGRV-HlGRV bat derived clade and WMV-GALV clade. The high sequence identity among viruses across the genome makes phylogenetic resolution difficult, emphasizing the very close relationship and minimal divergence among these viral sequences.

MelWMV is an endogenized WMV in an isolated *M. burtoni* subspecies (Halmahera) in Indonesia. It suffers large deletions in the *env* and *pol* genes, suggesting it is no longer capable of producing viral particles nor re-integrating in the host genome without a helper virus. Whereas cMWMV was detected in different tissue samples of *M. leucogaster* that were collected in different years from a subset of provinces (Southern Highland and Western Provinces) within its distribution in PNG, it was not detected in Chimbu, Gulf and Sandaun Provinces (*n = 5*). Identical integration sites were detected among multiple tissues from different individuals from the same region strongly indicating the virus is endogenous and that these individuals inherited the virus as a genomic locus. The absence of cMWMV from other populations indicates it is not yet fixed in *M. leucogaster* and, unlike MelWMV, has managed to retain intact open reading frames (ORFs) for all protein coding and non-coding viral sequences. This was also true for the cMWMV detected in *M. burtoni* (*n* = 7) from Queensland (204-MF). However, the other *M. burtoni* (201-MF) was too fragmented for a complete analysis (only the *gag* gene was characterized). Overall, the cMWMVs identified are more broadly distributed and retrain fully intact genomes.

In addition to the intact ORFs, cMWMV encoded polypeptides likely retained their original functions. This idea is supported by the minimal alterations in the predicted function and structure of the GAG, POL and ENV proteins because none of the identified mutations are predicted to result in functional changes between the cMWMV and WMV. This high level of structural and functional conservation could in turn suggest that cMWMV might employ the ubiquitous sodium-dependent phosphate transporter (PiT-1, also known as SLC20A1) protein as cellular receptor similar to WMV. This likely enabled cMWMV to infect such distinct rodent lineages such as *Melomys*, *R. norvegicus*, and *R. verecundus* as PiT-1 is highly conserved among vertebrates [26]. While we were unable to determine if cMWMV uses Pit1 exclusively or can also use Pit2 as a receptor, its ability to infect both NIH3T3 and HEK293 cells productively, form viral particles and bud from the cell membrane suggest this ERV is potentially infectious *in vivo*. The conservation of the Pit1 receptor could also explain why such diverse species including, rodents, bats, primates and marsupials have been infected by relatives of this virus clade.

Multiple germline colonizing retroviruses have been detected in mammals in the Australo-Papuan region, a very rarely observed process outside this region. In particular, replication-competent endogenous WMVs have been found frequently among *Melomys* across their biogeographical

distribution. In many cases these viruses have endogenized in their rodent hosts but only regionally. This finding is an additional confirmation that suggests, like KoRV in its koala host, the endogenization process in *Melomys* is at the earliest stages, providing additional wildlife models of the complex process of germline colonization by exogenous retroviruses. Bats show a more regionally distinct and discontinuous prevalence even within the same species and may only be sporadically infected by contact with rodent reservoirs. There is no evidence of endogenization of GALVs in bats. Nonetheless, this has enabled GALV-like viruses to be transmitted as far from the Wallace Line as Southeastern China and New Guinea. GALV-like viruses appear to be circulating, evolving and endogenizing in the endemic New Guinea rodent population and occasionally transmitting to other vertebrates resulting in viruses that are particularly apt at endogenizing, such as KoRV. The biodiversity within New Guinea is immense including within the genus *Melomys*, which contains many defined species and taxonomically uncharacterized populations most of which has yet to be screened for viruses. Our results suggest the region will be of particular interest for further identifying germline integration events and assessing the limits to which the Wallace Line prevents viral spillover into Southeast Asia and beyond.

Chapter III: Ongoing Retroviral Invasion and Adaptive Evolution of the Non-Model Organism, *Melomys* Rodents

3.1. Abstract

Rodentia is the largest order of mammals and one of the main reservoirs for zoonoses. However, our understanding of their antiviral defenses is mainly restricted to the model organisms *Mus musculus* and *Rattus norvegicus*, and coevolution with Murine Leukemia Viruses (MLV). Recent reports of Gibbon Ape Leukemia Viruses (GALVs), including our described cMWMV (complete melomys woolly monkey virus), a pathogenic *gammaretrovirus*, colonizing the genome of *Melomys* rodents, inspired this work to elucidate the evolution of five antiretroviral immune genes in *Melomys*. It is not clear whether antiviral genes of rodents are subject to the same selective pressures as orthologs in primates. Various substitution models were used to quantify the selection pressure in coding sequences of five rodent antiretroviral gene families. Our data suggest that, while these genes may have experienced positive selection at some codons (sites) in some *Mus Musculus* and *Rattus Norvegicus* lineages, the excess of synonymous sites asserts a long-term trend of purifying selection with episodic bursts of adaptive evolution. A weak intensified diversifying selection pattern in *Melomys* lineage of ZAP (zinc-finger CCCH-type antiviral protein 1) gene could indicate an effort to inhibit viral mRNA translation of endogenizing GALVs.

3.2. Introduction

Retroviruses are able to shift the burden of replication to a host's cellular machinery. Endogenous retroviruses (ERVs) represent remnants of these past infections that occupy a significant locus of the vertebrate genomes. Such antagonistic interactions trigger the innate immune system. As a result of this continuous arms race over evolutionary time, genetic variation which leads to phenotypic heterogeneity is produced. The role and fate of these variations is determined in a fitness (adaptive) landscape. As restriction factors represent the first line of defense against viral pathogens, mapping this dynamic fitness landscape is fundamental to understanding the mechanisms of inhibition and evolution of vertebrate defense systems.

The potentially unique status of bats and rodents as viral reservoirs has triggered increasing efforts to elucidate host-virus interactions. Our novel cMWMW (complete melomys woolly monkey virus),

represents a young Gibbon Ape Leukemia Virus (GALVs) that are known to be present in a wide variety of Australasian mammals [26,27,74,76]. As cMWMV is invading the genome of *Melomys* rodents, we can use this interaction to explain the development of the unusual portion of vertebrate genomes. The following is an overview of five potent restriction factors that have been shown to inhibit the complex genomes of the lentivirus HIV infection.

3.2.1. APOBEC3 (A3)

APOBEC3 (A3) genes are specific to placental mammals and are part of the vertebrate conserved AID/APOBEC (activation-induced cytidine deaminase/apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) superfamily of proteins. They are cytosine deaminases that inhibit nascent DNAs during reverse transcription by catalyzing cytosine to uracil (C→U), thereby causing guanine to adenine (G→A) hypermutations in the viral genomes, inactivating substitutions and premature stop codons [141,142]. This gene family has been under evolutionary pressure during speciation of marsupials and eutherian lineages, leaving marsupials such as koalas without APOBEC3 [143]. This restriction factor is capable of inhibiting several exogenous viral families, retroviruses, and endogenous retroelements as a consequence of duplicating APOBEC3 locus and adapting homologous genes with similar cytidine deamination functions through different inhibition mechanisms [142,144–146]. As a result of this expansion, the single copy of APOBEC3 gene in rodents (also known as mA3) ,which has not been able to efficiently block gammaretrovirus infections [141], has increased to seven copies in humans (A3A to A3H) [147]. The inability of mA3 to hypermutate gammaretroviral MLVs is reportedly caused by the antagonistic activity of the alternate glycosylated form of the viral GAG gPr80, which is conserved in gammaretroviruses [148–150]. The negative gPr80 mutants such as KoRV are shown to be infectious and replication competent [145,150]. Interestingly koalas, similar to other marsupials, lack this gene [145].

3.2.2. BST-2 (Tetherin /CD317/ HM1.24)

The interferon-stimulated bone marrow stromal antigen 2 (BST-2) gene encodes for the membrane associated glycoprotein tetherin. Although initially identified in bone marrow, this restriction factor is expressed in most mammalian cell types where its antiviral activity is related to the unique protein configuration rather than the amino acid sequence [151]. Tetherin has a bridge like topology that consist of a highly conserved cell membrane C-terminal glycosylphosphatidylinositol (GPI) anchor which inserts into the budding virus, a coiled-coil ectodomain and a diversified Transmembrane (TM) domain with a N-terminal cytoplasmic tail that remains in the host cell membrane to physically restrain (tether) the

broad spectrum of enveloped viruses from budding out of the cell membrane [152–154]. It is speculated that tetherin has aroused more than 450 million years ago in primates, rodents and a wide variety of placental mammals [155]. Tetherin is depleting in many bird species but is known to be under positive selection in primates, and even more so in bats, constantly being antagonized by several viral accessory proteins such as Vpu (viral protein U) in HIV-1 and Nef (negative factor) in simian immunodeficiency virus (SIV) amongst others [152,155–158]. Although the underlying tethering mechanism is not fully inferred, it seems that in mice it is not an essential gene and is expressed in response to interferon in most cells [159].

3.2.3. TRIM5α

The tripartite motif (TRIM) protein belongs to a large family of ubiquitin E3 ligases that are involved in many cellular processes from cell differentiations to apoptosis [160]. The TRIM5 gene exhibits antiviral activity before reverse transcription where the isoform alpha (TRIM5α) is known to be under positive selection in bats and human genomes and demonstrates varied specificity [160–164]. Post entry, viral capsid (CA) antagonizes this restriction factor as it targets viral core formation, disturbing CA uncoating in complex retroviral genomes [165] and pre-integration complex (PIC) in simple retroviruses such as MLV [166,167] which account for difference in susceptibility.

Primates have one copy with several splice variants but unlike APOBEC3 and tetherin, rodents have multiple copies [168]. This perplexing lack of gene expansion in primates is suggested to be a tradeoff strategy for the APOBEC3 family which inhibits a wider range of viruses [169]. Therefore in primates, with exception of *Pan troglodyte* ERV [162], TRIM5α largely acts as a barrier to CST events rather than retroviral inhibition. This may partly explain the distinct features of bats as reservoirs of viruses while maintaining protection against CST events.

3.2.4. SAMHD1 (Mg11)

The sterile alpha motif and histidine-aspartic acid domain-containing protein 1 are the two domains of SAMHD1 (murine ortholog Mg11) that mediates protein to protein interactions and hydrolyzes dNTPs (deoxynucleoside triphosphates) to deoxynucleosides and inorganic phosphate [170]. Intracellular dNTPs are precursors for DNA replication and repair, and a starting material for retroviral transcription. Thus by modulating dNTPs levels, SAMHD1 is able to block the early stages of HIV-1 reverse transcription, especially in dendritic and myeloid cells which are ubiquitously expressed [171,172]. SAMHD1 viral restriction mechanism differs not only for RNA and DNA viruses, but also for retroviruses with simple or

complex genomes, causing differentiated host-virus coevolutionary conflict [173]. As a result, SAMHD1 is known to be under diversifying selection across mammalian taxa [170].

3.2.5. ZAP (ZC3HAV1/PARP-13)

Poly(ADP-ribose) polymerases (PARPs) also known as ZAP (zinc-finger CCCH-type antiviral protein 1) are a gene superfamily with 17 members in humans. The most studied is PARP13 (ZC3HAV1), which has been shown to act as a zinc finger antiviral protein [174]. Via ADP-ribosylation, ZAPS are involved in a wide variety of cellular processes, post-translational regulations and genome integrity by DNA repair [175]. Because of this diversity, ZAPS are quite prevalent across all domains of life, including the last common ancestor of all eukaryotes that possessed at least five ZAPs and even horizontal gene transfer to some dsDNA viruses with homologues [174]. Therefore it is not surprising that ZAPs have evolved under positive selection in primates [176]. ZAP interacts with various host factors to achieve an optimal broad yet specific antiviral state against diverse viruses by direct degradation or viral translation inhibition [177]. Such as poly(A)-specific ribonuclease factor to shorten the viral mRNA poly(A) tail in HIV-1 infection whereas rat ZAP MLV inhibition is via direct mRNA degradation [178].

3.2.6. Signature of selection in *Melomys* rodents

To understand the evolutionary selective pressures on antiretroviral genes and specifically within the *Melomys* lineage, PacBio long sequencing read technology was used to sequence the genome of this rodent from Papua New Guinea. The sample chosen was shown to harbor the infectious endogenizing cMWMV (refer to Chapter II). The coding sequences (CDS) corresponding to antiretroviral APOBEC3, BST-2, TRIM5α, SAMHD1 and ZAP loci were constructed for this non-model organism. These restriction factors influence retrotransposition and interfere at different stages of the retroviral life cycle [179]. Also in vertebrates, these antiviral genes have a long arms race conflict with various virus families, including retroviruses, and are known to be under positive selection in primates and chiropteran [142,157,176,180,181].

The DGINN (Detection of Genetic INNovations) pipeline [182] was used for identifying ortholog sequences, identification of gene duplication and recombination events and optimizing codon-alignments which is needed for quantifying selection forces. Lastly, various substitution models from the HyPhy package [183], implemented in Datamonkey server [184] were tested to infer non-synonymous and synonymous substitution rates and to determine patterns of molecular evolution in these rodent antiviral genes.

3.3. Materials and Methods

3.3.1. Selected samples and PacBio sequencing

From the previous study (Chapter II, *Appendix* Table S2.1), extracted DNA of *Pteropus alecto* (259-BF), *M. burtoni* (204-MF) and two *M. leucogaster* (88-ME, 291-MF) from the South Australian Museum (SAM) were sent for PacBio library construction and Sequel II SMRT Cell sequencing at the Max Delbrück Center (MDC). DNA was extracted as described earlier, using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Germany). Sequencing failed due to low sample quality for 88-ME and 204-MF and for the purpose of this work, only *M. leucogaster* 291-MF that harbors cMWMV will be discussed.

3.3.2. Constructing the coding sequences (CDs)

Melomys is a genus in the Muridae family. Currently there is no annotated reference genome available for these rodents. Therefore using BWA-MEM with -x pacbio parameters [185], the PacBio circular consensus sequences (CCS) reads were aligned to *R. norvegicus* (Norway rat) genome assembly (GCF_000001895.5, Rnor_6.0). With SAMtools [186], the aligned BAM files were filtered for low quality reads, unmapped segments and reads that were not the primary alignment (supplementary alignments were allowed). The remaining reads were merged and visualized in Integrated Genomics Viewer (IGV) v2.6.2. [187]. The reference genome was distinct from *Melomys*, and instead of generating a consensus sequence, we used the PacBio reads to preserve the haplotypes within some sequencing reads and minimize alignment errors.

The longest open reading frame (ORF) is considered to be the coding sequence (CDS) [188]. To identify the longest ORF for the genes of interest in 291-MF sample, the coordinates spanning these five genes were used from the rat reference genome to annotate the aligned *M. leucogaster* reads, and using seqtk [189] those fasta sequences were extracted. The rat CDS for the genes of interest were downloaded from NCBI and using BLASTn [92] were searched against the extracted *Melomys* sequences. Where there were multiple sequence hits to the rat CDS query, the top sequence hit was taken. Pairwise sequence alignment GeneWise [190] was used to compare rat protein CDS to the *M. leucogaster* genomic DNA sequences. These assumed ORFs were examined manually and if genetic stop codons TAG, TAA or TGA were observed and the blast alignment had a single gap or runs of polynucleotides, those nucleotides were modified to “N” as ambiguous. The pairwise sequence alignment was repeated to piece together a CDS that would cover the majority of the rat CDS.

3.3.3. DGINN (Detection of Genetic INNovations) pipeline

The constructed nucleotide CDS for each gene was used as an input query for DGINN (Detect Genetic INNovations and adaptive evolution in protein-coding genes) pipeline [182] (Figure 3.1). The initial step in the pipeline is BLAST+ [191] to search for homologue genes against the NCBI nucleotide database. Customizing the pipeline is performed through a provided parameter file. For the purpose of this analysis, BLASTn was performed with default setting (e-value 10^{-4}, minimum 50% coverage and 70% identity) but the search was limited to Muridae taxa.

To shorten the analysis time, blast results were filtered by length to remove overly long sequences (sequences that are 3x longer than median, where median > 10,000 nucleotides). For each gene, the CDS (as the longest ORF) is identified with ORFinder from the EMBOSS package [192] and extracted. The codon alignment of these CDS is performed with PRANK [193] using default setting (prank -F -codon; version 150803). The phylogenetic gene tree is inferred with PhyML v3.2 [194] with DGINN's default model (HKY+G+I).

Along with the input query, DGINN requires the user provided species tree. This tree is based on ortholog genes where their Most Recent Common Ancestor (MRCA) is a speciation event [195]. Ortholog genes usually maintain the same function across species whereas gene trees evolve inside species trees and are a product of duplication, gene transfer or conversion events, giving rise to paralogs genes with various evolutionary rates [195]. Our species tree estimate was based on timetree.org database [196] for the Muridae and included 465 species (*Appendix,* Text file S3.1). To model the evolution of gene families, a reconciliation model between gene tree and species tree is performed. This step in DGINN is executed by TreeRecs [197] followed by PRANK re-alignment (*Appendix* Figure S3.1 and Figure S3.2).

Identifying recombination events, as another mechanism for variation in genomes, will limit bias in positive selection analysis [182]. To account for such events in these genes, DGINN uses default settings in GARD (Genetic Algorithm for Recombination Detection) [198] as part of the HyPhy (Hypothesis Testing using Phylogenies) package [183,199] to identify and cut recombinant fragments. After duplication and recombination analysis, the remaining non-recombinant orthologs fragments are used along with the reconciled tree for re-alignment with PRANK. This codon-aware nucleotide alignment is used as input for positive selection analysis (*Appendix*, Figure S3.2).

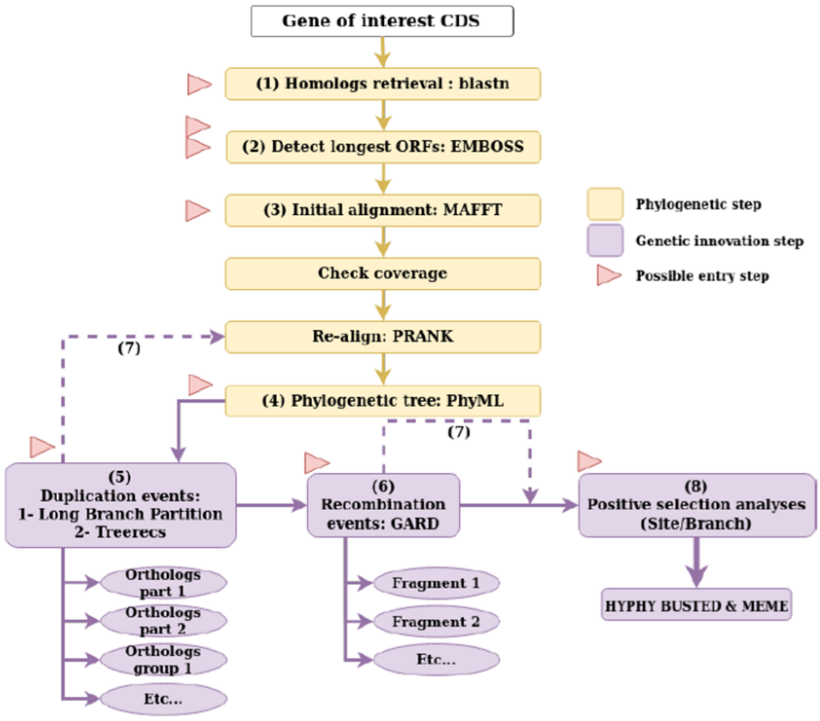


Figure 3.1. Modified Workflow of DGINN. Phylogenetic steps (yellow) happen sequentially from the entry point of the pipeline (Steps 1–4). Each genetic innovation step (purple, Steps 5, 6 and 7) is optional. All red arrowheads denote possible entry points into the pipeline. ©The Author(s) 2020. Published by Oxford University Press on behalf of Nucleic Acids Research [182].

3.3.4. Inferring selection by using substitution models

To detect signatures of selection operating in codons (sites) of these restriction factors, multiple substitution models in the HyPhy package [183] that are implemented in the Datamonkey web-server [184] were used. Two branch-site models, BUSTED (version 3.1) and aBSREL (version 2.2), and three site-level models, MEME (version 2.1.2), FUBAR (version 2.2) and FEL (version 2.1) were tested.

Given the phylogeny aware codon alignment, and the substitution model of choice, HyPhy obtains maximum likelihood parameter estimates, and assesses the estimation variability. Presented data are fitted to models that do not allow for positive selection (null model as default in HyPhy), and alternative models that allow for positive selection with varied rates of ω. Statistical significance of positive selection is determined through a chi-squared test of the LRT (likelihood ratio test) to derive p-values. Results counted as significance with p-value ≤ 0.05 for less conservative branch-site models and p-value ≤ 0.1 for site-level data (or posterior probability ≥ 0.9 for Bayesian approximation methods).

BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification) [200] is a random effect model that tests whether a gene (across the whole phylogeny) has experienced positive selection. This model is recommended for low-divergence alignment for a relatively small (less than 10 taxons) datasets. For each gene family, this model performed without a priori hypothesis (all branches in the phylogeny were tested).

ABSREL (adaptive Branch-Site Random Effects Likelihood) [201] is a branch-site model which uses both site-level and branch-level ω heterogeneity. This model is based on the idea that ω of a particular phylogeny branch is independent from other branches because of varying ω rate class. The optimal number of rate categories per branch is inferred, using a small-sample Akaike Infor-mation Criterion correction (AICc). Thereby aBSREL tests whether the proportion of sites at a particular phylogeny branch has a $\omega > 1$ where the significance is tested using a LRT. Exploratory analysis across the entire phylogeny was chosen by testing all branches for selection.

MEME (Mixed Effects Model of Evolution) [202] model employs a generalization of the maximum likelihood approach (p-value ≤ 0.1) to infer selection by two ω rate classes (i.e. one dS and two separate dN parameters, β- and β+) and corresponding weights representing the probability that a site evolves under each rate class at a given branch. To this end, MEME can identify sites that have undergone episodic positive selection in the past and it is one of the widely used site-level models to identify candidate sites influenced by selection dynamics.

FUBAR (Fast Unconstrained Bayesian AppRoximation) [203] is a bayesian model used to infer ω at each site and therefore unlike MEME that addresses the sites that are under selection, FUBAR scales the strength of pervasive positive or negative selection on individual sites. Thereby identifying candidate sites that would be subject to strong selective pressures (posterior probability = 0.9) across the whole gene phylogeny [204].

FEL (Fixed Effects Likelihood) [205] is a site-level model that is similar to FUBAR but with less statistical power (low values of $\omega > 1$) when compared to random effect approaches. To fit a codon model to each site, FEL estimates a value for dN and dS, asking whether the dN estimate is significantly greater than the inferred dS estimate and calculating the significance using the LRT. If we assume that some sites are only under episodic diversifying selection, MEME is more powerful than FEL but the latter is recommended for < 30 sequences in an alignment [184]. FEL was performed for comparison with a LRT of $p \leq 0.05$.

RELAX [206] is a hypothesis testing framework that asks whether the strength of natural selection has been relaxed ($K < 1$) or intensified ($K > 1$) along a specified set of test branches, where parameter K

serves as the selection intensity parameter. This model was applied to confirm aBSREL model analysis on *Melomys* ZAP branch (test branch).

3.4. Results

The constructed *Melomys* CDS for APOBEC3, BST-2, TRIM5α, SAMHD1 and ZAP have a length of 1188, 483, 1686, 1485 and 2946 bp respectively (*Appendix,* Text file S3.2 contains the nucleotide CDS). Input gene trees were corrected by minimizing the duplication and loss score [197]. Through reconciliation, 24 duplications with 103 losses identified for APOBEC3, 3 duplications with 111 losses identified for BST-2, 17 duplications with 97 losses identified for SAMHD1, 26 duplications with 188 losses identified for TRIM5 and 14 duplications with 113 losses identified for ZAP. Using GARD as part of the DGINN pipeline, recombination events were identified with 1, 2, 2, and 4 possible breakpoints for BST-2, SAMHD1, TRIM5 and ZAP respectively. It should be noted that all the reconciled gene trees had a short phylogenetic branch length, which reflects a relative low signal of substitution (*Appendix* Figure S3.1).

3.4.1. Branches under diversifying selection

The maximum-likelihood models of sequence evolution to infer evolutionary rates as a function of structural features found significant evidence of wide-episodic diversification for APOBEC3 (p-value = 0.032) and TRIM5 (p-value = 0.00) have been subject to positive diversifying selection, either pervasive (throughout the evolutionary tree) or episodic (only on some lineages). This indicates that at least one site on at least one test branch has experienced diversifying selection (LRT, p-value ≤ 0.05). The performed LRT of the unconstrained model returned a p-value which was smaller (better fit) than the model disallowing positive selection (constrained model), thereby rejecting the null hypothesis that no episodic positive selection occurred in the alignment. The fitted unconstrained model also specified that approximately 23.56 % of the sites in APOBEC3 lineage and 15.25% in TRIM5 had a diversifying ω rate of 2.52 and 3.03 respectively. The two class rate in APOBEC3 (ω1 = ω2) is probably caused by low-complexity data (*Appendix* Table S3.1). BUSTED returned a p-value of 0.006 for ZAP only when tested for selection on a specific *priori* (*Melomys*) of foreground branches. However, this result was deemed inconclusive because: (a) data best fitted to the constrained model where class rate ω1 and ω2 collapsed to zero and 47.89% of the branch had ω3 =1 which indicates lack of support on diversifying signal (ω value of zero typically means that a particular branch has no non-synonymous changes. A value of exactly one indicates that the branch has zero length and therefore ω is not inferred), (2) high rates of false positives reported with this model, especially when a proper *priori* is not set [207].

At a given branch in each gene phylogeny, aBSREL determined that few branches with a small proportion of sites in SAMHD1 (2 out of 41 branch: *M. musculus* AK151335 and *R. norvegicus* XM_039105087), TRIM5 (4 out of 71 branch: *M. musculus* XM_017312234, *Mastomys coucha* XM_031338209 and the two transmitting nodes) and ZAP (*Melomys* branch out of 32 branches) are under episodic diversifying selection (LRT at $p \leq 0.05$). The adaptive aBSREL model was the best fitted and refers to an optimized number of ω rate categories per branch thus rejecting the null model where branches are not allowed to have rate classes of $\omega > 1$. For rodent ZAP gene, the result of tests for episodic selection on individual *Melomys* branch, shows two ω rates with the diversifying class, taking on value 141.97 (0.46 % proportion of the mixture) (*Appendix* Table S3.1 and Figure 3.2). This outcome was congruence with the RELAX model, indicating that selection pressure acting on ZAP *Melomys* lineage has been intensified (K = 1.79 and is statistically significant p-value = 0.011, LTR = 6.39).

3.4.2. Sites (codons) under diversifying selection

In all the analyzed genes, MEME identified sites that experienced episodic diversifying selection with p-value $\leq$ 0.1. For all, the alternative MG94xREV codon model (a generalized extension of the MG94 model that allows for a full GTR mutation rate matrix) was best fitted, rejecting the null hypothesis. TRIM5 had the highest number of sites (58 out of 981 sites), followed by ZAP (28 out of 1072), APOBEC3 (24 out of 707), SAMHD1 (12 out of 933), and BST-2 (5 out of 183). FUBAR detected the sites evolving via pervasive diversifying and purifying selection with a posterior probability of 0.9. With this model 22 sites in TRIM5 were identified to be under pervasive diversifying selection followed by APOBEC3 (12 sites), SAMHD1 (8 sites), BST-2 (4 sites) and ZAP (4 sites). The overlapping sites from these two models are indicated in *Appendix* Table S3.2. FEL results were almost similar to FUBAR in detecting pervasive diversifying sites (APOBEC3 = 11, BST-2 = 2, SAMHD1 = 8, TRIM5 = 38 and ZAP = 12 sites). However, a greater number of sites were identified under purifying selection (APOBEC3 = 48, BST-2 = 29, SAMHD1 = 156, TRIM5 = 97 and ZAP = 152 sites)(*Appendix* Table S3.1).

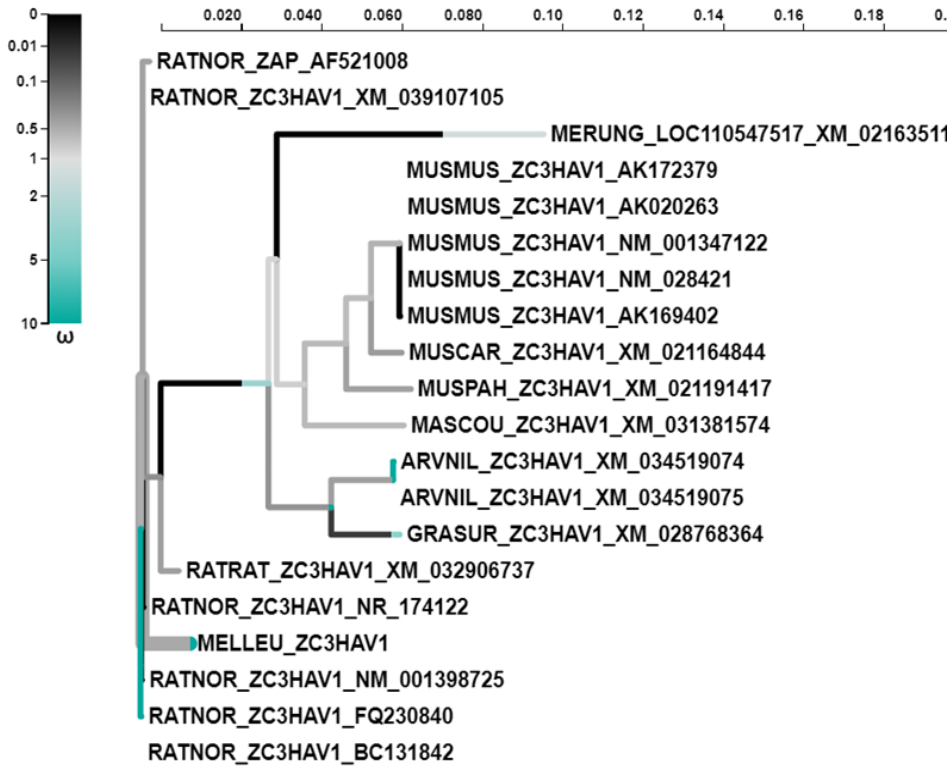


Figure 3.2

ABSREL (adaptive Branch-Site Random Effects Likelihood) analysis best fitted the adaptive model (rejecting null hypothesis) for ZAP gene family. Branches are colored by their inferred ω distribution, as indicated in the legend. Lineages identified as positive selection after correction for multiple testing are shown with thick branches, with color distributions representing the relative values and proportions of inferred ω categories. Only *M. leucogaster* (MELLEU_ZC3HAV1) lineage was determined to be significant (p-value < 0.05). The *Melomys* branch is reported by aBRSEL to be under episodic diversifying selection pressure with two ω rates (green and grey). Labels follows Spespe name trend that was required by DGINN pipeline (i.e. MELLEU corresponds to *M. leucogaster*), followed by the gene name and NCBI GenBank. Figure is automatically generated by Datamonkey server [184].

3.5. Discussion

The evolutionarily ancient orders of Chiroptera and Rodentia have vast diversity and peridomestic habitats, granting them important reservoir hosts for a number of zoonotic viral pathogens. Although rodents have twice as many species as bats and harbor more zoonotic viruses [208,209], for obvious reasons, most sequencing resources and research efforts are put into primates and bats.

The complex genome of lentiviruses HIV-1 and HIV-2 are among the most successful and thereby most studied retroviruses evading restriction factors APOBEC3 by expression of the viral accessory protein Vif [210], BST-2 by Vpu [211], TRIM5 by viral CA sequence modification [167], SAMHD1 by Vpx [171] and ZAP by mimicking low CG content of their hosts [212]. This chapter attempts to estimate selection patterns in these five restriction factors in Murid lineages and *Melomys* rodents specifically. These rodents are confined to northern Australian territories and New Guinea. They are a genus of the family Muridae and, like other genera, are species rich. However, unlike *Mus musculus* or *Rattus norvegicus*, little is known about their biology, community, and environmental pressures experienced. Given the numerous reports of ERVs in several species of this genus [26,74] and the endogenizing *gammaretrovirus* cMWMV that was described in Chapter II, we decided to perform whole genome sequencing (WGS) for a *Melomys* using PacBio long read technology. No reference genome is available for *Melomys* and this is the first time that a WGS has been performed for these rodents. For the purpose of this study, the whole genome was not assembled nor annotated, and, as described in the materials and methods, sequence alignment was performed only for genes of interest guided by the *R. norvegicus* genome assembly. Exons corresponding to these genes were extracted and coding sequences (CDS) were *de novo* assembled and manually curated. These CDS for each gene was used in DGINN pipeline and codon substitution models to infer selection pressure on gene-wide orthologs, phylogeny branches, and individual sites that would account specifically for *Melomys*.

With the explosion of empirical data, we have the ability to observe the evolutionary dynamics in population genetic processes. Variability is a frequently used factor to quantify selection pressure and infer the host-virus evolutionary conflict. A widely used method to test for selection is to compute the ratio of non-synonymous (dN or α) to synonymous (dS or β) substitution rates. This ratio is known as ω and diversifying positive and purifying negative selections are indicated when $\omega > 1$ and $\omega < 1$ respectively. The CodeML program in PAML package [213] and HyPhy [183] are the main resources for these analyses, each implementing different models in terms of their assumptions. However, the degree to which a mutation is calculated as neutral or detrimental also depends on the stability of the protein. Additionally, because viruses evolve faster than their hosts, perhaps this threshold should be drawn from statistical models that are customized for such parameters. These factors make the statistical inference for interpreting data, more real with lesser false positives, a challenging task that remains experimental.

In molecular evolutionary assessment of host-virus interactions, the different fitness landscapes that the two parties would engage with should also be considered. The small genomes of viruses packed tightly with functionally essential proteins demonstrate that their evolutionary potential is usually constrained.

This would translate to either maintaining a stabilizing neutral evolution or purifying those deleterious mutations. On the other hand, there could be two scenarios for the adaptive evolution of the restriction factors: either they will evolve towards recognizing the invading virus or evolve away from being recognized. In both scenarios, the restriction factor would need to diversify those engaging sites to increase affinity for binding to the viral proteins or decrease binding with the antagonists. The imprint of this arms race is evident with the expanded repertoire of molecular activities to create novelty in evolution, displayed as positive (diversifying) selection [179]. Persistence in this signal for sites that confer a fitness benefit would open windows of adaptive evolution and a temporary win over the antagonizing viral factors [214].

With our datasets, results from the branch-site unrestricted statistical test for episodic diversification (BUSTED) model was consistent with previous studies and also the primate orthologs, showing evidence of gene-wide episodic diversifying selection in rodent APOBEC3 (P= 0.032) and TRIM5 (P= 0.000) [169,215]. But unlike primate BST-2 [216] and mammalian SAMHD1 [170], no statistically significant gene-wide selection in these rodent genes was detected. Meanwhile, a gene-wide episodic diversifying signal was detected for ZAP only when the *Melomys* branch was selected as *priori* (foreground) against the background branches. This shows the great potential of detecting false positives when prior knowledge regarding which branches are under positive selection is not clear. This restriction profile also cannot be generalized for rodents as this has previously been shown to considerably vary between different rodent genera [217].

The second branch-site model used (aBSREL), detected episodic diversifying selection in *M. musculus* and *R. norvegicus* lineages of SAMHD1 which was in congruence of positive selection throughout mammalian evolution including Glires (rodents, rabbits, and hares) [170]. As to why this signal was not uniform across the Murid lineage that are known to be harboring ERVs, it might be explained by how deleterious polymorphisms in SAMHD1 are linked to high synthesis of viral DNA of ERVs [218]. This result was complemented by site-level models detecting a strong influence of a pervasive negative selection on a noticeable sites in SAMHD1 protein the highest number of sites, compared to other four genes, to be under pervasive purifying selection (from a total of 933 sites; FUBAR = 163 sites, FEL = 156 sites with ω = 0.26 in MEME) (*Appendix* Table S3.1). None of the lineages in Murids BST-2 and APOBEC3 was detected having experienced episodic diversification. For APOBEC3, this outcome, which is contrary to what BUSTED asserted, might be explained by aBSREL limitation on picking up signals at a lower proportion of branches [201]. Looking at the APOBEC3 Murids evolutionary history and the outcome of other site-level models, a BUSTED gene-wide diversifying signal without an expression profile should be viewed

cautiously or be re-sampled with a proper *priori* branch selection. This branch-site model (aBSREL) also detected an episodic diversifying selection in TRIM5 and 0.46% branch portion of ZAP *Melomys* lineage (*Appendix* Table S3.1). Although a canonical arms race was reported for primate ZAP blocking post-entry viral infection [176], the selection pressure endured against retroviruses is not clear. Furthermore, the RELAX model inferred an intensified selection signal for *Melomys* branch compared to the 31 reference branches. This finding is consistent with the evolutionary pattern that typically occurs during a viral infection to inhibit, in this case, viral mRNA translation. This is an intriguing finding worth exploring with future expression profiles to examine if young retroviruses such as cMWMV are the drivers of rodent ZAP evolution. In primates, the antagonizing viral factor that drove ZAP evolution is believed to be a member of the *Togaviridae* family [176].

The main advantage of the site-level MEME is that this model is sensitive to finding sites under selection in only some lineages or at only some times in the history of the gene tree (episodic), and, according to the developers, it could be used as complementary method to the branch-site aBSREl model. Because genes that have undergone positive selection are unlikely, as a whole, to have an overall $\omega > 1$, this would mean that most of their sites have undergone positive selection. As expected, all site-level models (FEL, FUBAR and MEME) showed that these front liner genes had experienced or are having an episodic diversifying selection. TRIM5 tested the highest number of sites that experienced episodic bursts of adaptive selection (MEME, $\omega = 0.75$) followed by APOBEC3 ($\omega = 0.81$), ZAP ($\omega = 0.5$), SAMHD1 ($\omega = 0.26$) and BST-2 ($\omega = 0.6$) (*Appendix* Table S3.1). This pattern in restriction factors that are involved in early stages of retroviral life cycle (from viral replication to translation) could be an indication of the ongoing retroviral colonization of rodent genomes as adaptive mutations are acquired during endogenization [17]. Furthermore, multiple copies of TRIM5 in rodents is a proof of a long antagonistic activity because gene duplication is an important driver of the acquisition of new functions that can be released from evolutionary constraints imposed on a gene.

APOBEC3 has a wider viral restriction range and a long history of coevolution across mammalian lineages. But with one gene copy, it seems that, unlike primates, rodents invested more into expanding the specialized, species-specific TRIM5 gene family. Rodents have a shorter lifespan, younger reproductive age, and, in general, smaller body size compared to primates. Therefore it could be that large populations of rodents can tolerate Mendelian disorders linked to purifying selections, or, with a short life span, they do not need to invest in costly immune gene repertoire to control all aspects of retroviral infection. This would result in a different innate immune strategy to battle the ongoing retroviral invasion than primates. It is worth mentioning that TRIM5 clusters in some species, including

rodents, are flanked by olfactory receptor genes, a main immune barrier against pathogens, that some species use as alternative adaptive landscapes [219,220].

Drawing strict conclusions from estimation of selection pressure from small scale datasets can be problematic. The identified residues were not mapped to the respective domains, which prevents predictions about biological consequences of these selected sites. Studies that have followed a conservative one-taxa approach and have manually curated datasets still detected false positives and artifacts in their results [157]. There are studies that have accounted for positive selection of an immune gene across the mammalian taxa [142,170]. However, because of the different selection pressures that different taxa endure, we have decided to limit taxon sampling to Murids. This may have reduced the power of these models to differentiate a relaxed selection from an episodic positive selection. Furthermore, we see that most phylogeny branches of the Murid gene families have a short length. These short branches are an indication of either a slow evolutionary rate or time with less selection signals. This was especially true for BST-2 where branch length varied between zero to 0.1 (*Appendix* Figure S3.1), and only five sites were detected to have experienced a diversifying selection (none overlapping with FUBAR or FEL models) (*Appendix* Table S3.2). As a result, complex models with saturating parameters are not needed to detect the evolutionary signals because maximum likelihood fitting of mixture models has numerous convergence problems, especially as the number of parameters increases. Therefore the evolutionary analysis of these rodent genes was performed with only the models in the HyPhy package that tends to employ fewer parameters, less complexity, and shorter running time than CodeML.

Overall, these results imply that the evolutionary history of rodents APOBEC3, BST-2, TRIM5, SAMHD1 and ZAP has been driven by recurrent positive selection on a small proportion of codons, against a background of strong purifying selection. A mild amplified diversifying signal in the *Melomys* lineage of the ZAP evolution is an interesting discovery that could reflect an early selective pressure exerted on this gene by the invading cMWMV.

Chapter IV: Concluding Remarks and Future Prospects

In this study, we described our novel *gammaretrovirus*, cMWMV (complete melomys woolly monkey virus), which is currently circulating in the Australo-Papuan region and endogenizing in the genome of *Melomys leucogaster*. We then quantified the selection pressure of different restriction factors in *Melomys* rodents.

This endogenizing retrovirus is the latest addition to the GALV (Gibbon Ape Leukemia Virus) group that are frequently found in wild as infectious exogenous viruses (XRV) in different bat species and defective ERVs in the Austrlo-Papuan *M. burtoni*. Although such wide cellular tropism is common in retroviruses, cMWMV holds a unique position in this spectrum. Such molecular characteristics can be beneficial not only to study early events of endogenization but also for genomic comparison to an older lineage, the koala retroviruses (KoRVs). We showed that Australo-Papuan region is a hotspot for ongoing invasions of viruses related to KoRV and GALV mediated by rodents. But as to why no other taxa within the region carry sequences with higher KoRV identity is unclear. *Melomys* rodents have a shorter lifespan and, unlike koalas, are not having a population bottleneck. *Melomys* as a new rodent model can be beneficial to elucidate how the germline of these rodents are responding compared to the koala model. How many generations does it take for cMWMV to be completely endogenized in the host genome population? Also why are GALV sequences in *M. burtoni* species so far found to be defective whereas in *M. leucogaster* are still active? Establishing *Melomys* cell line is an ambitious but extremely useful goal that can be used for future functional studies.

Endogenous retroviruses are a major contributor to the genetic diversity of vertebrate genomes. ERVs sequence insertions in the germline have consequences such as cancer, and triggers a variety of responses from the innate immune system such as transcription modification. These changes influence the host genome such as introduction of new genes or deleting the incompetent ones. Using cMWMV interaction with *Melomys* can provide insight into disease associated with purifying selection that is currently endured by these rodents.

Although we can not determine if ZAP gene evolution in *Melomys* is driven by cMWMV, mapping the identified sites to functional motifs and transcriptional experiments is needed to provide a more comprehensive view into this macroevolution pattern. In order to do so, *Melomys* reference genome assembly and annotation are necessary. Only then can we, as Barbara McClintock said, time the action of genes.

References

1. Robertson MP, Joyce GF. The origins of the RNA world. Cold Spring Harb Perspect Biol. 2012;4. doi:10.1101/cshperspect.a003608

2. Grosjean H. DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution. CRC Press; 2009.

3. White M. Faculty Opinions recommendation of The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. 2005. doi:10.3410/f.1028214.335086

4. Roy P, Fukusho A, Ritter GD, Lyon D. Evidence for genetic relationship between RNA and DNA viruses from the sequence homology of a putative polymerase gene of bluetongue virus with that of vaccinia virus: conservation of RNA polymerase genes from diverse species. Nucleic Acids Res. 1988;16: 11759–11767.

5. Gorbalenya AE, Koonin EV, Wolf YI. A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. FEBS Lett. 1990;262: 145–148.

6. Miller RH, Robinson WS. Common evolutionary origin of hepatitis B virus and retroviruses. Proc Natl Acad Sci U S A. 1986;83: 2531–2535.

7. Holmes EC. What does virus evolution tell us about virus origins? J Virol. 2011;85: 5247–5251.

8. Mustafin RN, Khusnutdinova EK. The Role of Reverse Transcriptase in the Origin of Life. Biochemistry . 2019;84: 870–883.

9. Moelling K, Broecker F, Russo G, Sunagawa S. RNase H As Gene Modifier, Driver of Evolution and Antiviral Defense. Front Microbiol. 2017;8: 1745.

10. Javier RT, Butel JS. The history of tumor virology. Cancer Res. 2008;68: 7693–7706.

11. Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, Gifford R. Larger Mammalian Body Size Leads to Lower Retroviral Activity. PLoS Pathog. 2014;10: e1004214.

12. Coffin JM, Fan H. The Discovery of Reverse Transcriptase. Annu Rev Virol. 2016;3: 29–51.

13. Hayward A, Cornwallis CK, Jern P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc Natl Acad Sci U S A. 2015;112: 464–469.

14. Greenwood AD, Ishida Y, O'Brien SP, Roca AL, Eiden MV. Transmission, Evolution, and Endogenization: Lessons Learned from Recent Retroviral Invasions. Microbiol Mol Biol Rev. 2018;82. doi:10.1128/MMBR.00044-17

15. Coffin JM. Structure and Classification of Retroviruses. The Retroviridae. Boston, MA: Springer US; 1992. pp. 19–49.

16. Boomer S, Eiden M, Burns CC, Overbaugh J. Three distinct envelope domains, variably present in subgroup B feline leukemia virus recombinants, mediate Pit1 and Pit2 receptor recognition. J Virol. 1997;71: 8116–8123.

17. Oliveira NM, Satija H, Kouwenhoven IA, Eiden MV. Changes in viral protein function that accompany retroviral endogenization. Proceedings of the National Academy of Sciences. 2007. pp. 17506–17511. doi:10.1073/pnas.0704313104

18. Urnovitz HB, Murphy WH. Human endogenous retroviruses: nature, occurrence, and clinical implications in human disease. Clinical microbiology reviews. 1996. pp. 72–99. doi:10.1128/cmr.9.1.72-99.1996

19. Maetzig T, Galla M, Baum C, Schambach A. Gammaretroviral vectors: biology, technology and application. Viruses. 2011;3: 677–713.

20. Overbaugh J, Miller AD, Eiden MV. Receptors and entry cofactors for retroviruses include single and multiple transmembrane-spanning proteins as well as newly described glycophosphatidylinositol-anchored and secreted proteins. Microbiol Mol Biol Rev. 2001;65: 371–89, table of contents.

21. Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2011.

22. Pimentel AC, Beraldo CS, Cogni R. Host-shift as the cause of emerging infectious diseases: Experimental approaches using Drosophila-virus interactions. Genet Mol Biol. 2020;44: e20200197.

23. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. Cold Spring Harb Perspect Med. 2011;1: a006841.

24. Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. Proc Natl Acad Sci U S A. 2013;110: 20146–20151.

25. Tarlinton RE, Meers J, Young PR. Retroviral invasion of the koala genome. Nature. 2006. pp. 79–81. doi:10.1038/nature04841

26. Alfano N, Michaux J, Morand S, Aplin K, Tsangaras K, Löber U, et al. Endogenous Gibbon Ape Leukemia Virus Identified in a Rodent (Melomys burtoni subsp.) from Wallacea (Indonesia). Journal of Virology. 2016. pp. 8169–8180. doi:10.1128/jvi.00723-16

27. Hayward JA, Tachedjian M, Kohl C, Johnson A, Dearnley M, Jesaveluk B, et al. Infectious KoRV-related retroviruses circulating in Australian bats. Proc Natl Acad Sci U S A. 2020;117: 9529–9536.

28. O'Hara B, Johann SV, Klinger HP, Blair DG, Rubinson H, Dunn KJ, et al. Characterization of a human gene conferring sensitivity to infection by gibbon ape leukemia virus. Cell Growth Differ. 1990;1: 119–127.

29. Oliveira NM, Farrell KB, Eiden MV. In vitro characterization of a koala retrovirus. J Virol. 2006;80: 3104–3107.

30. Shojima T, Yoshikawa R, Hoshino S, Shimode S, Nakagawa S, Ohata T, et al. Identification of a novel subgroup of Koala retrovirus from Koalas in Japanese zoos. J Virol. 2013;87: 9943–9948.

31. Xu W, Stadler CK, Gorman K, Jensen N, Kim D, Zheng H, et al. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proceedings of the National Academy of Sciences. 2013. pp. 11547–11552. doi:10.1073/pnas.1304704110

32. Xu W, Gorman K, Santiago JC, Kluska K, Eiden MV. Genetic diversity of koala retroviral envelopes. Viruses. 2015;7: 1258–1270.

33. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci U S A. 2012;109: 7385–7390.

34. Tsangaras K, Siracusa MC, Nikolaidis N, Ishida Y, Cui P, Vielgrader H, et al. Hybridization capture reveals evolution and conservation across the entire Koala retrovirus genome. PLoS One. 2014;9: e95633.

35. Smith BT, McCormack JE, Cuervo AM, Hickerson MJ, Aleixo A, Cadena CD, et al. The drivers of tropical speciation. Nature. 2014;515: 406–409.

36. Brodie JF, Helmy O, Pangau-Adam M, Ugiek G, Froese G, Granados A, et al. Crossing the (Wallace) line: local

abundance and distribution of mammals across biogeographic barriers. Biotropica. 2018. pp. 116–124. doi:10.1111/btp.12485

37. McCallum HI, Roshier DA, Tracey JP, Joseph L, Heinsohn R. Will Wallace's Line Save Australia from Avian Influenza? Ecology and Society. 2008. doi:10.5751/es-02620-130241

38. Sébastien Desfarges AC. Viral Integration and Consequences on Host Gene Expression. Viruses: Essential Agents of Life. : 147.

39. Katzourakis A, Gifford RJ. Endogenous Viral Elements in Animal Genomes. PLoS Genet. 2010;6: e1001191.

40. Vogt PK. Historical Introduction to the General Properties of Retroviruses. In: Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2011.

41. Weiss RA. On the concept and elucidation of endogenous retroviruses. Philos Trans R Soc Lond B Biol Sci. 2013;368: 20120494.

42. Denner J. Transspecies Transmission of Gammaretroviruses and the Origin of the Gibbon Ape Leukaemia Virus (GaLV) and the Koala Retrovirus (KoRV). Viruses. 2016;8. doi:10.3390/v8120336

43. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19: 199.

44. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. Mol Biol Evol. 2005;22: 814–817.

45. Bannert N, Kurth R. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet. 2006;7: 149–173.

46. Breed AC, Meers J, Sendow I, Bossart KN, Barr JA, Smith I, et al. The Distribution of Henipaviruses in Southeast Asia and Australasia: Is Wallace's Line a Barrier to Nipah Virus? PLoS ONE. 2013. p. e61316. doi:10.1371/journal.pone.0061316

47. Zhuo X, Feschotte C. Cross-Species Transmission and Differential Fate of an Endogenous Retrovirus in Three Mammal Lineages. PLoS Pathog. 2015;11: e1005279.

48. Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, et al. Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology. 2018;15: 59.

49. Stoye JP. Fv1, the mouse retrovirus resistance gene. Rev Sci Tech. 1998;17: 269–277.

50. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, et al. Paleovirology of "syncytins", retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci. 2013;368: 20120507.

51. Bock M, Stoye JP. Endogenous retroviruses and the human germline. Current Opinion in Genetics & Development. 2000. pp. 651–655. doi:10.1016/s0959-437x(00)00138-6

52. Aswad A, Katzourakis A. Paleovirology and virally derived immunity. Trends Ecol Evol. 2012;27: 627–636.

53. Löber U, Hobbs M, Dayaram A, Tsangaras K, Jones K, Alquezar-Planas DE, 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. 2018;115: 8609–8614.

54. Campos-Sánchez R, Cremona MA, Pini A, Chiaromonte F, Makova KD. Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. PLoS Comput Biol. 2016;12: e1004956.

55. McEwen GK, Alquezar-Planas DE, Dayaram A, Gillett A, Tarlinton R, Mongan N, et al. Retroviral integrations contribute to elevated host cancer rates during germline invasion. Nat Commun. 2021;12: 1316.

56. Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, et al. Lineage-Specific Expansions of Retroviral Insertions within the Genomes of African Great Apes but Not Humans and Orangutans. PLoS Biology. 2005. p. e110. doi:10.1371/journal.pbio.0030110

57. Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, et al. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol. 2012;13: R45.

58. Sharif J, Shinkai Y, Koseki H. Is there a role for endogenous retroviruses to mediate long-term adaptive phenotypic response upon environmental inputs? Philosophical Transactions of the Royal Society B: Biological Sciences. 2013. p. 20110340. doi:10.1098/rstb.2011.0340

59. Yu T, Koppetsch BS, Pagliarani S, Johnston S, Silverstein NJ, Luban J, et al. The piRNA Response to Retroviral Invasion of the Koala Genome. Cell. 2019;179: 632–643.e12.

60. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol. 2005;79: 12507–12514.

61. Joachim Denner RRT. Infection Barriers to Successful Xenotransplantation Focusing on Porcine Endogenous Retroviruses. Clin Microbiol Rev. 2012;25: 318.

62. Stocking C, Kozak CA. Murine endogenous retroviruses. Cell Mol Life Sci. 2008;65: 3383–3398.

63. Kawakami TG, Kollias GV Jr, Holmberg C. Oncogenicity of gibbon type-C myelogenous leukemia virus. Int J Cancer. 1980;25: 641–646.

64. Kawakami TG, Buckley PM. Antigenic studies on gibbon type-C viruses. Transplant Proc. 1974;6: 193–196.

65. Gallo RC, Gallagher RE, Wong-Staal F, Aoki T, Markham PD, Schetters H, et al. Isolation and tissue distribution of type-C virus and viral components from a gibbon ape (Hylobates lar) with lymphocytic leukemia. Virology. 1978;84: 359–373.

66. Todaro GJ, Lieber MM, Benveniste RE, Sherr CJ, Gibbs CJ, Carleton Gajdusek D. Infectious primate type C viruses: Three isolates belonging to a new subgroup from the brains of normal gibbons. Virology. 1975. pp. 335–343. doi:10.1016/0042-6822(75)90435-3

67. Snyder SP, Dungworth DL, Kawakami TG, Callaway E, Lau DT. Lymphosarcomas in two gibbons (Hylobates lar) with associated C-type virus. J Natl Cancer Inst. 1973;51: 89–94.

68. Parent I, Qin Y, Vandenbroucke AT, Walon C, Delferrière N, Godfroid E, et al. Characterization of a C-type retrovirus isolated from an HIV infected cell line: complete nucleotide sequence. Arch Virol. 1998;143: 1077–1092.

69. Burtonboy G, Delferriere N, Mousset B, Heusterspreute M. Isolation of a C-type retrovirus from an HIV infected cell line. Archives of Virology. 1993. pp. 289–300. doi:10.1007/bf01309661

70. Alfano N, Kolokotronis S-O, Tsangaras K, Roca AL, Xu W, Eiden MV, et al. Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses. J Virol. 2016;90: 1757–1772.

71. Brown K, Tarlinton RE. Is gibbon ape leukaemia virus still a threat? Mammal Review. 2017. pp. 53–61. doi:10.1111/mam.12079

72. McKee J, Clark N, Shapter F, Simmons G. A new look at the origins of gibbon ape leukemia virus. Virus Genes.

2017. pp. 165–172. doi:10.1007/s11262-017-1436-0

73. Siegal-Willott JL, Jensen N, Kimi D, Taliaferro D, Blankenship T, Malinsky B, et al. Evaluation of captive gibbons (Hylobates spp., Nomascus spp., Symphalangus spp.) in North American Zoological Institutions for Gibbon Ape Leukemia Virus (GALV). J Zoo Wildl Med. 2015;46: 27–33.

74. Simmons G, Clarke D, McKee J, Young P, Meers J. 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. 2014. p. e106954. doi:10.1371/journal.pone.0106954

75. Michaux B. Biogeology of Wallacea: geotectonic models, areas of endemism, and natural biogeographical units. Biol J Linn Soc Lond. 2010;101: 193–212.

76. McMichael L, Smith C, Gordon A, Agnihotri K, Meers J, Oakey J. A novel Australian flying-fox retrovirus shares an evolutionary ancestor with Koala, Gibbon and Melomys gamma-retroviruses. Virus Genes. 2019;55: 421–424.

77. Chappell KJ, Brealey JC, Amarilla AA, Watterson D, Hulse L, Palmieri C, et al. Phylogenetic Diversity of Koala Retrovirus within a Wild Koala Population. J Virol. 2017;91. doi:10.1128/JVI.01820-16

78. Simmons GS, Young PR, Hanger JJ, Jones K, Clarke D, McKee JJ, et al. Prevalence of koala retrovirus in geographically diverse populations in Australia. Aust Vet J. 2012;90: 404–409.

79. Xu W, Eiden MV. Koala Retroviruses: Evolution and Disease Dynamics. Annu Rev Virol. 2015;2: 119–134.

80. Murray SW, Campbell P, Kingston T, Zubaid A, Francis CM, Kunz TH. Molecular phylogeny of hipposiderid bats from Southeast Asia and evidence of cryptic diversity. Mol Phylogenet Evol. 2012;62: 597–611.

81. Kingston T, Rossiter SJ. Harmonic-hopping in Wallacea's bats. Nature. 2004. pp. 654–657. doi:10.1038/nature02487

82. Cui J, Tachedjian G, Wang L-F. Bats and Rodents Shape Mammalian Retroviral Phylogeny. Sci Rep. 2015;5: 16561.

83. Denner J. Transspecies transmissions of retroviruses: new cases. Virology. 2007;369: 229–233.

84. Bromham L. The human zoo: endogenous retroviruses in the human genome. Trends in Ecology & Evolution. 2002. pp. 91–97. doi:10.1016/s0169-5347(01)02394-1

85. Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev Microbiol. 2012;10: 395–406.

86. Geis FK, Goff SP. Silencing and Transcriptional Regulation of Endogenous Retroviruses: An Overview. Viruses. 2020;12. doi:10.3390/v12080884

87. Ali JR, Heaney LR. Wallace's line, Wallacea, and associated divides and areas: history of a tortuous tangle of ideas and labels. Biol Rev Camb Philos Soc. 2021. doi:10.1111/brv.12683

88. Rowe KC, Achmadi AS, Fabre P, Schenk JJ, Steppan SJ, Esselstyn JA. Oceanic islands of Wallacea as a source for dispersal and diversification of murine rodents. Journal of Biogeography. 2019. pp. 2752–2768. doi:10.1111/jbi.13720

89. Greenwood AD, Ishida Y, O'Brien SP, Roca AL, Eiden MV. Transmission, Evolution, and Endogenization: Lessons Learned from Recent Retroviral Invasions. Microbiol Mol Biol Rev. 2018;82. doi:10.1128/MMBR.00044-17

90. Hayward JA, Tachedjian M, Kohl C, Johnson A, Dearnley M, Jesaveluk B, et al. Infectious KoRV-related

retroviruses circulating in Australian bats. Proc Natl Acad Sci U S A. 2020;117: 9529–9536.

91. Ryan WBF, Carbotte SM, Coplan JO, O'Hara S, Melkonian A, Arko R, et al. Global multi-resolution topography synthesis. Geochem Geophys Geosyst. 2009;10. doi:10.1029/2008gc002332

92. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410.

93. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;2010: db.prot5448.

94. Alfano N, Courtiol A, Vielgrader H, Timms P, Roca AL, Greenwood AD. Variation in koala microbiomes within and between individuals: effect of body region and captivity status. Sci Rep. 2015;5: 10189.

95. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2012;40: e3.

96. Alfano N, Dayaram A, Axtner J, Tsangaras K, Kampmann M-L, Mohamed A, et al. Non-invasive surveys of mammalian viruses using environmental DNA. doi:10.1101/2020.03.26.009993

97. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis. 2012;6: e1485.

98. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011. p. 10. doi:10.14806/ej.17.1.200

99. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014. pp. 2114–2120. doi:10.1093/bioinformatics/btu170

100. Bushnell B, Rood J, Singer E. BBMerge – Accurate paired shotgun read merging via overlap. PLoS One. 2017;12: e0185056.

101. Li Y, Wang H, Nie K, Zhang C, Zhang Y, Wang J, et al. VIP: an integrated pipeline for metagenomics of virus identification and discovery. Scientific Reports. 2016. doi:10.1038/srep23774

102. Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics. 2018;35: 871–873.

103. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

104. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18: 821–829.

105. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12: 59–60.

106. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19: 455–477.

107. Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and epigenomes. Nat Rev Genet. 2015;16: 395–408.

108. McMichael L, Smith C, Gordon A, Agnihotri K, Meers J, Oakey J. A novel Australian flying-fox retrovirus shares an evolutionary ancestor with Koala, Gibbon and Melomys gamma-retroviruses. Virus Genes. 2019;55: 421–424.

109. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.

110. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008;25: 1253–1256.

111. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17: 754–755.

112. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313.

113. Humphries EM, Winker K. Working through polytomies: auklets revisited. Mol Phylogenet Evol. 2010;54: 88–96.

114. Talavera G, Castresana J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Systematic Biology. 2007. pp. 564–577. doi:10.1080/10635150701472164

115. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780.

116. Procter JB, Carstairs GM, Soares B, Mourão K, Ofoegbu TC, Barton D, et al. Correction to: Alignment of Biological Sequences with Jalview. Methods Mol Biol. 2021;2231: C1.

117. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46: W296–W303.

118. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011;27: 343–350.

119. DeLano WL. The PyMOL Molecular Graphics System. Schrödinger LLC www.pymol.org Version 2, http://www.pymol.org (2002). Available: www.pymol.org

120. Hess K, Oliverio R, Nguyen P, Le D, Ellis J, Kdeiss B, 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. 2018;8: 5082.

121. Oliverio R, Nguyen P, Kdeiss B, Ord S, Daniels AJ, Nikolaidis N. Functional characterization of natural variants found on the major stress inducible 70-kDa heat shock gene, HSPA1A, in humans. Biochem Biophys Res Commun. 2018;506: 799–804.

122. Nguyen P, Hess K, Smulders L, Le D, Briseno C, Chavez CM, et al. Origin and Evolution of the Human Bcl2-Associated Athanogene-1 (BAG-1). Int J Mol Sci. 2020;21. doi:10.3390/ijms21249701

123. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31: 3812–3814.

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

125. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015. pp. 2745–2747. doi:10.1093/bioinformatics/btv195

126. Wilson C, Reitz MS, Okayama H, Eiden MV. 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. 1989;63: 2374–2378.

127. Wilson CA, Farrell KB, Eiden MV. Comparison of cDNAs encoding the gibbon ape leukaemia virus receptor from susceptible and non-susceptible murine cells. J Gen Virol. 1994;75 (Pt 8): 1901–1908.

128. Timm RM, Weijola V, Aplin KP, Donnellan SC, Flannery TF, Thomson V, et al. A new species of Rattus (Rodentia: Muridae) from Manus Island, Papua New Guinea. J Mammal. 2016;97: 861–878.

129. Database MD. Mammal Diversity Database. 2020. doi:10.5281/zenodo.4139818

130. Fabre P-H, Fitriana YS, Semiadi G, Pagès M, Aplin K, Supriatna N, et al. New record of Melomys burtoni (Mammalia, Rodentia, Murinae) from Halmahera (North Moluccas, Indonesia): a review of Moluccan Melomys. Mammalia. 2018. pp. 218–247. doi:10.1515/mammalia-2016-0137

131. Rowe KC, Reno ML, Richmond DM, Adkins RM, Steppan SJ. Pliocene colonization and adaptive radiations in Australia and New Guinea (Sahul): multilocus systematics of the old endemic rodents (Muroidea: Murinae). Mol Phylogenet Evol. 2008;47: 84–101.

132. Bryant LM, Donnellan SC, Hurwood DA, Fuller SJ. Phylogenetic relationships and divergence date estimates among Australo-Papuan mosaic-tailed rats from the Uromys division (Rodentia: Muridae). Zoologica Scripta. 2011. pp. 433–447. doi:10.1111/j.1463-6409.2011.00482.x

133. Geffen E, Rowe KC, Yom-Tov Y. Reproductive rates in Australian rodents are related to phylogeny. PLoS One. 2011;6: e19199.

134. Flannery TF. Mammals of New Guinea. Cornell University Press; 1995.

135. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010. doi:10.1186/1471-2105-11-24

136. Sarker N, Fabijan J, Owen H, Seddon J, Simmons G, Speight N, et al. Koala retrovirus viral load and disease burden in distinct northern and southern koala populations. Sci Rep. 2020;10: 263.

137. Procter JB, Mungo Carstairs G, Soares B, Mourão K, Charles Ofoegbu T, Barton D, et al. Alignment of Biological Sequences with Jalview. Methods in Molecular Biology. 2021. pp. 203–224. doi:10.1007/978-1-0716-1036-7_13

138. Brown K, Tarlinton RE. Is gibbon ape leukaemia virus still a threat? Mammal Review. 2017. pp. 53–61. doi:10.1111/mam.12079

139. Lieber MM, Sherr CJ, Todaro GJ, Benveniste RE, Callahan R, Coon HG. Isolation from the asian mouse Mus caroli of an endogenous type C virus related to infectious primate type C viruses. Proc Natl Acad Sci U S A. 1975;72: 2315–2319.

140. Ishida Y, Zhao K, Greenwood AD, Roca AL. Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. Mol Biol Evol. 2015;32: 109–120.

141. Bishop KN, Holmes RK, Sheehy AM, Davidson NO, Cho S-J, Malim MH. Cytidine deamination of retroviral DNA by diverse APOBEC proteins. Curr Biol. 2004;14: 1392–1396.

142. Ito J, Gifford RJ, Sato K. Retroviruses drive the rapid evolution of mammalian genes. Proc Natl Acad Sci U S A. 2020;117: 610–618.

143. Sawyer SL, Emerman M, Malik HS. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol. 2004;2: E275.

144. Refsland EW, Harris RS. The APOBEC3 family of retroelement restriction factors. Curr Top Microbiol Immunol. 2013;371: 1–27.

145. Nitta T, Ha D, Galvez F, Miyazawa T, Fan H. Human and murine APOBEC3s restrict replication of koala retrovirus by different mechanisms. Retrovirology. 2015;12: 1–12.

146. Esnault C, Heidmann O, Delebecque F, Dewannieux M, Ribet D, Hance AJ, et al. APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. Nature. 2005;433: 430–433.

147. Conticello SG, Thomas CJF, Petersen-Mahrt SK, Neuberger MS. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol Biol Evol. 2005;22: 367–377.

148. Rosales Gerpe MC, Renner TM, Bélanger K, Lam C, Aydin H, Langlois M-A. N-linked glycosylation protects gammaretroviruses against deamination by APOBEC3 proteins. J Virol. 2015;89: 2342–2357.

149. Stavrou S, Nitta T, Kotla S, Ha D, Nagashima K, Rein AR, et al. Murine leukemia virus glycosylated Gag blocks apolipoprotein B editing complex 3 and cytosolic sensor access to the reverse transcription complex. Proceedings of the National Academy of Sciences. 2013. pp. 9078–9083. doi:10.1073/pnas.1217399110

150. Low A, Datta S, Kuznetsov Y, Jahid S, Kothari N, McPherson A, et al. Mutation in the Glycosylated Gag Protein of Murine Leukemia Virus Results in Reduced In Vivo Infectivity and a Novel Defect in Viral Budding or Release. J Virol. 2007;81: 3685.

151. Erikson E, Adam T, Schmidt S, Lehmann-Koch J, Over B, Goffinet C, et al. In vivo expression profile of the antiviral restriction factor and tumor-targeting antigen CD317/BST-2/HM1.24/tetherin in humans. Proceedings of the National Academy of Sciences. 2011. pp. 13688–13693. doi:10.1073/pnas.1101684108

152. Neil SJD, Zang T, Bieniasz PD. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature. 2008;451: 425–430.

153. Blanco-Melo D, Venkatesh S, Bieniasz PD. Origins and evolution of tetherin, an orphan antiviral gene. Cell Host Microbe. 2016;20: 189.

154. Liu J, Chen K, Wang J-H, Zhang C. Molecular evolution of the primate antiviral restriction factor tetherin. PLoS One. 2010;5: e11904.

155. Heusinger E, Kluge SF, Kirchhoff F, Sauter D. Early Vertebrate Evolution of the Host Restriction Factor Tetherin. J Virol. 2015;89: 12154.

156. Gupta RK, Mlcochova P, Pelchen-Matthews A, Petit SJ, Mattiuzzo G, Pillay D, et al. Simian immunodeficiency virus envelope glycoprotein counteracts tetherin/BST-2/CD317 by intracellular sequestration. Proc Natl Acad Sci U S A. 2009;106: 20889–20894.

157. van der Lee R, Wiel L, van Dam TJP, Huynen MA. Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. Nucleic Acids Res. 2017;45: 10634.

158. Hayward JA, Tachedjian M, Johnson A, Irving AT, Gordon TB, Cui J, et al. Unique Evolution of Antiviral Tetherin in Bats. bioRxiv. 2020. p. 2020.04.08.031203. doi:10.1101/2020.04.08.031203

159. Rachel A. Liberatore PDB. Tetherin is a key effector of the antiretroviral activity of type I interferon in vitro and in vivo. Proc Natl Acad Sci U S A. 2011;108: 18097.

160. Nisole S, Stoye JP, Saïb A. TRIM family proteins: retroviral restriction and antiviral defence. Nat Rev Microbiol. 2005;3: 799–808.

161. Blanco-Melo D, Venkatesh S, Bieniasz PD. Intrinsic cellular defenses against human immunodeficiency viruses. Immunity. 2012;37: 399–411.

162. Kaiser SM, Malik HS, Emerman M. Restriction of an extinct retrovirus by the human TRIM5alpha antiviral

protein. Science. 2007;316: 1756–1758.

163. Sawyer SL, Wu LI, Emerman M, Malik HS. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A. 2005;102: 2832–2837.

164. Fernandes AP, Águeda-Pinto A, Pinheiro A, Rebelo H, Esteves PJ. Evolution of TRIM5 and TRIM22 in Bats Reveals a Complex Duplication Process. Viruses. 2022;14: 345.

165. Nadine Laguette MB. How Samhd1 changes our view of viral restriction. Trends Immunol. 2012;33: 26.

166. Stremlau M, Perron M, Lee M, Li Y, Song B, Javanbakht H, et al. Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor. Proc Natl Acad Sci U S A. 2006;103: 5514–5519.

167. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, Sodroski J. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. Nature. 2004;427: 848–853.

168. Semih U. Tareen ME. Human Trim5α has additional activities that are uncoupled from retroviral capsid recognition. Virology. 2011;409: 113.

169. Tareen SU, Sawyer SL, Malik HS, Emerman M. An Expanded Clade of Rodent Trim5 Genes. Virology. 2009;385: 473.

170. Monit C, Morris ER, Ruis C, Szafran B, Thiltgen G, Tsai M-HC, et al. Positive selection in dNTPase SAMHD1 throughout mammalian evolution. Proc Natl Acad Sci U S A. 2019;116: 18647–18654.

171. Laguette N, Sobhian B, Casartelli N, Ringeard M, Chable-Bessia C, Ségéral E, et al. SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. Nature. 2011;474: 654–657.

172. Franzolin E, Pontarin G, Rampazzo C, Miazzi C, Ferraro P, Palumbo E, et al. The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. Proc Natl Acad Sci U S A. 2013;110: 14272–14277.

173. Coggins SA, Mahboubi B, Schinazi RF, Kim B. SAMHD1 Functions and Human Diseases. Viruses. 2020;12. doi:10.3390/v12040382

174. Perina D, Mikoč A, Ahel J, Ćetković H, Žaja R, Ahel I. Distribution of protein poly(ADP-ribosyl)ation systems across all domains of life. DNA Repair . 2014;23: 4–16.

175. Hayakawa S, Shiratori S, Yamato H, Kameyama T, Kitatsuji C, Kashigi F, et al. ZAPS is a potent stimulator of signaling mediated by the RNA helicase RIG-I during antiviral responses. Nat Immunol. 2010;12: 37–44.

176. Kerns JA, Emerman M, Malik HS. Positive Selection and Increased Antiviral Activity Associated with the PARP-Containing Isoform of Human Zinc-Finger Antiviral Protein. PLoS Genet. 2008;4: e21.

177. Li MMH, Lau Z, Cheung P, Aguilar EG, Schneider WM, Bozzacco L, et al. TRIM25 Enhances the Antiviral Action of Zinc-Finger Antiviral Protein (ZAP). PLoS Pathog. 2017;13: e1006145.

178. Zhu Y, Chen G, Lv F, Wang X, Ji X, Xu Y, et al. Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. Proc Natl Acad Sci U S A. 2011;108: 15834–15839.

179. Goodier JL. Restricting retrotransposons: a review. Mob DNA. 2016;7. doi:10.1186/s13100-016-0070-z

180. Morrison JH, Miller C, Bankers L, Crameri G, Wang L-F, Poeschla EM. A Potent Postentry Restriction to Primate Lentiviruses in a Yinpterochiropteran Bat. MBio. 2020;11. doi:10.1128/mBio.01854-20

181. Boso G, Kozak CA. Retroviral Restriction Factors and Their Viral Targets: Restriction Strategies and Evolutionary

Adaptations. Microorganisms. 2020;8. doi:10.3390/microorganisms8121965

182. Picard L, Ganivet Q, Allatif O, Cimarelli A, Guéguen L, Etienne L. DGINN, an automated and highly-flexible pipeline for the detection of genetic innovations on protein-coding genes. Nucleic Acids Res. 2020;48: e103.

183. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21: 676–679.

184. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol Biol Evol. 2018;35: 773–777.

185. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

186. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

187. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011. pp. 24–26. doi:10.1038/nbt.1754

188. Furuno M, Kasukawa T, Saito R, Adachi J, Suzuki H, Baldarelli R, et al. CDS annotation in full-length cDNA sequence. Genome Res. 2003;13: 1478–1487.

189. Website. Available: seqtk, Toolkit for processing sequences in FASTA/Q formats. Available from: https://github.com/lh3/seqtk.

190. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47: W636–W641.

191. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST : architecture and applications. BMC Bioinformatics. 2009. doi:10.1186/1471-2105-10-421

192. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277.

193. Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008;320: 1632–1635.

194. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010. pp. 307–321. doi:10.1093/sysbio/syq010

195. Toni Gabaldón EVK. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 2013;14: 360.

196. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34: 1812–1819.

197. Comte N, Morel B, Hasić D, Guéguen L, Boussau B, Daubin V, et al. Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. Bioinformatics. 2020;36: 4822–4824.

198. Pond SLK, Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: a genetic algorithm for recombination detection. Bioinformatics. 2006. pp. 3096–3098. doi:10.1093/bioinformatics/btl474

199. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Mol Biol Evol. 2020;37: 295–299.

200. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, et al. Gene-wide identification of episodic selection. Mol Biol Evol. 2015;32: 1365–1371.

201. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Molecular Biology and Evolution. 2015. pp. 1342–1353. doi:10.1093/molbev/msv022

202. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8: e1002764.

203. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 2013;30: 1196–1205.

204. Anisimova M. Evolutionary Genomics: Statistical and Computational Methods. Springer New York; 2019.

205. Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22: 1208–1222.

206. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Molecular Biology and Evolution. 2015. pp. 820–832. doi:10.1093/molbev/msu400

207. Wisotsky SR, Kosakovsky Pond SL, Shank SD, Muse SV. Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril. Mol Biol Evol. 2020;37: 2430–2439.

208. Luis AD, Hayman DTS, O'Shea TJ, Cryan PM, Gilbert AT, Pulliam JRC, et al. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? Proc Biol Sci. 2013;280: 20122753.

209. Han BA, Schmidt JP, Bowden SE, Drake JM. Rodent reservoirs of future zoonotic diseases. Proceedings of the National Academy of Sciences. 2015. pp. 7039–7044. doi:10.1073/pnas.1501598112

210. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature. 2002;418: 646–650.

211. Van Damme N, Goff D, Katsura C, Jorgenson RL, Mitchell R, Johnson MC, et al. The interferon-induced protein BST-2 restricts HIV-1 release and is downregulated from the cell surface by the viral Vpu protein. Cell Host Microbe. 2008;3: 245–252.

212. Meagher JL, Takata M, Gonçalves-Carneiro D, Keane SC, Rebendenne A, Ong H, et al. Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences. Proc Natl Acad Sci U S A. 2019;116: 24303–24309.

213. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24: 1586–1591.

214. Mustonen V, Lässig M. Fitness flux and ubiquity of adaptive evolution. Proc Natl Acad Sci U S A. 2010;107: 4248–4253.

215. Sanville B, Dolan MA, Wollenberg K, Yan Y, Martin C, Yeung ML, et al. Adaptive Evolution of Mus Apobec3 Includes Retroviral Insertion and Positive Selection at Two Clusters of Residues Flanking the Substrate Groove. PLoS Pathog. 2010;6: e1000974.

216. Kobayashi T, Takeuchi JS, Ren F, Matsuda K, Sato K, Kimura Y, et al. Characterization of red-capped mangabey tetherin: implication for the co-evolution of primates and their lentiviruses. Sci Rep. 2014;4: 5529.

217. Yap MW, Young GR, Varnaite R, Morand S, Stoye JP. Duplication and divergence of the retrovirus restriction gene Fv1 in Mus caroli allows protection from multiple retroviruses. PLoS Genet. 2020;16.

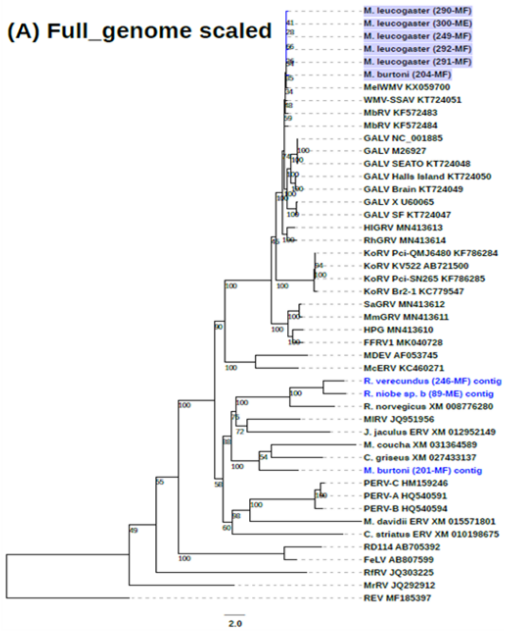
doi:10.1371/journal.pgen.1008471

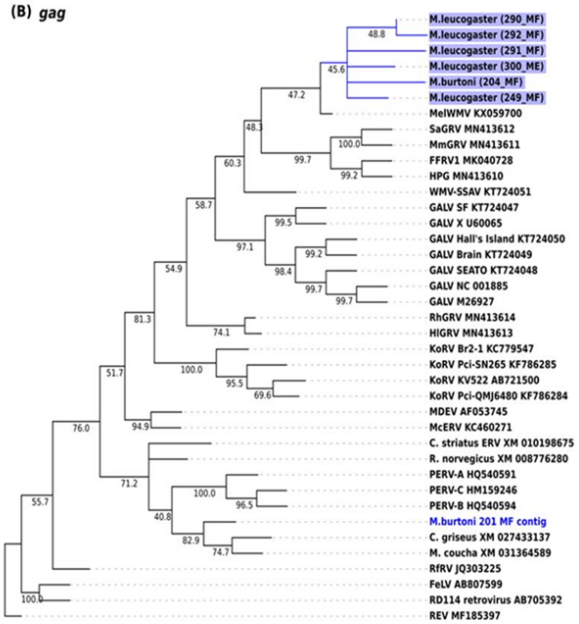
218. Herrmann A, Wittmann S, Thomas D, Shepard CN, Kim B, Ferreirós N, et al. The SAMHD1-mediated block of LINE-1 retroelements is regulated by phosphorylation. Mob DNA. 2018;9: 1–17.

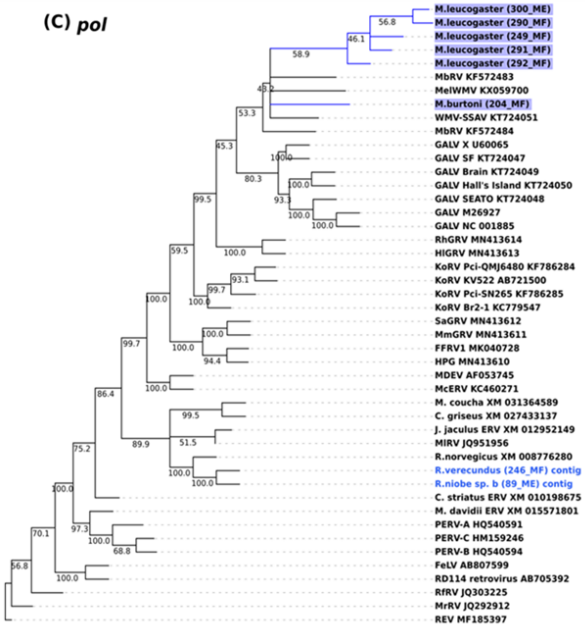
219. Boso G, Shaffer E, Liu Q, Cavanna K, Buckler-White A, Kozak CA. Evolution of the rodent Trim5 cluster is marked by divergent paralogous expansions and independent acquisitions of TrimCyp fusions. Sci Rep. 2019;9: 1–14.

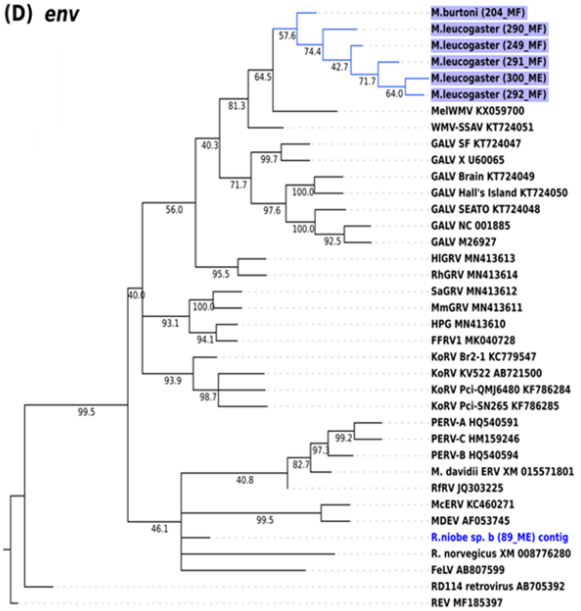
220. Sawyer SL, Emerman M, Malik HS. Discordant Evolution of the Adjacent Antiretroviral Genes TRIM22 and TRIM5 in Mammals. PLoS Pathog. 2007;3. doi:10.1371/journal.ppat.0030197

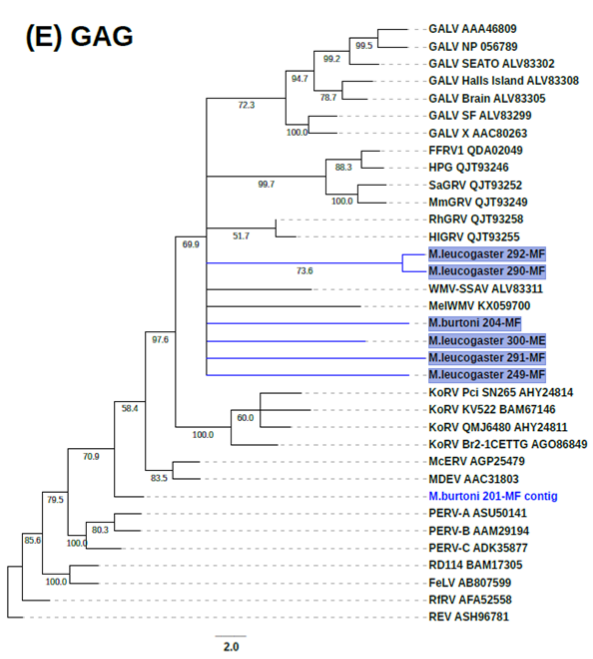
Appendix

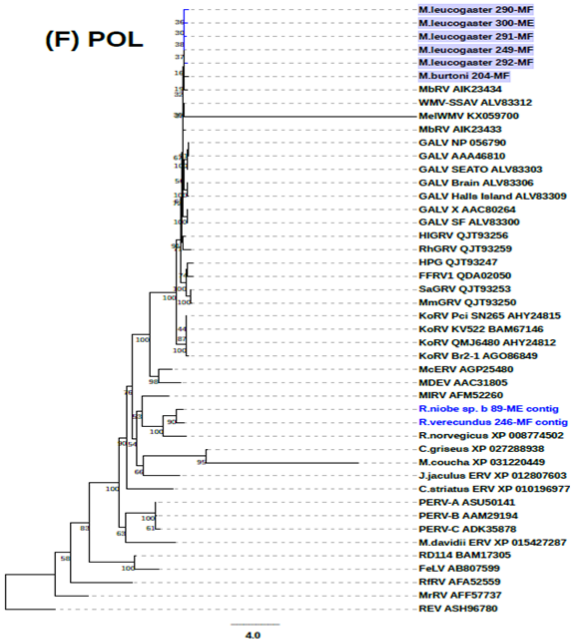

(A) Full_genome scaled
M. leucogaster (290-MF)
M. leucogaster (300-ME)
M. leucogaster (249-MF)
M. leucogaster (292-MF)
M. leucogaster (291-MF)
M. burtoni (204-MF)
MelWMV KX059700
WMV-SSAV KT724051
MbRV KF572483
MbRV KF572484
GALV NC_001885
GALV M26927
GALV SEATO KT724048
GALV Halls Island KT724050
GALV Brain KT724049
GALV X U60065
GALV SF KT724047
HIGRV MN413613
RhGRV MN413614
KoRV Pci-QMJ6480 KF786284
KoRV KV522 AB721500
KoRV Pci-SN265 KF786285
KoRV Br2-1 KC779547
SaGRV MN413612
MmGRV MN413611
HPG MN413610
FFRV1 MK040728
MDEV AF053745
McERV KC460271
R. verecundus (246-MF) contig
R. niobe sp. b (89-ME) contig
R. norvegicus XM 008776280
MlRV JQ951956
J. jaculus ERV XM 012952149
M. coucha XM 031364589
C. griseus XM 027433137
M. burtoni (201-MF) contig
PERV-C HM159246
PERV-A HQ540591
PERV-B HQ540594
M. davidii ERV XM 015571801
C. striatus ERV XM 010198675
RD114 AB705392
FeLV AB807599
RfRV JQ303225
MrRV JQ292912
REV MF185397
2.0

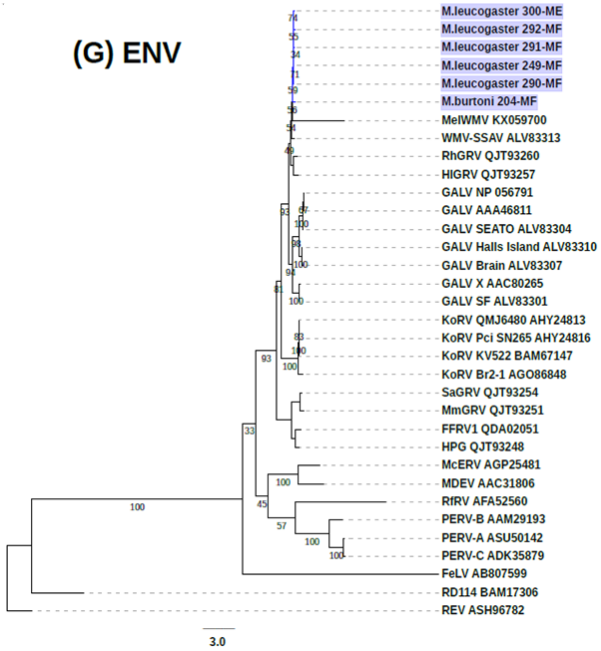

(B) gag
M.leucogaster (290_MF)
M.leucogaster (292_MF)
M.leucogaster (291_MF)
M.leucogaster (300_ME)
M.burtoni (204_MF)
M.leucogaster (249_MF)
MelWMV KX059700
SaGRV MN413612
MmGRV MN413611
FFRV1 MK040728
HPG MN413610
WMV-SSAV KT724051
GALV SF KT724047
GALV X U60065
GALV Hall's Island KT724050
GALV Brain KT724049
GALV SEATO KT724048
GALV NC 001885
GALV M26927
RhGRV MN413614
HIGRV MN413613
KoRV Br2-1 KC779547
KoRV Pci-SN265 KF786285
KoRV KV522 AB721500
KoRV Pci-QMJ6480 KF786284
MDEV AF053745
McERV KC460271
C. striatus ERV XM 010198675
R. norvegicus XM 008776280
PERV-A HQ540591
PERV-C HM159246
PERV-B HQ540594
M.burtoni 201 MF contig
C. griseus XM 027433137
M. coucha XM 031364589
RfRV JQ303225
FeLV AB807599
RD114 retrovirus AB705392
REV MF185397

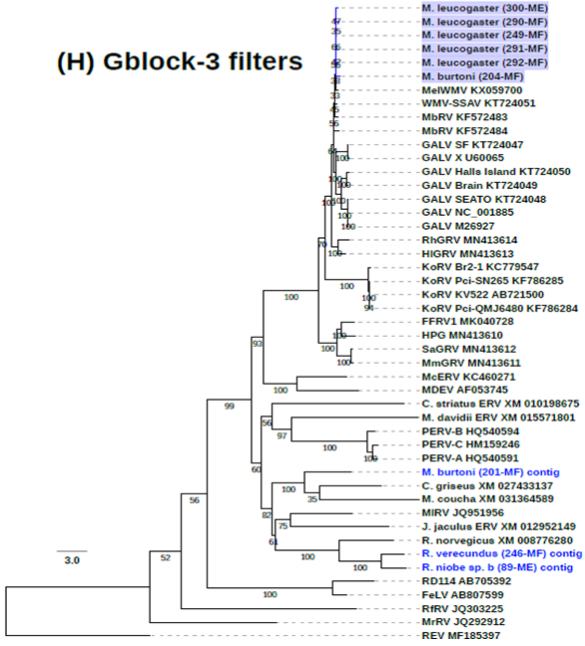

(C) pol
M.leucogaster (300_ME)
M.leucogaster (290_MF)
M.leucogaster (249_MF)
M.leucogaster (291_MF)
M.leucogaster (292_MF)
MbRV KF572483
MelWMV KX059700
M.burtoni (204_MF)
WMV-SSAV KT724051
MbRV KF572484
GALV X U60065
GALV SF KT724047
GALV Brain KT724049
GALV Hall's Island KT724050
GALV SEATO KT724048
GALV M26927
GALV NC 001885
RhGRV MN413614
HIGRV MN413613
KoRV Pci-QMJ6480 KF786284
KoRV KV522 AB721500
KoRV Pci-SN265 KF786285
KoRV Br2-1 KC779547
SaGRV MN413612
MmGRV MN413611
FFRV1 MK040728
HPG MN413610
MDEV AF053745
McERV KC460271
M. coucha XM 031364589
C. griseus XM 027433137
J. jaculus ERV XM 012952149
MlRV JQ951956
R.norvegicus XM 008776280
R.verecundus (246_MF) contig
R.niobe sp. b (89_ME) contig
C. striatus ERV XM 010198675
M. davidii ERV XM 015571801
PERV-A HQ540591
PERV-C HM159246
PERV-B HQ540594
FeLV AB807599
RD114 retrovirus AB705392
RfRV JQ303225
MrRV JQ292912
REV MF185397


(D) env
M.burtoni (204_MF)
M.leucogaster (290_MF)
M.leucogaster (249_MF)
M.leucogaster (291_MF)
M.leucogaster (300_ME)
M.leucogaster (292_MF)
MelWMV KX059700
WMV-SSAV KT724051
GALV SF KT724047
GALV X U60065
GALV Brain KT724049
GALV Hall's Island KT724050
GALV SEATO KT724048
GALV NC 001885
GALV M26927
HIGRV MN413613
RhGRV MN413614
SaGRV MN413612
MmGRV MN413611
HPG MN413610
FFRV1 MK040728
KoRV Br2-1 KC779547
KoRV KV522 AB721500
KoRV Pci-QMJ6480 KF786284
KoRV Pci-SN265 KF786285
PERV-A HQ540591
PERV-C HM159246
PERV-B HQ540594
M. davidii ERV XM 015571801
RfRV JQ303225
McERV KC460271
MDEV AF053745
R.niobe sp. b (89_ME) contig
R. norvegicus XM 008776280
FeLV AB807599
RD114 retrovirus AB705392
REV MF185397


(E) GAG
GALV AAA46809
GALV NP 056789
GALV SEATO ALV83302
GALV Halls Island ALV83308
GALV Brain ALV83305
GALV SF ALV83299
GALV X AAC80263
FFRV1 QDA02049
HPG QJT93246
SaGRV QJT93252
MmGRV QJT93249
RhGRV QJT93258
HlGRV QJT93255
M.leucogaster 292-MF
M.leucogaster 290-MF
WMV-SSAV ALV83311
MelWMV KX059700
M.burtoni 204-MF
M.leucogaster 300-ME
M.leucogaster 291-MF
M.leucogaster 249-MF
KoRV Pci SN265 AHY24814
KoRV KV522 BAM67146
KoRV QMJ6480 AHY24811
KoRV Br2-1CETTG AGO86849
McERV AGP25479
MDEV AAC31803
M.burtoni 201-MF contig
PERV-A ASU50141
PERV-B AAM29194
PERV-C ADK35877
RD114 BAM17305
FeLV AB807599
RfRV AFA52558
REV ASH96781
2.0


(F) POL
M.leucogaster 290-MF
M.leucogaster 300-ME
M.leucogaster 291-MF
M.leucogaster 249-MF
M.leucogaster 292-MF
M.burtoni 204-MF
MbRV AIK23434
WMV-SSAV ALV83312
MelWMV KX059700
MbRV AIK23433
GALV NP 056790
GALV AAA46810
GALV SEATO ALV83303
GALV Brain ALV83306
GALV Halls Island ALV83309
GALV X AAC80264
GALV SF ALV83300
HlGRV QJT93256
RhGRV QJT93259
HPG QJT93247
FFRV1 QDA02050
SaGRV QJT93253
MmGRV QJT93250
KoRV Pci SN265 AHY24815
KoRV KV522 BAM67146
KoRV QMJ6480 AHY24812
KoRV Br2-1 AGO86849
McERV AGP25480
MDEV AAC31805
MlRV AFM52260
R.niobe sp. b 89-ME contig
R.verecundus 246-MF contig
R.norvegicus XP 008774502
C.griseus XP 027288938
M.coucha XP 031220449
J.jaculus ERV XP 012807603
C.striatus ERV XP 010196977
PERV-A ASU50141
PERV-B AAM29194
PERV-C ADK35878
M.davidii ERV XP 015427287
RD114 BAM17305
FeLV AB807599
RfRV AFA52559
MrRV AFF57737
REV ASH96780
4.0


(G) ENV
M.leucogaster 300-ME
M.leucogaster 292-MF
M.leucogaster 291-MF
M.leucogaster 249-MF
M.leucogaster 290-MF
M.burtoni 204-MF
MelWMV KX059700
WMV-SSAV ALV83313
RhGRV QJT93260
HlGRV QJT93257
GALV NP 056791
GALV AAA46811
GALV SEATO ALV83304
GALV Halls Island ALV83310
GALV Brain ALV83307
GALV X AAC80265
GALV SF ALV83301
KoRV QMJ6480 AHY24813
KoRV Pci SN265 AHY24816
KoRV KV522 BAM67147
KoRV Br2-1 AGO86848
SaGRV QJT93254
MmGRV QJT93251
FFRV1 QDA02051
HPG QJT93248
McERV AGP25481
MDEV AAC31806
RfRV AFA52560
PERV-B AAM29193
PERV-A ASU50142
PERV-C ADK35879
FeLV AB807599
RD114 BAM17306
REV ASH96782
3.0


(H) Gblock-3 filters
M. leucogaster (300-ME)
M. leucogaster (290-MF)
M. leucogaster (249-MF)
M. leucogaster (291-MF)
M. leucogaster (292-MF)
M. burtoni (204-MF)
MelWMV KX059700
WMV-SSAV KT724051
MbRV KF572483
MbRV KF572484
GALV SF KT724047
GALV X U60065
GALV Halls Island KT724050
GALV Brain KT724049
GALV SEATO KT724048
GALV NC_001885
GALV M26927
RhGRV MN413614
HlGRV MN413613
KoRV Br2-1 KC779547
KoRV Pci-SN265 KF786285
KoRV KV522 AB721500
KoRV Pci-QMJ6480 KF786284
FFRV1 MK040728
HPG MN413610
SaGRV MN413612
MmGRV MN413611
McERV KC460271
MDEV AF053745
C. striatus ERV XM 010198675
M. davidii ERV XM 015571801
PERV-B HQ540594
PERV-C HM159246
PERV-A HQ540591
M. burtoni (201-MF) contig
C. griseus XM 027433137
M. coucha XM 031364589
MlRV JQ951956
J. jaculus ERV XM 012952149
R. norvegicus XM 008776280
R. verecundus (246-MF) contig
R. niobe sp. b (89-ME) contig
RD114 AB705392
FeLV AB807599
RfRV JQ303225
MrRV JQ292912
REV MF185397
3.0

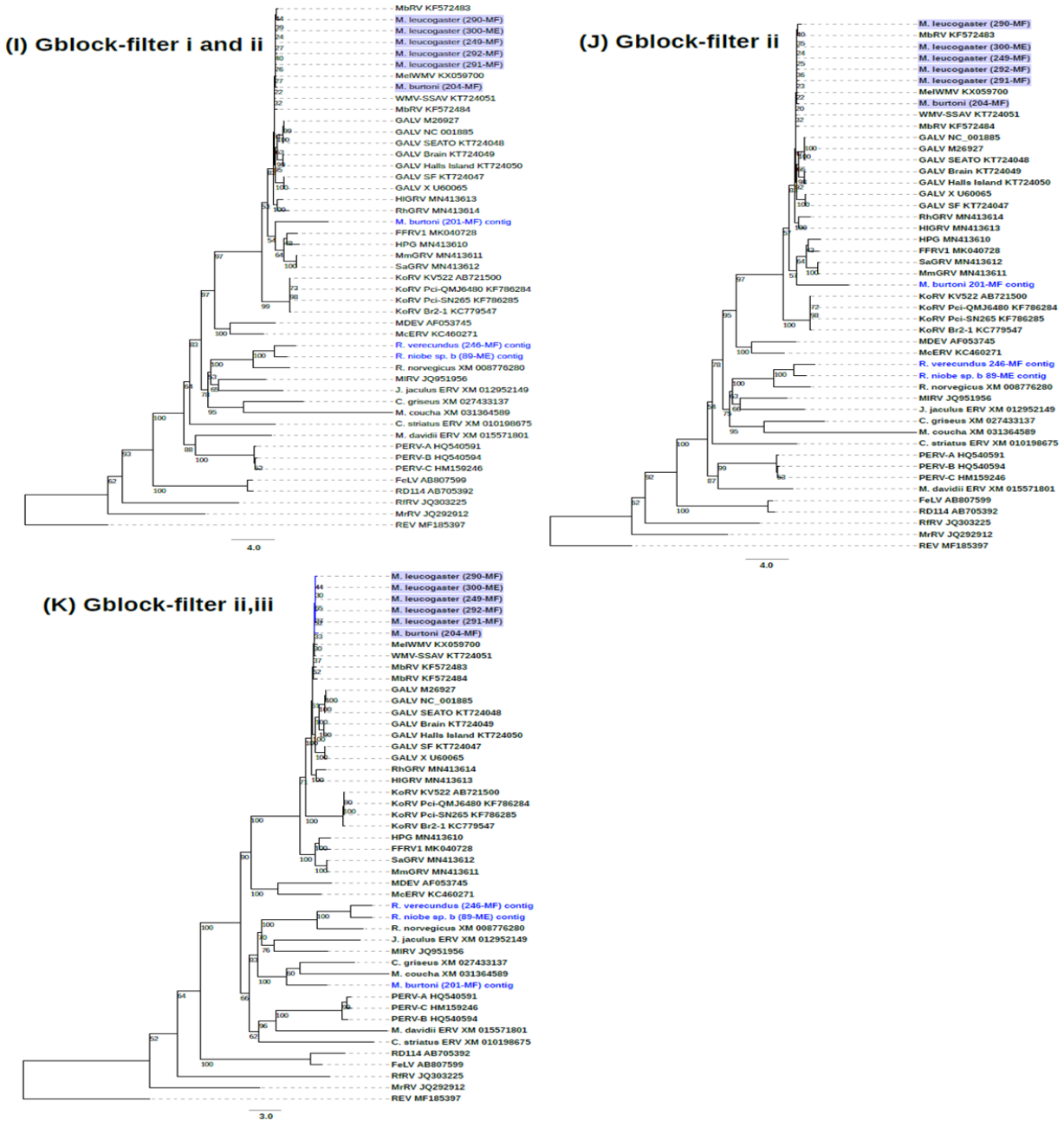
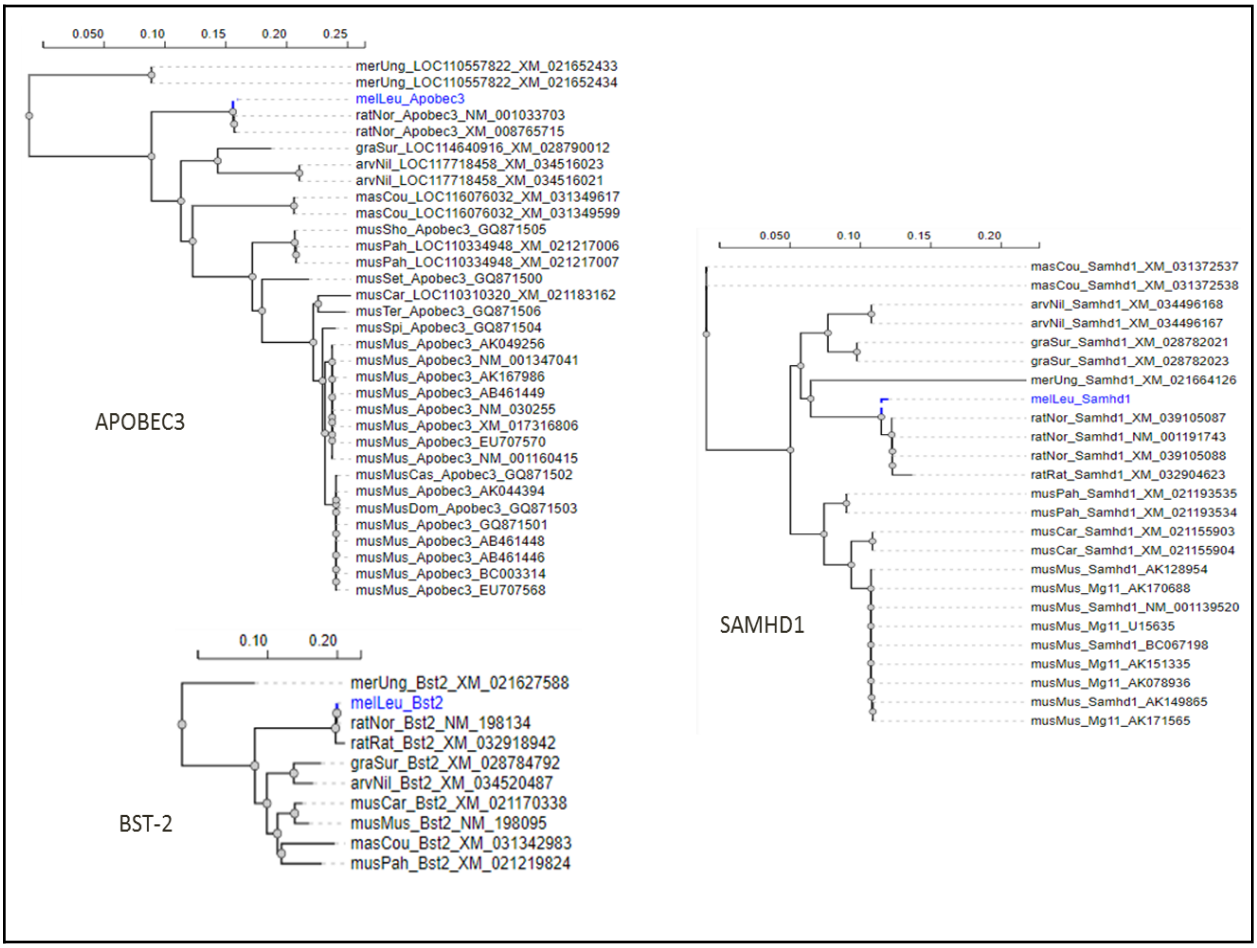


Figure S2.1 Maximum likelihood phylogenetic relationship of cMWMV inferred from gammaretroviruses for **(A)** complete nucleotide sequences with scaled branch length, **(B)** *gag*, **(C)** *pol*, **(D)** *env* genes and the proteins **(E-G)**. Combination of Gblock parameters were applied to the whole-genome alignments (refer to materials and methods) and the relationship of these viral sequences were constructed when three Gblock parameters **(H)**, first and second parameter **(I)**, second parameter **(J)**, second and third parameters **(K)** were applied. The viral trees were rooted using avian reticuloendotheliosis virus (REV) and bootstrap values are depicted on branches. The viral sequences identified in this study are marked in blue and the cMWMV clade is highlighted, showing the evolutionary relationship of these sequences remains largely consistent.


0.050
0.10
0.15
0.20
0.25
merUng_LOC110557822_XM_021652433
merUng_LOC110557822_XM_021652434
melLeu_Apobec3
ratNor_Apobec3_NM_001033703
ratNor_Apobec3_XM_008765715
graSur_LOC114640916_XM_028790012
arvNil_LOC117718458_XM_034516023
arvNil_LOC117718458_XM_034516021
masCou_LOC116076032_XM_031349617
masCou_LOC116076032_XM_031349599
musSho_Apobec3_GQ871505
musPah_LOC110334948_XM_021217006
musPah_LOC110334948_XM_021217007
musSet_Apobec3_GQ871500
musCar_LOC110310320_XM_021183162
musTer_Apobec3_GQ871506
musSpi_Apobec3_GQ871504
musMus_Apobec3_AK049256
musMus_Apobec3_NM_001347041
musMus_Apobec3_AK167986
musMus_Apobec3_AB461449
musMus_Apobec3_NM_030255
musMus_Apobec3_XM_017316806
musMus_Apobec3_EU707570
musMus_Apobec3_NM_001160415
musMusCas_Apobec3_GQ871502
musMus_Apobec3_AK044394
musMusDom_Apobec3_GQ871503
musMus_Apobec3_GQ871501
musMus_Apobec3_AB461448
musMus_Apobec3_AB461446
musMus_Apobec3_BC003314
musMus_Apobec3_EU707568
APOBEC3
0.10
0.20
merUng_Bst2_XM_021627588
melLeu_Bst2
ratNor_Bst2_NM_198134
ratRat_Bst2_XM_032918942
graSur_Bst2_XM_028784792
arvNil_Bst2_XM_034520487
musCar_Bst2_XM_021170338
musMus_Bst2_NM_198095
masCou_Bst2_XM_031342983
musPah_Bst2_XM_021219824
BST-2
0.050
0.10
0.15
0.20
masCou_Samhd1_XM_031372537
masCou_Samhd1_XM_031372538
arvNil_Samhd1_XM_034496168
arvNil_Samhd1_XM_034496167
graSur_Samhd1_XM_028782021
graSur_Samhd1_XM_028782023
merUng_Samhd1_XM_021664126
melLeu_Samhd1
ratNor_Samhd1_XM_039105087
ratNor_Samhd1_NM_001191743
ratNor_Samhd1_XM_039105088
ratRat_Samhd1_XM_032904623
musPah_Samhd1_XM_021193535
musPah_Samhd1_XM_021193534
musCar_Samhd1_XM_021155903
musCar_Samhd1_XM_021155904
musMus_Samhd1_AK128954
musMus_Mg11_AK170688
musMus_Samhd1_NM_001139520
musMus_Mg11_U15635
musMus_Samhd1_BC067198
musMus_Mg11_AK151335
musMus_Mg11_AK078936
musMus_Samhd1_AK149865
musMus_Mg11_AK171565
SAMHD1

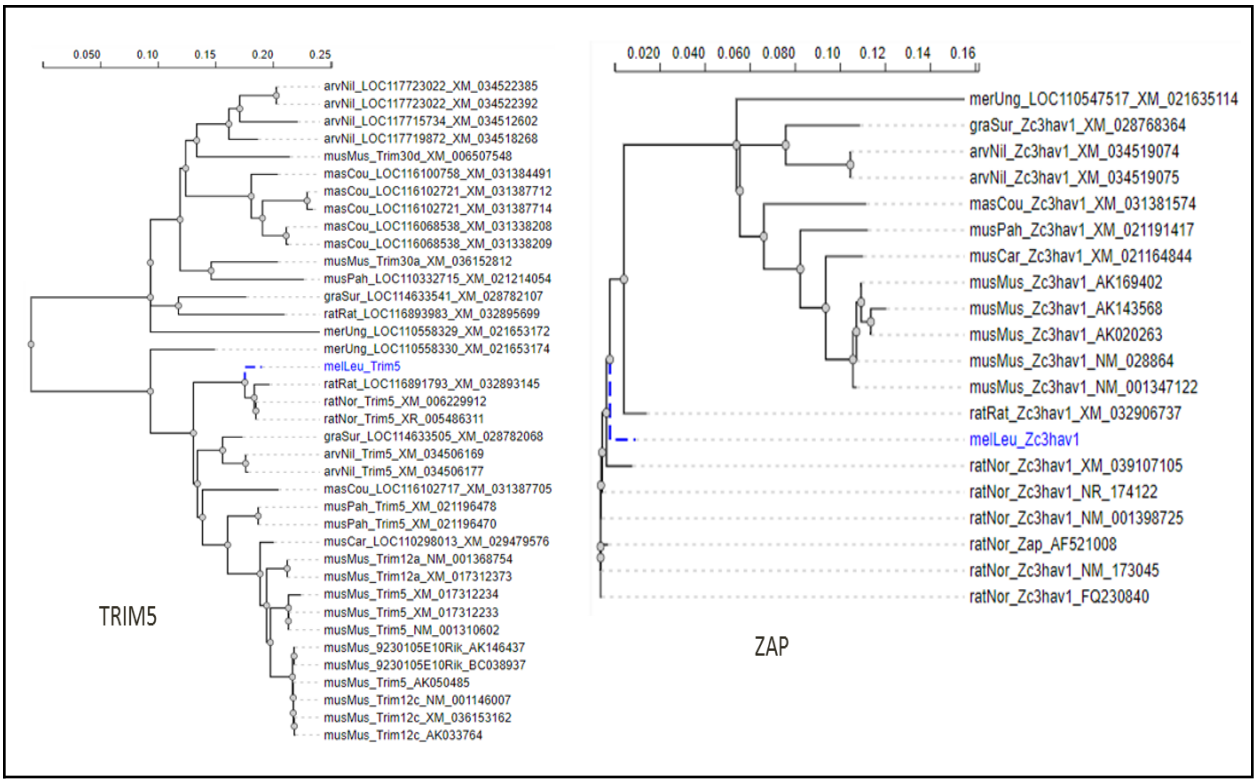
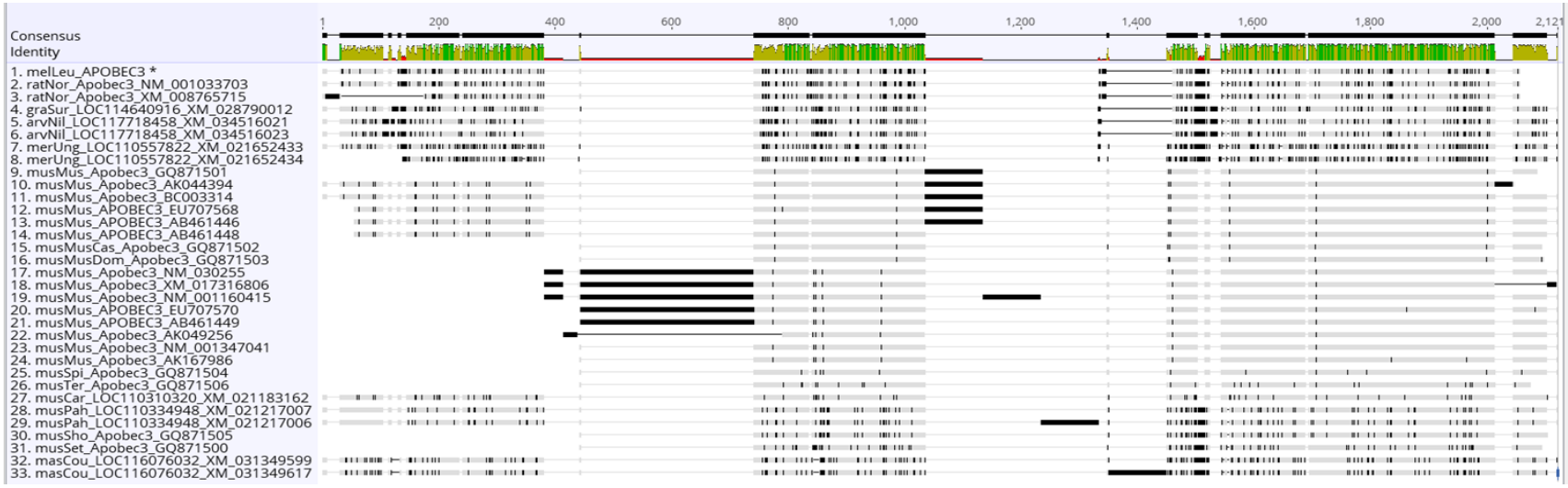
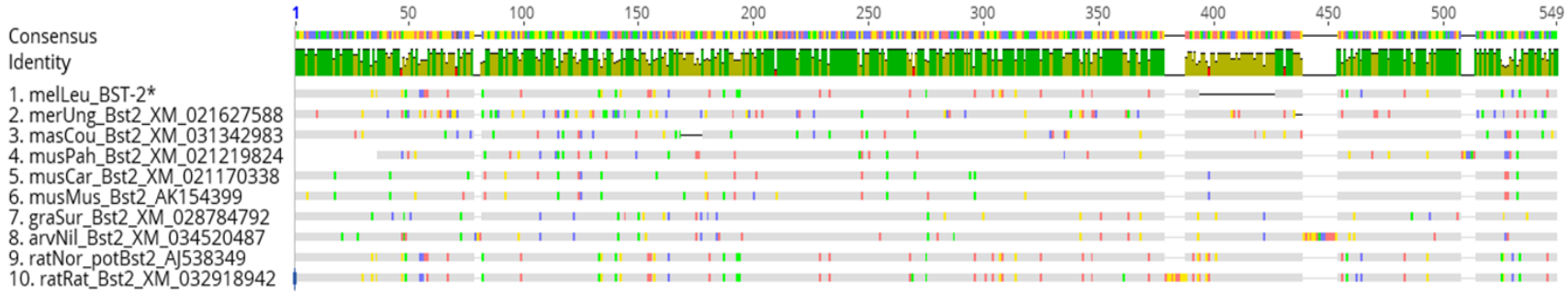


Figure S3.1 For positive selection analysis, it is required to compare groups of orthologous sequences. This figure shows reconciled gene trees that share a common evolutionary history in rodent species. These trees are based on retrieved homologous sequences from the DGINN pipeline and were inferred from the species tree (Text file S3.1) to form groups of orthologues based on ancestral duplication events. Because these trees are used to sort sequences, no bootstrap value is shown. Sequence IDs obeys DGINN parameter requirement of speSpe naming, followed by GenBank IDs. Phylogenetic trees were visualized using PRESTO (Phylogenetic tReE viSualisaTiOn) implemented in PhyML [194]. A scale bar for each gene tree is stated and positions of *Melomys* query sequences (melLeu) are displayed in blue.

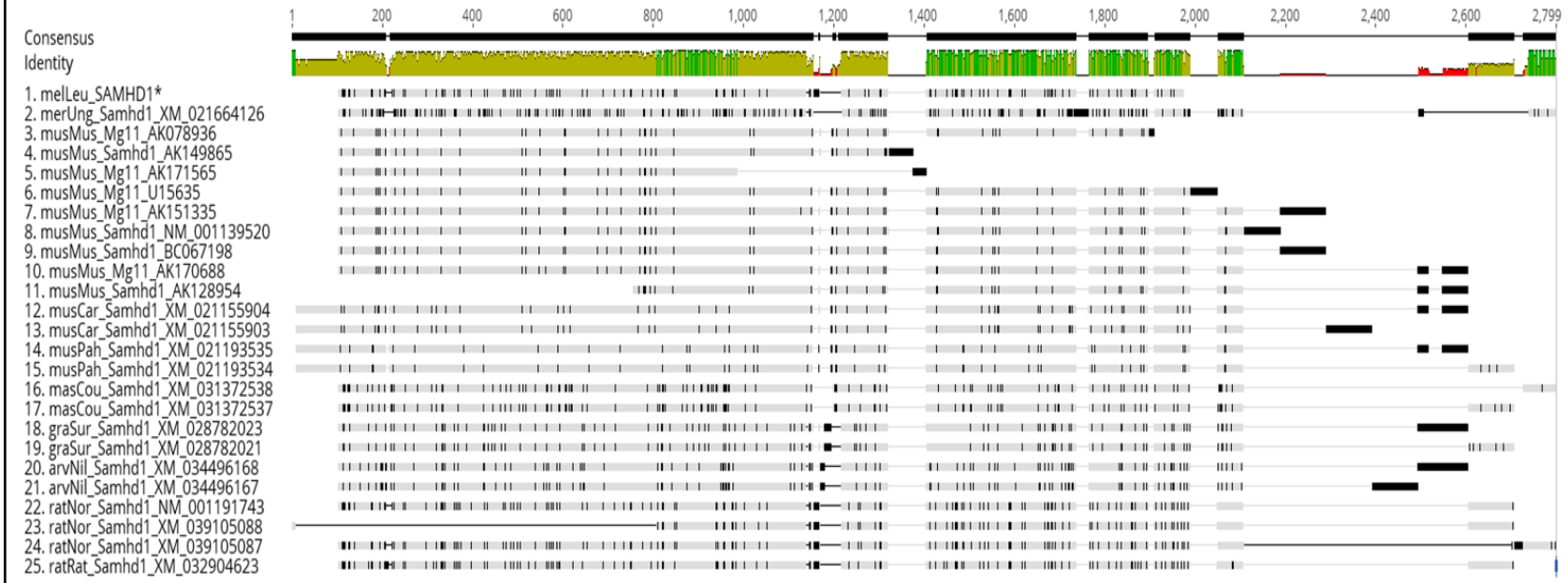
APOBEC3



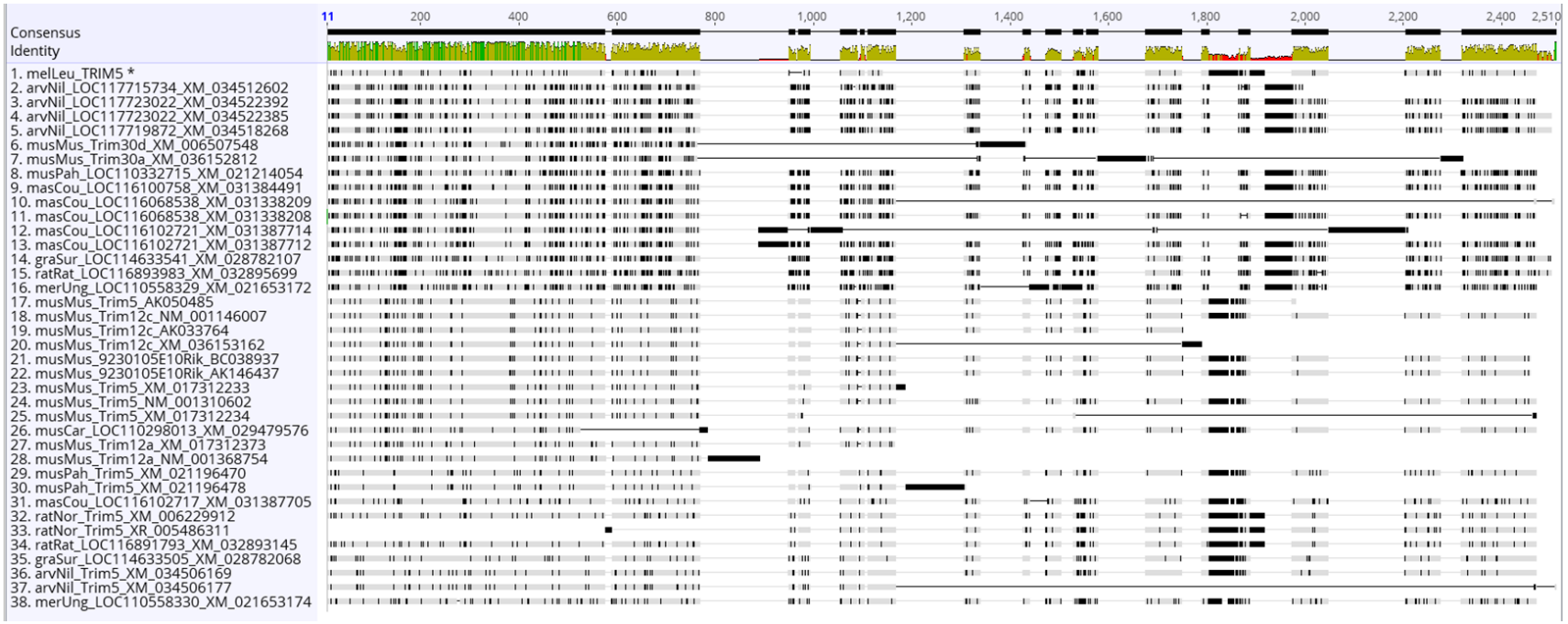
BST-2



SAMHD1


1
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
2,600
2,799
Consensus
Identity
1. melLeu_SAMHD1*
2. merUng_Samhd1_XM_021664126
3. musMus_Mg11_AK078936
4. musMus_Samhd1_AK149865
5. musMus_Mg11_AK171565
6. musMus_Mg11_U15635
7. musMus_Mg11_AK151335
8. musMus_Samhd1_NM_001139520
9. musMus_Samhd1_BC067198
10. musMus_Mg11_AK170688
11. musMus_Samhd1_AK128954
12. musCar_Samhd1_XM_021155904
13. musCar_Samhd1_XM_021155903
14. musPah_Samhd1_XM_021193535
15. musPah_Samhd1_XM_021193534
16. masCou_Samhd1_XM_031372538
17. masCou_Samhd1_XM_031372537
18. graSur_Samhd1_XM_028782023
19. graSur_Samhd1_XM_028782021
20. arvNil_Samhd1_XM_034496168
21. arvNil_Samhd1_XM_034496167
22. ratNor_Samhd1_NM_001191743
23. ratNor_Samhd1_XM_039105088
24. ratNor_Samhd1_XM_039105087
25. ratRat_Samhd1_XM_032904623

TRIM5


11
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
2,510
Consensus
Identity
1. melLeu_TRIM5 *
2. arvNil_LOC117715734_XM_034512602
3. arvNil_LOC117723022_XM_034522392
4. arvNil_LOC117723022_XM_034522385
5. arvNil_LOC117719872_XM_034518268
6. musMus_Trim30d_XM_006507548
7. musMus_Trim30a_XM_036152812
8. musPah_LOC110332715_XM_021214054
9. masCou_LOC116100758_XM_031384491
10. masCou_LOC116068538_XM_031338209
11. masCou_LOC116068538_XM_031338208
12. masCou_LOC116102721_XM_031387714
13. masCou_LOC116102721_XM_031387712
14. graSur_LOC114633541_XM_028782107
15. ratRat_LOC116893983_XM_032895699
16. merUng_LOC110558329_XM_021653172
17. musMus_Trim5_AK050485
18. musMus_Trim12c_NM_001146007
19. musMus_Trim12c_AK033764
20. musMus_Trim12c_XM_036153162
21. musMus_9230105E10Rik_BC038937
22. musMus_9230105E10Rik_AK146437
23. musMus_Trim5_XM_017312233
24. musMus_Trim5_NM_001310602
25. musMus_Trim5_XM_017312234
26. musCar_LOC110298013_XM_029479576
27. musMus_Trim12a_XM_017312373
28. musMus_Trim12a_NM_001368754
29. musPah_Trim5_XM_021196470
30. musPah_Trim5_XM_021196478
31. masCou_LOC116102717_XM_031387705
32. ratNor_Trim5_XM_006229912
33. ratNor_Trim5_XR_005486311
34. ratRat_LOC116891793_XM_032893145
35. graSur_LOC114633505_XM_028782068
36. arvNil_Trim5_XM_034506169
37. arvNil_Trim5_XM_034506177
38. merUng_LOC110558330_XM_021653174

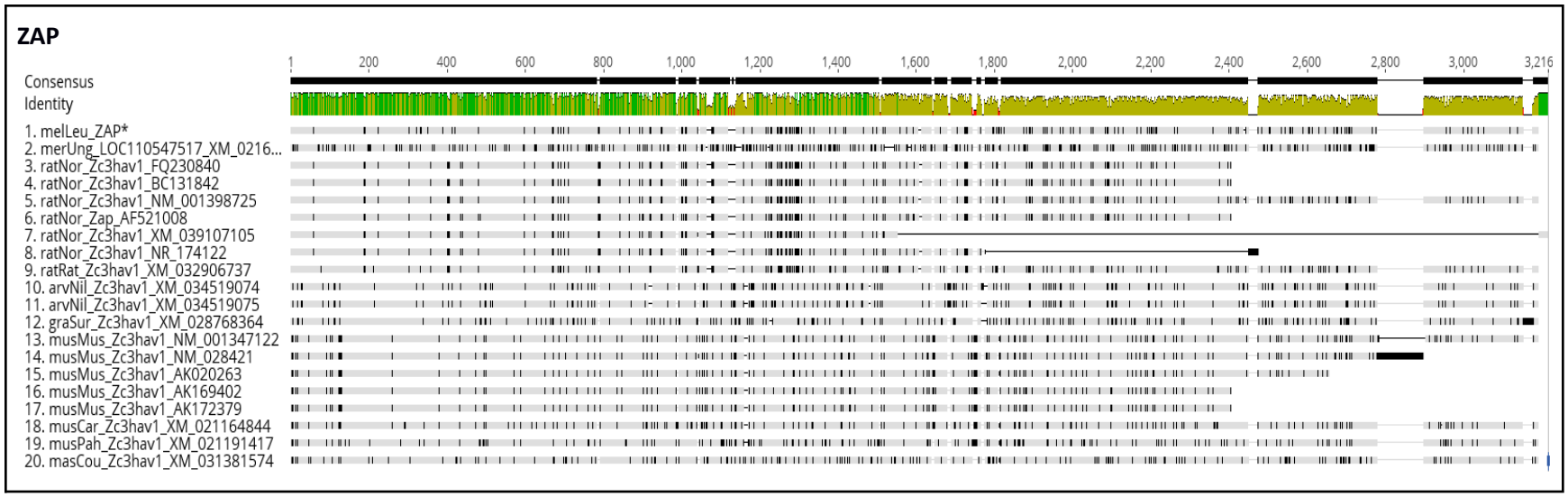


Figure S3.2 Final codon-wise alignment for each gene family is based on the reconciled gene trees (Figure S3.1) and re-alignments using PRANK [193]. These alignments were used further to provide the input for positive selection analysis. Alignments are visualized using Geneious R9.1, where character states not matching are indicated in black. Sequence IDs obeys DGINN parameter requirement of speSpe naming, followed by GenBank ID. *Melomys* sequences are displayed at number one with an asterix (melLeu*).

Table S2.1 The sample set used for PCR screening in this study. The approximate localities are shown in Figure 2.1. Samples that yielded an amplicon and were used for Illumina library preparations are marked *. Abbreviations for sample numbering are B= bat, M= muridae, F= frozen and E= ethanol preserved.

SAM ID		Sample no.		Order	Family	Genus	Species	Country	State/Province/District	Tissue	Collection year	Geographic units	Used for library prep
ABTC	44618	236	BF	Chiroptera	Emballonuridae	*Emballonura*	not described	Papua New Guinea	Southern Highlands Province	Kidney	1985	Australo-Papuan	
ABTC	44758	238	BF	Chiroptera	Emballonuridae	*Emballonura*	not described	Papua New Guinea	Southern Highlands Province	Liver	1987	Australo-Papuan	
ABTC	46964	242	BF	Chiroptera	Emballonuridae	*Emballonura*	not described	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	
ABTC	49362	267	BF	Chiroptera	Emballonuridae	*Emballonura*	not described	Papua New Guinea	Madang Province	Liver	1987	Australo-Papuan	
ABTC	49442	268	BF	Chiroptera	Emballonuridae	*Emballonura*	not described	Papua New Guinea	Manus Province	Not described	1987	Australo-Papuan	
ABTC	141839	123	BE	Chiroptera	Emballonuridae	*Mosia*	not described	Papua New Guinea	Gulf Province	Liver	2016	Australo-Papuan	
ABTC	5753	146	BF	Chiroptera	Emballonuridae	*Saccolaimus*	*flaviventris*	Australia	Northern Territory	Liver	1982	Australo-Papuan	
ABTC	128339	100	BE	Chiroptera	Emballonuridae	*Saccolaimus*	*saccolaimus*	Indonesia	Jasinga, Ranghas Bitung	Liver	1992	Asia	
ABTC	111441	23	BE	Chiroptera	Emballonuridae	*Taphozous*	*australis*	Australia	Queensland	Wing	2009	Australo-Papuan	
ABTC	111414	22	BE	Chiroptera	Emballonuridae	*Taphozous*	*australis/georgianus*	Australia	Queensland	Liver	2009	Australo-Papuan	
ABTC	5742	145	BF	Chiroptera	Emballonuridae	*Taphozous*	*hilli*	Australia	Northern Territory	Liver	1982	Australo-Papuan	
ABTC	122277	26	BE	Chiroptera	Hipposideridae	*Aselliscus*	*tricuspidatus*	Papua New Guinea	West New Britain Province	Liver	N/A	Australo-Papuan	
ABTC	51374	271	BF	Chiroptera	Hipposideridae	*Hipposideros*	*ater*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	141912	126	BE	Chiroptera	Hipposideridae	*Hipposideros*	*ater*	Papua New Guinea	Gulf Province	Liver	2016	Australo-Papuan	
ABTC	141626	120	BE	Chiroptera	Hipposideridae	*Hipposideros*	*calcaratus*	Papua New Guinea	Gulf Province	Wing	2016	Australo-Papuan	
ABTC	5809	148	BF	Chiroptera	Hipposideridae	*Hipposideros*	*cervinus*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	49491	269	BF	Chiroptera	Hipposideridae	*Hipposideros*	*cervinus*	Papua New Guinea	Sandaun Province	Liver	1987	Australo-Papuan	
ABTC	44262	232	BF	Chiroptera	Hipposideridae	*Hipposideros*	*corynophyllus*	Papua New Guinea	Sandaun Province	Liver	1985	Australo-Papuan	
ABTC	137181	115	BE	Chiroptera	Hipposideridae	*Hipposideros*	*diadema*	Papua New Guinea	Western Province	Liver	2014	Australo-Papuan	
ABTC	91986	101	BE	Chiroptera	Hipposideridae	*Hipposideros*	*inornatus*	Indonesia	West Papua	Liver	N/A	Australo-Papuan	
ABTC	42636	212	BF	Chiroptera	Hipposideridae	*Hipposideros*	*inornatus*	Papua New Guinea	Sandaun Province	Not described	1984	Australo-Papuan	
ABTC	117632	103	BE	Chiroptera	Hipposideridae	*Hipposideros*	*maggietaylorae*	Papua New Guinea	Sandaun Province	Liver	2011	Australo-Papuan	
ABTC	137233	119	BE	Chiroptera	Hipposideridae	*Hipposideros*	*muscinus*	Papua New Guinea	Western Province	Liver	2014	Australo-Papuan	
ABTC	22855	174	BF	Chiroptera	Hipposideridae	*Hipposideros*	*semoni*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	50967	270	BF	Chiroptera	Hipposideridae	*Hipposideros*	*stenotis*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	121253	25	BE	Chiroptera	Hipposideridae	*Hipposideros*	*wollastoni*	Papua New Guinea	Western Province	Liver	2009	Australo-Papuan	
ABTC	5706	143	BF	Chiroptera	Miniopteridae	*Miniopterus*	*australis*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	63365	273	BF	Chiroptera	Miniopteridae	*Miniopterus*	*australis*	Indonesia	Kai Besar	Liver	N/A	Wallacea	
ABTC	46437	241	BF	Chiroptera	Miniopteridae	*Miniopterus*	*australis*	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	
ABTC	44013	230	BF	Chiroptera	Miniopteridae	*Miniopterus*	*magnater*	Papua New Guinea	Sandaun Province	Liver	1986	Australo-Papuan	
ABTC	47147	252	BF	Chiroptera	Miniopteridae	*Miniopterus*	*medius*	Papua New Guinea	Sandaun Province	Kidney	1986	Australo-Papuan	
ABTC	5730	144	BF	Chiroptera	Miniopteridae	*Miniopterus*	*orianae*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	43604	2	BE	Chiroptera	Miniopteridae	*Miniopterus*	*schreibersii*	Papua New Guinea	National Capital District	Liver	1983	Australo-Papuan	
ABTC	129157	104	BE	Chiroptera	Miniopteridae	*Miniopterus*	*sp (48 khz)*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	128715	111	BE	Chiroptera	Miniopteridae	*Miniopterus*	*sp a*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	

SAM ID		Sample no.		Order	Family	Genus	Species	Country	State/Province/District	Tissue	Collection year	Geographic units	Used for library prep
ABTC	128716	112	BE	Chiroptera	Miniopteridae	*Miniopterus*	*sp b*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	137201	116	BE	Chiroptera	Miniopteridae	*Miniopterus*	*sp. d*	Papua New Guinea	Western Province	Liver	2014	Australo-Papuan	
ABTC	44511	235	BF	Chiroptera	Miniopteridae	*Miniopterus*	*tristis*	Papua New Guinea	Southern Highlands Province	Kidney	1987	Australo-Papuan	
ABTC	102003	20	BE	Chiroptera	Molossidae	*Chaerephon*	*jobensis*	Australia	Northern Territory	Liver	2002	Australo-Papuan	
ABTC	5568	139	BF	Chiroptera	Molossidae	*Chaerephon*	*jobensis*	Australia	Queensland	Liver	1983	Australo-Papuan	
ABTC	5451	136	BF	Chiroptera	Molossidae	*Mormopterus*	*eleryi*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	5382	131	BF	Chiroptera	Molossidae	*Mormopterus*	*lumsdenae*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	5402	133	BF	Chiroptera	Molossidae	*Mormopterus*	*lumsdenae*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	5598	141	BF	Chiroptera	Molossidae	*Mormopterus*	not described	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	5503	137	BF	Chiroptera	Molossidae	*Mormopterus*	*petersi*	Australia	Queensland	Fetus	N/A	Australo-Papuan	
ABTC	81341	4	BE	Chiroptera	Molossidae	*Mormopterus*	*ridei*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	99196	14	BE	Chiroptera	Pteropodidae	*Acerodon*	*celebensis*	Indonesia	North Sulawesi	Hair	2007	Wallacea	
ABTC	99206	16	BE	Chiroptera	Pteropodidae	*Acerodon*	*humilis*	Indonesia	North Sulawesi	Hair	2007	Wallacea	
ABTC	48023	259	BF	Chiroptera	Pteropodidae	*Aethalops*	*alecto*	Indonesia	West Java	Liver	N/A	Asia	
ABTC	42678	213	BF	Chiroptera	Pteropodidae	*Aproteles*	*bulmerae*	Papua New Guinea	Sandaun Province	Not described	1984	Australo-Papuan	
ABTC	48016	257	BF	Chiroptera	Pteropodidae	*Chironax*	*melanocephalus*	Indonesia	Cibodas forest, Java	Liver	N/A	Asia	
ABTC	48051	261	BF	Chiroptera	Pteropodidae	*Cynopterus*	*brachyotis*	Indonesia	West Java	Liver	N/A	Asia	
ABTC	47976	254	BF	Chiroptera	Pteropodidae	*Cynopterus*	*sphinx*	Indonesia	Yogyakarta	Liver	N/A	Asia	
ABTC	44298	233	BF	Chiroptera	Pteropodidae	*Dobsonia*	*anderseni*	Papua New Guinea	East New Britain Province	Blood	1985	Australo-Papuan	
ABTC	103262	21	BE	Chiroptera	Pteropodidae	*Dobsonia*	*magna*	Papua New Guinea	Western Province	Not described	2008	Australo-Papuan	
ABTC	92001	102	BE	Chiroptera	Pteropodidae	*Dobsonia*	*minor*	Indonesia	West Papua	Liver	N/A	Australo-Papuan	
ABTC	128723	113	BE	Chiroptera	Pteropodidae	*Dobsonia*	*minor*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	99274	18	BE	Chiroptera	Pteropodidae	*Dobsonia*	*peronii*	Indonesia	Sumba	Hair	2007	Wallacea	
ABTC	44305	234	BF	Chiroptera	Pteropodidae	*Dobsonia*	*praedatrix*	Papua New Guinea	East New Britain Province	Blood	1985	Australo-Papuan	
ABTC	119115	24	BE	Chiroptera	Pteropodidae	*Eonycteris*	not described	Lao PDR		Liver	N/A	Asia	
ABTC	137138	91	BE	Chiroptera	Pteropodidae	*Macroglossus*	*minimus*	Indonesia	South Sulawesi	Liver	1990	Wallacea	
ABTC	91922	12	BE	Chiroptera	Pteropodidae	*Macroglossus*	*minimus*	Papua New Guinea	Western Province	Not described	2005	Australo-Papuan	
ABTC	137143	92	BE	Chiroptera	Pteropodidae	*Macroglossus*	*sobrinus*	Indonesia	Sinabag, Mentawai	Liver	1991	Asia	
ABTC	130318	108	BE	Chiroptera	Pteropodidae	*Nyctimene*	*aello*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	126484	169	BF	Chiroptera	Pteropodidae	*Nyctimene*	*albiventer*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	49024	253	BF	Chiroptera	Pteropodidae	*Nyctimene*	*cyclotis*	Papua New Guinea	Morobe Province	Kidney	1987	Australo-Papuan	
ABTC	147325	277	BE	Chiroptera	Pteropodidae	*Nyctimene*	*draconilla*	Papua New Guinea	Gulf Province	Wing	2016	Australo-Papuan	
ABTC	130291	106	BE	Chiroptera	Pteropodidae	*Paranyctimene*	*raptor*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	130313	107	BE	Chiroptera	Pteropodidae	*Paranyctimene*	*tenax*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	91915	11	BE	Chiroptera	Pteropodidae	*Pteropus*	*alecto*	Papua New Guinea	Western Province	Hair	2005	Australo-Papuan	

SAM ID		Sample no.		Order	Family	Genus	Species	Country	State/Province/District	Tissue	Collection year	Geographic units	Used for library prep
ABTC	83062	6	BE	Chiroptera	Pteropodidae	*Pteropus*	*conspicillatus*	Australia	Queensland	Hair	2005	Australo-Papuan	
ABTC	91927	13	BE	Chiroptera	Pteropodidae	*Pteropus*	*griseus*	Timor Leste		Not described	2005	Wallacea	
ABTC	91914	10	BE	Chiroptera	Pteropodidae	*Pteropus*	*macrotis*	Papua New Guinea	Western Province	Hair	2005	Australo-Papuan	
ABTC	46874	9	BF	Chiroptera	Pteropodidae	*Pteropus*	*neohibernicus*	Papua New Guinea	Southern Highlands Province	Liver	1987	Australo-Papuan	
ABTC	83209	7	BE	Chiroptera	Pteropodidae	*Pteropus*	*vampyrus*	Timor Leste		Hair	2004	Wallacea	
ABTC	99269	17	BE	Chiroptera	Pteropodidae	*Rousettus*	*amplexicaudatus*	Indonesia	Sumba	Hair	2007	Wallacea	
ABTC	147340	278	BE	Chiroptera	Pteropodidae	*Rousettus*	*amplexicaudatus*	Papua New Guinea	Gulf Province	Liver	2016	Australo-Papuan	
ABTC	83217	8	BE	Chiroptera	Pteropodidae	*Rousettus*	*amplexicaudatus*	Timor Leste		Hair	2004	Wallacea	
ABTC	48001	256	BF	Chiroptera	Pteropodidae	*Rousettus*	*celebensis*	Indonesia	North Sulawesi	Serum	N/A	Wallacea	
ABTC	47979	255	BF	Chiroptera	Pteropodidae	*Rousettus*	*leschenaultii*	Indonesia	Yogyakarta	Liver	N/A	Asia	
ABTC	137148	93	BE	Chiroptera	Pteropodidae	*Syconycteris*	*australis*	Indonesia	Ambon	Liver	1992	Wallacea	
ABTC	133746	30	BE	Chiroptera	Pteropodidae	*Syconycteris*	*hobbit*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	133603	28	BE	Chiroptera	Pteropodidae	*Syconycteris*	*sp lowland*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	133749	31	BE	Chiroptera	Pteropodidae	*Syconycteris*	*sp montane*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	129187	105	BE	Chiroptera	Rhinolophidae	*Rhinolophus*	*arcuatus*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	45149	239	BF	Chiroptera	Rhinolophidae	*Rhinolophus*	*euryotis*	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	
ABTC	5835	149	BF	Chiroptera	Rhinolophidae	*Rhinolophus*	*intermediate*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	5703	142	BF	Chiroptera	Rhinolophidae	*Rhinolophus*	*megaphyllus*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	141263	114	BE	Chiroptera	Rhinolophidae	*Rhinolophus*	*megaphyllus*	Papua New Guinea	Southern Highlands Province	Liver	N/A	Australo-Papuan	
ABTC	44757	237	BF	Chiroptera	Rhinolophidae	*Rhinolophus*	*philippinensis*	Papua New Guinea	Southern Highlands Province	Liver	1987	Australo-Papuan	
ABTC	130462	109	BE	Chiroptera	Rhinolophidae	*Rhinolophus*	*sp. a*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	130491	110	BE	Chiroptera	Rhinolophidae	*Rhinolophus*	*sp. b*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	104412	94	BE	Chiroptera	Vespertilionidae	*Arielulus*	*circumdatus*	Indonesia	Kebun Raya, Candikuning	Liver	1991	Asia	
ABTC	24593	175	BF	Chiroptera	Vespertilionidae	*Chalinolobus*	*dwyeri*	Australia	New South Wales	Liver	N/A	Australo-Papuan	
ABTC	79278	178	BF	Chiroptera	Vespertilionidae	*Chalinolobus*	*gouldii*	Australia	Northern Territory	Liver	2000	Australo-Papuan	
ABTC	5907	150	BE	Chiroptera	Vespertilionidae	*Chalinolobus*	*morio*	Australia	Queensland	Kidney	N/A	Australo-Papuan	
ABTC	5777	147	BE	Chiroptera	Vespertilionidae	*Chalinolobus*	*nigrogriseus*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	43583	217	BF	Chiroptera	Vespertilionidae	*Chalinolobus*	*nigrogriseus*	Papua New Guinea	National Capital District	Liver	1983	Australo-Papuan	
ABTC	100473	19	BE	Chiroptera	Vespertilionidae	*Chalinolobus*	*picatus*	Australia	Queensland	Liver	2005	Australo-Papuan	
ABTC	104413	99	BE	Chiroptera	Vespertilionidae	*Hypsugo*	*imbricatus*	Indonesia	Batu Koq	Liver	1987	Wallacea	
ABTC	104432	96	BE	Chiroptera	Vespertilionidae	*Hypsugo*	*macrotis*	Indonesia	Candidasa	Liver	1991	Asia	
ABTC	147239	276	BE	Chiroptera	Vespertilionidae	*Kerivoula*	*muscina*	Papua New Guinea	Gulf Province	Liver	2016	Australo-Papuan	
ABTC	136735	32	BE	Chiroptera	Vespertilionidae	*Kerivoula*	*myrella*	Papua New Guinea	Manus Province	Liver	2014	Australo-Papuan	
ABTC	43588	218	BF	Chiroptera	Vespertilionidae	*Kerivoula*	not described	Papua New Guinea	National Capital District	Liver	1983	Australo-Papuan	
ABTC	137202	117	BE	Chiroptera	Vespertilionidae	*Kerivoula*	*sp a*	Papua New Guinea	Western Province	Liver	2014	Australo-Papuan	

SAM ID		Sample no.		Order	Family	Genus	Species	Country	State/Province/District	Tissue	Collection year	Geographic units	Used for library prep
ABTC	137203	118	BE	Chiroptera	Vespertilionidae	*Kerivoula*	*sp b*	Papua New Guinea	Western Province	Liver	2014	Australo-Papuan	
ABTC	6084	158	BF	Chiroptera	Vespertilionidae	*Murina*	*florium*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	141792	122	BE	Chiroptera	Vespertilionidae	*Murina*	not described	Papua New Guinea	Gulf Province	Liver	2016	Australo-Papuan	
ABTC	48019	258	BF	Chiroptera	Vespertilionidae	*Murina*	*suilla*	Indonesia	West Java	Liver	N/A	Asia	
ABTC	43390	214	BF	Chiroptera	Vespertilionidae	*Myotis*	*adversus*	Papua New Guinea	Chimbu Province	Not described	1983	Australo-Papuan	
ABTC	6071	153	BF	Chiroptera	Vespertilionidae	*Myotis*	*macropus*	Australia	Queensland	Liver	1982	Australo-Papuan	
ABTC	81364	5	BE	Chiroptera	Vespertilionidae	*Myotis*	*moluccarum*	Australia	Queensland	Not described	N/A	Australo-Papuan	
ABTC	63124	272	BF	Chiroptera	Vespertilionidae	*Myotis*	not described	Indonesia	Alor	Liver	N/A	Wallacea	
ABTC	43619	222	BF	Chiroptera	Vespertilionidae	*Myotis*	not described	Papua New Guinea	National Capital District	Kidney	1983	Australo-Papuan	
ABTC	6239	160	BF	Chiroptera	Vespertilionidae	*Nyctophilus*	*arnhemensis*	Australia	Northern Territory	Heart	N/A	Australo-Papuan	
ABTC	6269	163	BF	Chiroptera	Vespertilionidae	*Nyctophilus*	*bifax*	Australia	Queensland	Heart	N/A	Australo-Papuan	
ABTC	6241	161	BF	Chiroptera	Vespertilionidae	*Nyctophilus*	*daedalus*	Australia	Northern Territory	Heart	N/A	Australo-Papuan	
ABTC	6387	164	BF	Chiroptera	Vespertilionidae	*Nyctophilus*	*geoffroyi*	Australia	Northern Territory	Kidney	N/A	Australo-Papuan	
ABTC	6396	165	BF	Chiroptera	Vespertilionidae	*Nyctophilus*	*gouldi*	Australia	Queensland	Kidney	N/A	Australo-Papuan	
ABTC	43566	215	BF	Chiroptera	Vespertilionidae	*Nyctophilus*	*microdon*	Papua New Guinea	Chimbu Province	Not described	1983	Australo-Papuan	
ABTC	43655	224	BF	Chiroptera	Vespertilionidae	*Nyctophilus*	*microtis*	Papua New Guinea	National Capital District	Kidney	1983	Australo-Papuan	
ABTC	6243	162	BF	Chiroptera	Vespertilionidae	*Nyctophilus*	*walkeri*	Australia	Northern Territory	Heart	N/A	Australo-Papuan	
ABTC	46149	240	BF	Chiroptera	Vespertilionidae	*Philetor*	*brachypterus*	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	
ABTC	43629	223	BF	Chiroptera	Vespertilionidae	*Phoniscus*	not described	Papua New Guinea	National Capital District	Kidney	1983	Australo-Papuan	
ABTC	6083	157	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*adamsi*	Australia	Queensland	Heart	N/A	Australo-Papuan	
ABTC	6013	293	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*adamsi*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	44108	231	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*angulatus*	Papua New Guinea	Sandaun Province	Liver	1985	Australo-Papuan	
ABTC	44109	294	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*angulatus*	Papua New Guinea	Sandaun Province	Liver	1985	Australo-Papuan	
ABTC	133658	29	BE	Chiroptera	Vespertilionidae	*Pipistrellus*	*collinus*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	104423	95	BE	Chiroptera	Vespertilionidae	*Pipistrellus*	*javanicus*	Indonesia	Batudulang	Liver	1988	Wallacea	
ABTC	6020	152	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*murrayi*	Australia	Western Australia	Liver	N/A	Australo-Papuan	
ABTC	43596	220	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*papuanus*	Papua New Guinea	National Capital District	Kidney	1983	Australo-Papuan	
ABTC	132515	27	BE	Chiroptera	Vespertilionidae	*Pipistrellus*	*sp cf. papuanus*	Papua New Guinea	Western Highlands Province	Liver	2013	Australo-Papuan	
ABTC	43593	219	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*sp dark*	Papua New Guinea	National Capital District	Kidney	1983	Australo-Papuan	
ABTC	43617	221	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*sp orange*	Papua New Guinea	National Capital District	Kidney	1983	Australo-Papuan	
ABTC	141888	124	BE	Chiroptera	Vespertilionidae	*Pipistrellus*	*sp_large*	Papua New Guinea	Gulf Province	Liver	2016	Australo-Papuan	
ABTC	141889	125	BE	Chiroptera	Vespertilionidae	*Pipistrellus*	*sp_small*	Papua New Guinea	Gulf Province	Liver	2016	Australo-Papuan	
ABTC	141752	121	BE	Chiroptera	Vespertilionidae	*Pipistrellus*	*sp. 2*	Papua New Guinea	Gulf Province	Wing	2016	Australo-Papuan	
ABTC	104440	97	BE	Chiroptera	Vespertilionidae	*Pipistrellus*	*tenuis*	Indonesia	Batudulang	Liver	1988	Wallacea	
ABTC	72558	176	BF	Chiroptera	Vespertilionidae	*Pipistrellus*	*westralis*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	

SAM ID		Sample no.		Order	Family	Genus	Species	Country	State/Province/District	Tissue	Collection year	Geographic units	Used for library prep
ABTC	5513	138	BF	Chiroptera	Vespertilionidae	*Scoteanax*	*rueppellii*	Australia	New South Wales	Liver	N/A	Australo-Papuan	
ABTC	48038	260	BF	Chiroptera	Vespertilionidae	*Scotophilus*	*kuhlii*	Indonesia	Bogor Botanic Gardens Java	Liver	N/A	Asia	
ABTC	5389	132	BF	Chiroptera	Vespertilionidae	*Scotorepens*	*balstoni*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	5424	135	BF	Chiroptera	Vespertilionidae	*Scotorepens*	*greyii*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	5595	140	BF	Chiroptera	Vespertilionidae	*Scotorepens*	*orion*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	5413	134	BF	Chiroptera	Vespertilionidae	*Scotorepens*	*sanborni*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	43574	216	BF	Chiroptera	Vespertilionidae	*Scotorepens*	*sanborni*	Papua New Guinea	National Capital District	Kidney	1983	Australo-Papuan	
ABTC	62457	274	BF	Chiroptera	Vespertilionidae	*Scotorepens*	*sanborni*	Timor Leste		Liver	N/A	Wallacea	
ABTC	6046	155	BF	Chiroptera	Vespertilionidae	*Vespadelus*	*baverstocki*	Australia	Queensland	Heart	N/A	Australo-Papuan	
ABTC	29447	177	BF	Chiroptera	Vespertilionidae	*Vespadelus*	*caurinus*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	6053	156	BF	Chiroptera	Vespertilionidae	*Vespadelus*	*darlingtoni*	Australia	New South Wales	Heart	N/A	Australo-Papuan	
ABTC	6074	154	BF	Chiroptera	Vespertilionidae	*Vespadelus*	*finlaysoni*	Australia	Queensland	Heart	N/A	Australo-Papuan	
ABTC	24234	203	BF	Chiroptera	Vespertilionidae	*Vespadelus*	*pumilus*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	22846	173	BF	Chiroptera	Vespertilionidae	*Vespadelus*	*sp nov*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	81323	3	BE	Chiroptera	Vespertilionidae	*Vespadelus*	*troughtoni*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	13523	167	BF	Chiroptera	Vespertilionidae	*Vespadelus*	*vulturnus*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	121660	60	ME	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Australia	Western Australia	Muscle	N/A	Australo-Papuan	
ABTC	5312	181	MF	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Australia	South Australia	Liver	1986	Australo-Papuan	
ABTC	7432	182	MF	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	41165	207	MF	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	130495	86	ME	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	42569	210	MF	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Papua New Guinea	Sandaun Province	Liver	1984	Australo-Papuan	
ABTC	44494	243	MF	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Papua New Guinea	Southern Highlands Province	Not described	1987	Australo-Papuan	
ABTC	48908	265	MF	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Papua New Guinea	Madang Province	Liver	1987	Australo-Papuan	
ABTC	48932	266	MF	Rodentia	Muridae	*Hydromys*	*chrysogaster*	Papua New Guinea	Manus Province	Liver	1987	Australo-Papuan	
ABTC	29485	180	MF	Rodentia	Muridae	*Melomys*	*burtoni*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	8231	184	MF	Rodentia	Muridae	*Melomys*	*burtoni*	Australia	Western Australia	Muscle	N/A	Australo-Papuan	
ABTC	24239	201	MF	Rodentia	Muridae	*Melomys*	*burtoni*	Australia	Queensland	Liver	N/A	Australo-Papuan	*
ABTC	24240	204	MF	Rodentia	Muridae	*Melomys*	*burtoni*	Australia	Queensland	Liver	N/A	Australo-Papuan	*
ABTC	8233	298	MF	Rodentia	Muridae	*Melomys*	*burtoni*	Australia	Western Australia	Liver	N/A	Australo-Papuan	
ABTC	8363	C3288	ME	Rodentia	Muridae	*Melomys*	*burtoni*	Australia	Queensland	Hair	N/A	Australo-Papuan	
ABTC	116408	Cat.15	MF	Rodentia	Muridae	*Melomys*	*burtoni*	Australia	Queensland	Liver	1977	Australo-Papuan	
ABTC	8197	183	MF	Rodentia	Muridae	*Melomys*	*capensis*	Australia	Queensland	Liver	1983	Australo-Papuan	
ABTC	8336	185	MF	Rodentia	Muridae	*Melomys*	*cervinipes*	Australia	New South Wales	Not described	N/A	Australo-Papuan	
ABTC	24245	202	MF	Rodentia	Muridae	*Melomys*	*cervinipes*	Australia	Queensland	Liver	N/A	Australo-Papuan	

SAM ID		Sample no.		Order	Family	Genus	Species	Country	State/Province/District	Tissue	Collection year	Geographic units	Used for library prep
ABTC	130427	84	ME	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	137206	88	ME	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Western Province	Liver	2014	Australo-Papuan	
ABTC	8398	186	MF	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Sandaun Province	Liver	N/A	Australo-Papuan	
ABTC	42761	225	MF	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Chimbu Province	Blood	1983	Australo-Papuan	
ABTC	45752	249	MF	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	*
ABTC	45799	290	MF	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Southern Highlands Province	Kidney	1985	Australo-Papuan	*
ABTC	46056	291	MF	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Southern Highlands Province	Kidney	1985	Australo-Papuan	*
ABTC	44487	292	MF	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Southern Highlands Province	Liver	1987	Australo-Papuan	*
ABTC	130456	299	ME	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	137299	300	ME	Rodentia	Muridae	*Melomys*	*leucogaster*	Papua New Guinea	Western Province	Liver	2014	Australo-Papuan	*
ABTC	8393	187	MF	Rodentia	Muridae	*Melomys*	*lutillus*	Papua New Guinea	Western Province	Liver	N/A	Australo-Papuan	
ABTC	8394	188	MF	Rodentia	Muridae	*Melomys*	*lutillus*	Papua New Guinea	National Capital District	Liver	1983	Australo-Papuan	
ABTC	45120	248	MF	Rodentia	Muridae	*Melomys*	*lutillus*	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	
ABTC	48893	263	MF	Rodentia	Muridae	*Melomys*	*lutillus*	Papua New Guinea	Manus Province	Liver	1987	Australo-Papuan	
ABTC	8415	189	MF	Rodentia	Muridae	*Melomys*	*rubicola*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	91955	76	ME	Rodentia	Muridae	*Melomys*	*rufescens*	Indonesia	West Papua	Liver	N/A	Australo-Papuan	
ABTC	117667	79	ME	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Sandaun Province	Liver	2011	Australo-Papuan	
ABTC	121349	80	ME	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Western Province	Not described	N/A	Australo-Papuan	
ABTC	129133	81	ME	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	42505	208	MF	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Central Province	Not described	1981-2	Australo-Papuan	
ABTC	43059	227	MF	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Chimbu Province	Liver	1983	Australo-Papuan	
ABTC	44587	245	MF	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Southern Highlands Province	Kidney	1987	Australo-Papuan	
ABTC	48907	264	MF	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Manus Province	Not described	1987	Australo-Papuan	
ABTC	147240	281	ME	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Gulf Province	Not described	2016	Australo-Papuan	
ABTC	42615	283	MF	Rodentia	Muridae	*Melomys*	*rufescens*	Papua New Guinea	Sandaun Province	Liver	1984	Australo-Papuan	
ABTC	137205	87	ME	Rodentia	Muridae	*Melomys*	*sp (rufescens a)*	Papua New Guinea	Western Province	Liver	2014	Australo-Papuan	
ABTC	133597	71	ME	Rodentia	Muridae	*Melomys*	*sp cf niobe*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	132480	70	ME	Rodentia	Muridae	*Melomys*	*sp cf rufescens*	Papua New Guinea	Western Highlands Province	Liver	2013	Australo-Papuan	
ABTC	110210	45	ME	Rodentia	Muridae	*Rattus*	*argentiventer*	Indonesia	Gunung Mutis, Timor	Liver	N/A	Wallacea	
ABTC	110227	46	ME	Rodentia	Muridae	*Rattus*	*argentiventer*	Indonesia	Kalabahi, Alor	Liver	N/A	Wallacea	
ABTC	110231	47	ME	Rodentia	Muridae	*Rattus*	*argentiventer*	Indonesia	Apui, Alor	Liver	N/A	Wallacea	
ABTC	110235	48	ME	Rodentia	Muridae	*Rattus*	*argentiventer*	Indonesia	Ubud, Bali	Liver	N/A	Asia	
ABTC	8572	192	MF	Rodentia	Muridae	*Rattus*	*colletti*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	30680	282	MF	Rodentia	Muridae	*Rattus*	*colletti*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	123588	63	ME	Rodentia	Muridae	*Rattus*	*exulans*	Australia	Western Australia	Liver	2012	Australo-Papuan	

SAM ID		Sample no.		Order	Family	Genus	Species	Country	State/Province/District	Tissue	Collection year	Geographic units	Used for library prep
ABTC	115123	50	ME	Rodentia	Muridae	*Rattus*	*exulans*	Bangladesh	Feni	Liver	N/A	Asia	
ABTC	140332	58	ME	Rodentia	Muridae	*Rattus*	*exulans*	Indonesia	West Sulawesi	Liver	2012	Wallacea	
ABTC	48012	262	MF	Rodentia	Muridae	*Rattus*	*exulans*	Indonesia	West Java	Not described	N/A	Asia	
ABTC	125159	66	ME	Rodentia	Muridae	*Rattus*	*exulans*	Myanmar	Bago Division	Liver	N/A	Asia	
ABTC	130449	85	ME	Rodentia	Muridae	*Rattus*	*exulans*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	42509	209	MF	Rodentia	Muridae	*Rattus*	*exulans*	Papua New Guinea	Central Province	Not described	1981-2	Australo-Papuan	
ABTC	43078	228	MF	Rodentia	Muridae	*Rattus*	*exulans*	Papua New Guinea	Chimbu Province	Liver	1983	Australo-Papuan	
ABTC	142254	74	ME	Rodentia	Muridae	*Rattus*	*exulans*	Solomon Islands		Liver	N/A	Australo-Papuan	
ABTC	119382	56	ME	Rodentia	Muridae	*Rattus*	*exulans*	Thailand	Kanchanaburi	Liver	2002	Asia	
ABTC	100157	34	ME	Rodentia	Muridae	*Rattus*	*fuscipes*	Australia	Queensland	Liver	2004	Australo-Papuan	
ABTC	117078	52	ME	Rodentia	Muridae	*Rattus*	*fuscipes*	Australia	New South Wales	Liver	2011	Australo-Papuan	
ABTC	107436	196	MF	Rodentia	Muridae	*Rattus*	*fuscipes*	Australia	Queensland	Kidney	N/A	Australo-Papuan	
ABTC	107441	197	MF	Rodentia	Muridae	*Rattus*	*fuscipes*	Australia	Queensland	Kidney	N/A	Australo-Papuan	
ABTC	51715	279	MF	Rodentia	Muridae	*Rattus*	*fuscipes*	Australia	New South Wales	Liver	N/A	Australo-Papuan	
ABTC	87301	33	ME	Rodentia	Muridae	*Rattus*	*giluwensis*	Papua New Guinea	Enga Province	Liver	2005	Australo-Papuan	
ABTC	121690	61	ME	Rodentia	Muridae	*Rattus*	*hainaldi*	Indonesia	Ruteng, Flores	Liver	N/A	Wallacea	
ABTC	124323	64	ME	Rodentia	Muridae	*Rattus*	*hoffmanni*	Indonesia	Sulawesi	Liver	N/A	Wallacea	
ABTC	8447	190	MF	Rodentia	Muridae	*Rattus*	*leucopus*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	8493	191	MF	Rodentia	Muridae	*Rattus*	*leucopus*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	130331	83	ME	Rodentia	Muridae	*Rattus*	*leucopus*	Papua New Guinea	Gulf Province	Liver	2012	Australo-Papuan	
ABTC	42762	226	MF	Rodentia	Muridae	*Rattus*	*leucopus*	Papua New Guinea	Chimbu Province	Liver	1983	Australo-Papuan	
ABTC	44508	244	MF	Rodentia	Muridae	*Rattus*	*leucopus*	Papua New Guinea	Southern Highlands Province	Kidney	1987	Australo-Papuan	
ABTC	45966	250	MF	Rodentia	Muridae	*Rattus*	*leucopus*	Papua New Guinea	Oro Province	Liver	1985	Australo-Papuan	
ABTC	109040	38	ME	Rodentia	Muridae	*Rattus*	*losea*	Cambodia	Kampong Cham	Liver	2004	Asia	
ABTC	108956	37	ME	Rodentia	Muridae	*Rattus*	*losea*	Lao PDR	Khammouane Province	Liver	2007	Asia	
ABTC	119389	57	ME	Rodentia	Muridae	*Rattus*	*losea*	Thailand	Loei	Liver	2003	Asia	
ABTC	119187	54	ME	Rodentia	Muridae	*Rattus*	*losea*	Vietnam	Vinh Phu Province	Liver	2002	Asia	
ABTC	124325	65	ME	Rodentia	Muridae	*Rattus*	*marmosurus*	Indonesia	Sulawesi	Liver	N/A	Wallacea	
ABTC	92057	78	ME	Rodentia	Muridae	*Rattus*	*niobe*	Indonesia	West Papua	Liver	N/A	Australo-Papuan	
ABTC	121321	59	ME	Rodentia	Muridae	*Rattus*	*niobe*	Papua New Guinea	Western Province	Liver	N/A	Australo-Papuan	
ABTC	141256	302	ME	Rodentia	Muridae	*Rattus*	*niobe*	Papua New Guinea	Southern Highlands Province	Liver	N/A	Australo-Papuan	
ABTC	140467	89	ME	Rodentia	Muridae	*Rattus*	*niobe sp. b*	Papua New Guinea	Southern Highlands Province	Liver	N/A	Australo-Papuan	*
ABTC	125297	68	ME	Rodentia	Muridae	*Rattus*	*nitidus*	Lao PDR	Luang Numtha	Liver	N/A	Asia	
ABTC	43216	229	MF	Rodentia	Muridae	*Rattus*	*nitidus*	Papua New Guinea	Chimbu Province	Liver	1983	Australo-Papuan	
ABTC	27454	179	MF	Rodentia	Muridae	*Rattus*	*norvegicus*	Australia	South Australia	Liver	1998	Australo-Papuan	

SAM ID		Sample no.		Order	Family	Genus	Species	Country	State/Province/District	Tissue	Collection year	Geographic units	Used for library prep
ABTC	125189	67	ME	Rodentia	Muridae	*Rattus*	*norvegicus*	Cambodia	Takéo Province	Liver	N/A	Asia	
ABTC	119314	55	ME	Rodentia	Muridae	*Rattus*	*norvegicus*	Philippines		Tail	N/A	Philippines	
ABTC	119179	53	ME	Rodentia	Muridae	*Rattus*	*norvegicus*	Vietnam	Hung Yen	Liver	2002	Asia	
ABTC	46853	251	MF	Rodentia	Muridae	*Rattus*	*novaeguineae*	Papua New Guinea	Southern Highlands Province	Liver	1987	Australo-Papuan	
ABTC	91989	77	ME	Rodentia	Muridae	*Rattus*	*praetor*	Indonesia	West Papua	Liver	N/A	Australo-Papuan	
ABTC	42614	211	MF	Rodentia	Muridae	*Rattus*	*praetor*	Papua New Guinea	Sandaun Province	Not described	1984	Australo-Papuan	
ABTC	149256	75	ME	Rodentia	Muridae	*Rattus*	*rattus*	Australia	Northern Territory	Ear	2011	Australo-Papuan	
ABTC	107492	198	MF	Rodentia	Muridae	*Rattus*	*rattus*	Australia	Queensland	Kidney	N/A	Australo-Papuan	
ABTC	141604	199	MF	Rodentia	Muridae	*Rattus*	*rattus*	Papua New Guinea	Gulf Province	Liver in EtoH	2016	Australo-Papuan	
ABTC	100495	36	ME	Rodentia	Muridae	*Rattus*	*sordidus*	Australia	Queensland	Liver	2005	Australo-Papuan	
ABTC	41160	206	MF	Rodentia	Muridae	*Rattus*	*sordidus*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	
ABTC	133598	72	ME	Rodentia	Muridae	*Rattus*	*sp cf niobe*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	133606	73	ME	Rodentia	Muridae	*Rattus*	*sp cf rubex*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	140466	287	ME	Rodentia	Muridae	*Rattus*	*sp longnose*	Papua New Guinea	Southern Highlands Province	Liver	N/A	Australo-Papuan	
ABTC	140468	288	ME	Rodentia	Muridae	*Rattus*	*sp longnose*	Papua New Guinea	Southern Highlands Province	Liver	N/A	Australo-Papuan	
ABTC	140469	289	ME	Rodentia	Muridae	*Rattus*	*sp longnose*	Papua New Guinea	Southern Highlands Province	Liver	N/A	Australo-Papuan	
ABTC	140468	301	ME	Rodentia	Muridae	*Rattus*	*sp longnose*	Papua New Guinea	Southern Highlands Province	Liver	N/A	Australo-Papuan	
ABTC	129180	82	ME	Rodentia	Muridae	*Rattus*	*sp. a*	Papua New Guinea	Western Province	Liver	2013	Australo-Papuan	
ABTC	140470	90	ME	Rodentia	Muridae	*Rattus*	*sp. spiny*	Papua New Guinea	Southern Highlands Province	Liver	N/A	Australo-Papuan	
ABTC	141617	200	MF	Rodentia	Muridae	*Rattus*	*steini*	Papua New Guinea	Gulf Province	Liver	2016	Australo-Papuan	
ABTC	109482	39	ME	Rodentia	Muridae	*Rattus*	*tanezumi*	Australia	Queensland	Liver	N/A	Australo-Papuan	
ABTC	109667	40	ME	Rodentia	Muridae	*Rattus*	*tanezumi*	Australia	Northern Territory	Liver	2006	Australo-Papuan	
ABTC	110191	41	ME	Rodentia	Muridae	*Rattus*	*tanezumi*	Indonesia	Belang Watokobu, Lembata	Liver	N/A	Wallacea	
ABTC	110206	44	ME	Rodentia	Muridae	*Rattus*	*tanezumi*	Indonesia	Pelangan, Lombok	Liver	N/A	Wallacea	
ABTC	44857	247	MF	Rodentia	Muridae	*Rattus*	*tanezumi*	Papua New Guinea	National Capital District	Liver	1985	Australo-Papuan	
ABTC	110242	49	ME	Rodentia	Muridae	*Rattus*	*tiomanicus*	Indonesia	Selo Boyolali, Java	Liver	N/A	Asia	
ABTC	100494	35	ME	Rodentia	Muridae	*Rattus*	*tunneyi*	Australia	Queensland	Liver	2005	Australo-Papuan	
ABTC	44796	246	MF	Rodentia	Muridae	*Rattus*	*verecundus*	Papua New Guinea	Southern Highlands Province	Liver	1987	Australo-Papuan	*
ABTC	45161	285	MF	Rodentia	Muridae	*Rattus*	*verecundus*	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	
ABTC	46375	286	MF	Rodentia	Muridae	*Rattus*	*verecundus*	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	
ABTC	43424	295	MF	Rodentia	Muridae	*Rattus*	*verecundus*	Papua New Guinea	Chimbu Province	Liver	1983	Australo-Papuan	
ABTC	45161	296	MF	Rodentia	Muridae	*Rattus*	*verecundus*	Papua New Guinea	Southern Highlands Province	Liver	1985	Australo-Papuan	
ABTC	47140	297	MF	Rodentia	Muridae	*Rattus*	*verecundus*	Papua New Guinea	Sandaun Province	Kidney	1986	Australo-Papuan	
ABTC	8600	193	MF	Rodentia	Muridae	*Rattus*	*villosissimus*	Australia	Queensland	Heart	N/A	Australo-Papuan	
ABTC	41146	205	MF	Rodentia	Muridae	*Rattus*	*villosissimus*	Australia	Northern Territory	Liver	N/A	Australo-Papuan	

Table S2.2 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 (refer to materials and 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
R269G	ENV	-2	-3	83.2	N/A	N/A	No	tolerated	neutral	N/A

Replacement Mutations	Protein	BLOSUM 62*	BLOSUM 80*	Mismatches (%)	Motif	Conserved Domain Consensus Amino Acid	3D model	SIFT prediction	PROVEAN prediction	Major Structural change
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

Table S3.1 Summary of the statistical models used from the HyPhY package to infer selection pressures on five rodent antiretroviral gene families (APOBEC3, BST-2, SAMHD1, TRIM5 and ZAP). The length (bp) of *Melomys* coding sequences (CDS) used as a query to retrieve the homologous sequences are indicated. In branch-site models, *b* indicates number of branches in each gene phylogeny. *N* = number of sequences in each alignment. *c* = number of codons (sites) in each alignment. pp = posterior probability threshold. LTR = Likelihood test ratio. logL = the log likelihood of the fitted model. AIC-c= the small-sample AIC for the fitted model. dN/dS (ω) = rate of non-synonymous over synonymous change < 1 shows a constrain purifying selection, whereas the excess of non-synonymous changes is an indication of diversifying positive selection trend.

		branch-site models								site-models						
		BUSTED					aBSREL		FUBAR			MEME		FEL	
			rate distribution				LTR ≤ 0.05		pp= 0.9 # sites under selection			p-value threshold 0.1		p-value ≤ 0.1	
Genes	*Melomys* CDS length (bp)	Episodic diversification?	negative (ω1)	neutral (ω2)	positive (ω3)	LTR ≤ 0.05	*b*	# *b* with episodic (diversifying) selection	*N*	pervasive (diversifying)	negative	*c*	# sites under diversifying (+) selection	Global MG94xREV | AICC | logL	diversifying (+)	Purifying (-)
APOBEC3	1,188	Yes	0.37 (60.64%)	0.37 (15.80%)	2.52 (23.56%)	0.032	50	0	33	12	22 (3%)	707	24	13709.99 | -6789.81 | ω= 0.815	11	48
BST-2	483	No				0.472	17	0	10	4	21 (11%)	189	5	4191.29 | -2063.06 | ω= 0.598	2	29
SAMHD1	1,686	No				0.446	41	2	25	8	163 (17%)	933	12	14972.43 | -7430.08 | ω= 0.261	8	156
TRIM5	1,485	Yes	0.12 (24.36%)	0.55 (60.39%)	3.03 (15.25%)	0	71	4	38	22	65 (7%)	981	58	27723.49 | -13775.54 | ω= 0.756	38	97
ZAP	2,946	No				0.056	32	1 (*Melomys*)	20	4	66 (6%)	1072	28	21731.87 | -10818.83 | ω= 0.5	12	152

Table S3.2 Sites identified by MEME (episodic selection with p-value < 0.1) and FUBAR (with posterior probability > 0.9) models. *N* corresponds to the number of codons (sites), (+) corresponds to sites identified as pervasive diversifying (positive) selection, (-) corresponds to sites identified as pervasive negative (purifying) selection. Sites in alignment identified by both models are indicated.

Genes	*N* sites (total)	Sites identified by MEME and FUBAR	*N* sites MEME	*N* sites (+) FUBAR	*N* sites (-) FUBAR
APOBEC3	707	51, 282, 288, 331, 484, 531	24	12	22
BST-2	189	0	5	4	21
SAMHD1	933	39, 40, 66, 314, 562, 695	12	8	163
TRIM5	981	100, 147, 188, 189, 231, 311, 322, 333, 533, 567, 588, 601, 606, 609, 714, 732, 735, 737, 748, 763, 887	58	22	65
ZAP	1072	260, 460, 602	28	4	66

Text file S2.1

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

>cMWMV_consensus|complete_genome|8459bp|

TGTTTTTCAAGCTAGCTGCAGTAACGCCATTTTGCAAGGCACGGAAAATTACCCTGGTAAAAAGCCCAAAGCATAGGGAAAGTACAGCTAAAGGTCAAGTCGAGA
AAAACAAGGAGAACAGGGCCAAACAGGATATCTGTGGTCATGCACCTGGGCCCCGGCCCAGGGCCAAAGACAGATGGTTCCCAGAAATAGATGAGTCAACAGCA
GTTTCCAGGGTGCCCCTCAACTGTTTCAAGAACTCCCACATGACCGGAGCTCACCCCTGGCCTTATTTGAACTGACCAATTACCTTGCTTCTCGCTTCTGTACCCGCG
CTTTTTGCTATAAAATGAGCTCAGAAACTCCACTCGGCGCGCCAGTCCTCCGAGAGACTGAGTCGCCCGGGTACTCGTGTGTTCAATAAAACCTCTTGCTATTTGCAT
CCGAAGCCGTGGTCTCGTTGTTCCTTGGGAGGGTCCCTCCTAACTGATTGACTGCCCACCTCGGGGGTCTTTCATTTGGGGGCTCGTCCGGGATCGGAGACCCCCA
CCCAGGGACCACCGACCCACCAACGGGAGGTAAGCTGGCCAGCGATCGCTCTGTGTCTCGTTTCTGTGTCTAACTCCGTAACTCTGACTGTCCCTCTGAGTGCGCG
CATTTTGGTTTCAGTTTGTTCCGGGCTGATCGCTCTGTGAGCGACGTGTGAGTAGCGAGCAGACGTGTTCGGGGGCTCACCGCCCGATAATCCTGGGAGACGTCCC
AGGATCAGGGGAGGACCAGGGACGCCTGGTGGACCCCTCGGCAGAGGATCATTGTGTTCTGATCCCACTGCCGCGTCTAGAGAGGCGCGCTCTGCCATCTGACTC
TTTCTTTTGCCTCTACGCTACTCGATCTCGCCGCCGTTTCTGGTTTCTTTTTGTTTTGTTTCTGATAAGCCTCTGTGTCGTAGTCCCTCTCTTCGAAATCTTCGAAATGG
GACAAGATAACTCTACCCCTCTCTCCCTTACTCTAGATCACTGGAAAGATGTGAGAACAAGGGCTCACAATCTGTCCGTGGAAATCAGAAAGGGAAAATGGCAGAC
TTTCTGTTCCTCCGAGTGGCCCACATTCGGCGTGGGGTGGCCGCCGGAGGGAACTTTTAATCTCTCTGTCATTTTTGCAGTTAAAAGGATTGTCTTTCAGGAAACCG
GAGGACACCCGGACCAAGTTCCATACATCGTGGTGTGGCAGGACCTCGCCCAGAGTCCCCCACCATGGGTGCCGCCCTCCGTCAAGATCGCTGTTGTCTCTAGTCC
AGAGAACACTCGAGGACCAGCTGCGGGGAGGCCATCCGCTCCTCCCCGACCCCCCATCTACCCGGCAACAGACGACTTGCTCCTTCTCTCTGAGCCCCCGCCCTAT
CCGGCGGCCCTGCCACCTCCTCTGGCCCCTCAGGCGGTCGGACCGGCGCCGGGCCAGGCGCCCGATAGTTCCGATCCTGAGGGACCAGCCGCTGGGACCAGGAG
TCGCCGTGCCCGCAGTCCGGCAGACGACTCGGGTCCTGACTCCACTGTGATTTTGCCCCTCCGAGCCATAGGACCCCCAGCCGAGCCCAACGGCCTGGTCCCTCTA
CAATATTGGCCTTTTTCCTCAGCAGATCTTTATAATTGGAAATCTAACCATCCTTCTTTTTCTGAAAATCCAGCAGGACTCACGGGGCTCCTTGAGTCTCTTATGTTTTC
TCATCAGCCCACTTGGGACGATTGCCAACAGCTCCTACAGATTCTCTTCACCACTGAGGAGCGGGAAAGGATTCTCCTGGAGGCCCGCAAGAATGTCCTTGGGGAC
AACGGGGCCCCTACTCAACTCGAGAACCTCATTAATGAGGCCTTCCCCCTCAATCGACCTCAGTGGGATTACAACACGGCCGCAGGTAGGGAGCGTCTCCTGGTCT
ACCGCCGGACTCTAGTGGCAGGTCTCAAAGGGGCAGCTCGGCGCCCCACCAATTTGGCTAAGGTAAGAGAGGTCTTGCAGGGACCGGCAGAACCCCCTTCGGTT
TTCTTAGAACGCCTAATGGAGGCCTATAGGAGATACACTCCGTTTGAACCCTCTTCTGAGGGGCAGCAGGCTGCGGTTGCCATGGCCTTCATCGGACAGTCAGCCC
CAGATATCAAGAAAAAGTTACAGAGGCTAGAGGGGCTCCAAGATTATTCCTTACAAGATTTAGTGAGAGAGGCAGAGAAGGTGTACCACAAGAGAGAGACAGAA
GAAGAAAGACAAGAAAGAGAAAAGAAAGAGGCAGAAGAGAGAGAGAGGCGGCGCGATAAGCGCCAAGAGAAAAACTTGACTAGGATTTTGGCCGCAGTGGT
AAGTGAAAGAGGGTCTAGAGATAGGCAGACAGGGAACCTGAGCAACCGGGCAAGGAAGACACCTAGGGATGGAAGACCTCCTCTAGACAAAGACCAGTGCGCG
TACTGTAAAGAGAAGGGTCACTGGGCAAGAGAATGTCCCCGAAAGAAGAACGTCAGAGAAGCCAAGGTTCTGTCCCTAGATGACTAGGGGAGTCGGGGTTCGG
ACCCCCTCCCCGAACCTAGGGTAACACTGACTGTGGAGGGGACCCCCATTGAGTTCCTGGTCGATACCGGGGCTGAACATTCGGTATTGACCCAACCCATGGGAAA
GGTAGGGTCCAGACGGACAGTCGTGGAAGGAGCGACAGGAAGCAAAGTCTACCCCTGGACCACAAAGAGACTTTTAAAAGTTGGACATAAACAAGTGACCCACT
CCTTTCTGGTCATACCCGAGTGCCCCGCTCCTCTGTTGGGCAGGGACCTCCTAACCAAACTAAAGGCCCAGATCCAGTTTTCTGCTGAGGGCCCACAGGTAACATG
GGAAGACCGCCCTACTATGTGCCTGGTCCTAAACCTAGAAGAAGAATACCGGCTACATGAAAAGCCAGTTCCCTCCTCTATCGACCCATCCTGGCTCCAGCTTTTTCC
CACTGTATGGGCAGAGAGGGCAGGCATGGGACTGGCCAATCAAGTCCCACCAGTGGTAGTAGAACTGAGATCAGGTGCCTCACCGGTGGCTGTTCGACAATATCC
AATGAGCAAAGAAGCCCGGGAAGGTATCAGACCCCACATCCAAAGGTTCTTAGACCTAGGGGTCTTGGTGCCCTGTCAGTCGCCCTGGAATACCCCTCTACTACCT
GTAAAGAAGCCAGGGACCAATGACTATCGGCCAGTCCAAGACCTGAGAGAAATTAATAAGAGGGTACAGGATATTCATCCCACAGTCCCAAACCCTTACAACCTTCT
GAGTTCCCTTCCGCCTAGCCACACTTGGTACTCAGTCTTAGATCTCAAGGATGCCTTTTTCTGCCTCAAGCTACATCCCAACAGCCAGCCGCTGTTCGCGTTCGAGTG
GAGAGACCCAGAAAAAGGTAACACAGGTCAGCTGACCTGGACACGGCTACCACAAGGGTTCAAGAACTCTCCCACTCTCTTCGACGAGGCCCTCCACCGAGATTT
GGCTCCCTTTAGGGCCCTCAACCCCCAGGTAGTGTTACTCCAATATGTAGACGACCTCCTGGTGGCGGCCCCCACATATAGAGACTGCAAAGAAGGGACACAGAAG
CTCCTACAGGAATTGAGTAAGTTGGGGTACCGGGTATCGGCTAAGAAGGCCCAGCTCTGCCAGAAAGAGGTCACCTATCTGGGGTACTTGCTCAAGGAAGGGAAA
AGATGGCTGACCCCGGCCCGAAAGGCTACTGTTATGAAGATCCCCGCTCCCACGACCCCCAGACAGGTCCGTGAATTTCTGGGCACTGCTGGATTCTGCAGGCTCT
GGATCCCTGGGTTTGCTTCCCTGGCTGCACCCTTGTACCCCCTAACAAAGGAAAGCATCCCTTTTATCTGGACTGAGGAACATCAGAAGGCTTTTGACCGCATAAAA
GAAGCCTTGCTGTCAGCCCCCGCTTTGGCCCTCCCAGACCTCACCAAACCATTTACTCTATACGTAGATGAGAGGGCCGGCGTGGCCCGGGGAGTGCTTACTCAGA
CTTTAGGACCCTGGCGGCGGCCGGTAGCTTATCTATCGAAGAAACTGGATCCGGTGGCCAGCGGGTGGCCAACCTGCCTGAAAGCGGTTGCAGCAGTGGCACTCC
TTCTCAAGGACGCTGATAAGTTAACCTTGGGACAAAATGTGACTGTGATTGCTTCCCATAGCCTCGAAAGCATCGTGCGGCAGCCCCCCGACCGGTGGATGACCAA
TGCCAGGATGACTCATTACCAGAGCCTGCTGCTAAATGAAAGGGTATCGTTCGCGCCCCCTGCCGTCCTGAACCCAGCTACCCTACTACCAGTCGAGTCGGAAGCCA
CCCCAGTGCACAGGTGCTCAGAAATCCTCGCCGAAGAAACTGGAACTCGACGAGACCTGAAGGACCAGCCATTGCCCGGGGTGCCAGCCTGGTATACGGACGGT
AGCAGTTTCATCGCGGAAGGTAAACGGAGAGCAGGGGCCGCCATCGTAGATGGCAAGCGGACGGTGTGGGCAAGCAGCCTGCCAGAAGGTACGTCAGCCCAGA
AGGCCGAACTAGTGGCCTTGACGCAGGCATTACGCCTGGCCGAAGGAAGAGACATCAACATCTACACAGACAGCAGGTATGCTTTTGCCACTGCTCATATTCATGG
GGCAATATATAAACAGAGGGGGCTGCTCACTTCTGCTGGAAAAGACATTAAAAACAAAGAAGAAATTTTGGCCCTGCTAGAGGCCATTCACCTTCCTAAGCGGGTC
GCCATTATCCACTGCCCCGGCCACCAGAGGGGAAATGACCCTGTGGCCACCGGGAACCGGAGGGCCGACGAGGCTACAAAACAAGCCGCCCTGTCGACCAGAGT
GCTGGCAGAAACTACAAAACCTCAAGAGCTAATCGAACCCGCTCAGGTTAAGACCAGGCCAGGAGAGCTCACCCCTGACCGGGGGAAGGAATTCATTCAGCGGT
TACATCAGTTAACACACTTAGGACCAGAGAAGCTTCTCCAACTAGTAAACCGCACCAGCCTCCTCATCCCGAACCTCCAGTCTGCAGTTCGCGAAGTCACCAGTCAG
TGTCAGGCTTGTGCCATGACTAATGCGGTCACAACCTACCGAGAGACCGGAAAGAGGCAACGAGGAGATCGACCCGGCGTGTACTGGGAGGTAGACTTCACAGA
GGTGAAGCCTGGCCGGTATGGAAACAGGTATCTGCTGGTATTCATAGATACTTTTTCCGGATGGATAGAAGCTTTTCCTACCAAAACTGAAACGGCCCTGACCGTCT
GCAAGAAGATATTAGAAGAAATTCTACCCCGCTTCGGGATCCCTAAGGTACTCGGGTCAGACAATGGCCCAGCCTTTGTTGCTCAGGTAAGTCAGGGACTGGCCAC
TCAACTGGGGATAAATTGGAAGTTACATTGTGCGTATAGACCCCAGAGCTCAGGTCAGGTAGAGAGAATGAACAGGACAATCAAAGAGACCTTGACCAAATTAGCC
TTAGAGACCGGTGGGAAAGACTGGGTGGCCCTCCTTCCCTTAGCGCTGCTCAGAGCCAGGAATACCCCTGGCCGGTTTGGTCTAACTCCTTATGAAATTCTCTATGG
GGGACCGCCCCCCATACTTGAGTCTGGAGGGACATTGGGTCCCGATGATAATTTTCTCCCTGTCTTATTTACTCATTTAAAGGCTTTAGAAGTTGTGAGGACCCAGAT
CTGGGACCAGATCAAGGAGGTGTACAAGCCCGGTACCGTGGCAATGCCCCACCCGTTCCAGGTCGGGGACCAAGTGCTTGTCAGACGCCATCGATCCAGCAGCCT
TGAGCCTCGGTGGAAAGGCCCGTACCTGGTGTTGCTGACCACCCCGACCGCGGTAAAAGTCGACGGTATCGCTGCCTGGATCCATGCTTCTCACCTCAAGCCTGCA
CCACCCTCGGCACCAGATGAGTCCTGGGAGCTGGAAAAGACTGATCATCCTCTTAAGCTGCGTATTCGGCGGCGGCGGAACGAGTCTGCAAAATAAGAACCCCCA
CCAGCCTATGACCCTCACCTGGCAGGTACTGTCCCAAACTGGAGACGTTGTCTGGGTCACAAAGGCAGTCCAGCCCCTTTGGACTTGGTGGCCCTCTCTTGAACCT
GATGTATGTGCCCTGGCGGCCGGTCTTGAGTCCTGGGATATCCCGGGATCCGATGTATCGGCCTCTAAAAGAATCAGACCCCCTGACTCAAACTATAATGACGCTAAT

AAGCAGATCAGCTGGGGAGCCATAGGATGCAGCTACCCTCGGGCTAGGACCAGGATTGCAAATTCCCCCTTCTACGTGTGTCCCCGGGATGGCCGGACCCTTTCAG
AAACTAGAAGGTGCGGGGGGCTAGAATCCCTGTACTGTAAAGAATGGGGTTGTGAGACCACGGGGACCGTTCATTGGCAACCTAGGTCCTCATGGGACCTCATAA
CTGTAAAATGGGGCCGAAATCGCCAATGGGAGCAAAATATGCTGTCAGTCTGTGAACAAACCGGCTGGTGTAACCCCCTCAAGATAGATTTCACAGAAAAAGGGA
AACACTCCAGGGATTGGATAAAGGGGAGAACCTGGGGATTGAGATTCAATGTGGCTGGACATCCAGGCGTACAGTTGACCATTCGCTTGAAGGTCACCAGCATGC
CAGCTGTGGCAGTGGGCCCCGACCCCGTCCTTGCGGAACAAGGACCTCCTAGCAAGCCCCTCCCTCTCCCCCCGAGGGAAGCGCCGCCCACCTCTCTACCCCCGG
CGGCTAGTGGGCAAGCCCCCACGGTGCGGGAGAGAACTGTTACCCTAAGCACTCCGCCTCCCACCACGGGCGACAGACTCTTTGGCCTTGTGCAGGGGGCCTTCC
TAGCCTTGAATGCTACCAACCCAGGGGCCACAGAGTCTTGCTGGCTCTGTTTGGCCATGGGCCCCCCTTATTATGAAGGAATAGCCTCTTTAGGAGAGGTCGCTTAT
ACCTCCGACCATACCCGGTGCCACTGGGGGGCCCAAGGAAAGCTTACCCTCACTGAGGTCTCAGGACACGGGTTGTGCATAGGAAAGGTGCCCTCTACCCATCAG
CATCTTTGCAATCAGACCCTACCCATCAATTCCTCCAAGGACCATCAGTACCTGCTCCCCTTCAACCATAGCTGGTGGTCCTGCAGCACTGGCCTCACCCCCTGCCTCT
CCACCTCAGTTTTTAATCAGTCTAGAGATTTCTGTATCCATATCCAGCTGATCCCTCGCATCTATTACCATCCTGAAGGAACTTTGTTGCAGGCCTATGACAATCCTCAC
CCCAGGCTTAAAAGAGAGCCTGTCTCACTTACCCTAGCTGTTTTACTGGGGTTGGGGATCGCGACAGGTATAGGTACTGGCTCAGCCGCCCTAATTAAAGGGCCCAT
GGACCTCCAGCAAGGCCTGACCAGCCTCCAGACCGCCATGGATACTGACCTCCGGGCCCTCCAGGACTCAATCAGCAAGCTGGAGGACTCGCTGACTTCCCTATCT
GAGGTAGTGCTCCAAAATAGGAGAGGCCTTGACTTACTGTTTCTAAAAGAAGGAGGCCTCTGCGCGGCCCTAAAAGAAGAGTGCTGTTTTTATGTGGACCACTCA
GGTGCAGTACGAGACTCCATGAGAAAGCTCAAAGAAAGACTAGATAAGAGACAGTTAGAGCGCCAGAAGAACCAAAACTGGTATGAAGGGTGGTTCAATAGCTC
CCCTTGGTTCACTACCCTACTATCAACCATCGCCGGGCCCCTATTACTCCTCCTTCTGTTGCTCACCCTCGGGCCCTGCATCATCAATAGGTTAGTCCAATTCATCAATG
ATAGGGTAAGTGCAGTTAAAATTCTGGTCCTTAGACAGAAATATCAGACCCTAGATAACGAAGATAACCTTTGATTCCGCTCTAAGATTAGAGCTATCCACAAGAGAA
ATGGGGGAATGAAGGAAGTGTTTTTCAAGCTAGCTGCAGTAACGCCATTTTGCAAGGCACGGAAAATTACCCTGGTAAAAAGCCCAAAGCATAGGGAAAGTACAG
CTAAAGGTCAAGTCGAGAAAAACAAGGAGAACAGGGCCAAACAGGATATCTGTGGTCATGCACCTGGGCCCCGGCCCAGGGCCAAAGACAGATGGTTCCCAGAA
ATAGATGAGTCAACAGCAGTTTCCAGGGTGCCCCTCAACTGTTTCAAGAACTCCCACATGACCGGAGCTCACCCCTGGCCTTATTTGAACTGACCAATTACCTTGCTT
CTCGCTTCTGTACCCGCGCTTTTTGCTATAAAATGAGCTCAGAAACTCCACTCGGCGCGCCAGTCCTCCGAGAGACTGAGTCGCCCGGGTACTCGTGTGTTCAATAA
AACCTCTTGCTATTTGCATCCGAAGCCGTGGTCTCGTTGTTCCTTGGGAGGGTCCCTCCTAACTGATTGACTGCCCACCTCGGGGGTCTTTC

Text file S3.1

The newick format of Muridae species tree. This tree was used by DGINN pipeline to separate the sequences into orthologous groups and reconcile gene trees. This tree was extracted from TimeTree.org and includes a total of 465 species. The nomenclature displayed in gene trees, follows DGINN parameter requirement of speSpe.

((((((((((Tonkinomys_daovantieni:2.66522000,Saxatilomys_paulinae:2.66522000)'14':8.53478000,(((((((((Bunomys_andrewsi:3.09000000,Bunomys_chrysocomus:3.09000000)'13':0.00000000,Halmaheramys_bokimekot:3.09000000)'11':3.80076556,Sundamys_muelleri:6.89076556)'10':2.40923444,((Diplothrix_legata:6.43731600,(Bandicota_savilei:2.40784000,(Bandicota_bengalensis:1.98807600,Bandicota_indica:1.98807600)'19':0.41976400)'9':4.02947600)'22':2.86268400,((Rattus_nativitatis:5.66704000,Rattus_mordax:5.66704000)'8':3.63296000,(((((((Rattus_argentiventer:2.35237400,Rattus_hoffmanni:2.35237400)'6':0.20744933,(((Rattus_rattus:1.42589042,Rattus_tanezumi:1.42589042)'30':0.60871792,Rattus_losea:2.03460833)'29':0.05413667,Rattus_tiomanicus:2.08874500)'27':0.47107833)'35':0.15931333,Rattus_andamanensis:2.71913667)'43':0.69074913,Rattus_exulans:3.40988579)'42':2.45374171,((Rattus_baluensis:5.78352000,Rattus_satarae:5.78352000)'40':0.00000000,Rattus_pyctoris:5.78352000)'48':0.08010750)'51':0.00000000,((((((Rattus_tunneyi:2.54511400,Rattus_lutreolus:2.54511400)'47':0.13280367,(((Rattus_villosissimus:1.81101329,Rattus_sordidus:1.81101329)'39':0.44500043,Rattus_leucopus:2.25601371)'56':0.22797462,Rattus_giluwensis:2.48398833)'55':0.19392933)'61':0.30621667,Rattus_fuscipes:2.98413433)'60':0.05451200,((Rattus_steini:2.20797600,Rattus_praetor:2.20797600)'54':0.13051900,(Rattus_morotaiensis:1.19961000,Rattus_novaeguineae:1.19961000)'38':1.13888500)'34':0.70015133)'26':0.27578167,Rattus_colletti:3.31442800)'66':0.89560200,Rattus_everetti:4.21003000)'75':1.65359750)'80':0.67756000,(Rattus_nitidus:2.45605333,Rattus_norvegicus:2.45605333)'78':4.08513417)'74':0.36885250,((Rattus_macleari:3.38823000,Rattus_niobe:3.38823000)'83':1.13290800,Rattus_verecundus:4.52113800)'73':2.38890200)'88':2.38996000)'86':0.00000000)'72':0.00000000)'93':0.00000000,(((Tarsomys_apoensis:3.70478143,(Limnomys_bryophilus:1.05380500,Limnomys_sibuanus:1.05380500)'92':2.65097643)'91':2.09206524,(Bullimus_luzonicus:2.00927000,(Bullimus_gamay:1.23959750,Bullimus_bagobus:1.23959750)'71':0.76967250)'69':3.78757667)'65':2.55000000,(Taeromys_celebensis:1.65068000,Paruromys_dominator:1.65068000)'25':6.69616667)'5':0.95315333)'102':0.00000000,(Berylmys_berdmorei:3.61605000,Berylmys_bowersi:3.61605000)'100':5.68395000)'107':0.00000000,Nesokia_indica:9.30000000)'111':0.00000000,Srilankamys_ohiensis:9.30000000)'110':1.90000000,((((Niviventer_tenaster:9.06000000,((((Nivivneter_brahma:4.35958000,Niviventer_eha:4.35958000)'106':4.60715667,(Niviventer_bukit:3.00452000,Niviventer_coninga:3.00452000)'129':5.96221667)'128':0.09326333,Niviventer_andersoni:9.06000000)'133':0.00000000,(((Niviventer_excelsior:6.41367500,(Niviventer_confucianus:5.89565333,Niviventer_culturatus:5.89565333)'127':0.51802167)'136':0.41971500,(((Niviventer_niviventer:1.15681000,Niviventer_huang:1.15681000)'126':3.51817857,Niviventer_fulvescens:4.67498857)'142':0.53554976,((Niviventer_sp._2_MP-2010:4.60945000,Niviventer_sp._1_MP-2010:4.60945000)'146':0.00000000,Niviventer_cremoriventer:4.60945000)'150':0.60108833)'149':1.62285167)'145':0.90747000,Niviventer_rapit:7.74086000)'141':1.31914000)'140':0.00000000)'139':0.00000000,Margaretamys_elegans:9.06000000)'125':2.14000000,(Chiromyscus_chiropus:11.20000000,Chiromyscus_langbianis:11.20000000)'124':0.00000000)'160':0.00000000,((Leopoldamys_milleti:5.33506000,(((Leopoldamys_revertens:1.98830000,Leopoldamys_edwardsi:1.98830000)'159':2.28084625,Leopoldamys_neilli:4.26914625)'158':0.58784708,Leopoldamys_sabanus:4.85699333)'157':0.47806667)'123':2.91699333,Dacnomys_millardi:8.25205333)'122':2.94794667)'121':0.00000000)'120':0.00000000)'119':0.20963833,Melasmothrix_naso:11.40963833)'118':0.99036167,((Crunomys_suncoides:0.70943250,Crunomys_melanius:0.70943250)'117':11.69056750,(((Maxomys_bartelsii:7.50748000,Maxomys_surifer:7.50748000)'115':1.14317000,Maxomys_whiteheadi:8.65065000)'105':3.74935000,(((Maxomys_rajah:12.40000000,Maxomys_hellwaldii:12.40000000)'99':0.00000000,(Maxomys_moi:12.40000000,Maxomys_ochraceiventer:12.40000000)'177':0.00000000)'185':0.00000000,((Maxomys_pagensis:12.40000000,Maxomys_musschenbroekii:12.40000000)'184':0.00000000,Maxomys_wattsi:12.40000000)'183':0.00000000)'182':0.00000000)'180':0.00000000)'176':0.00000000)'192':1.86251

909,Micromys_minutus:14.26251909)'195':6.62489831,((((((((((((((Paramelomys_lorentzii:8.25000000,((Paramelomys_rubex:8.15971000,Paramelomys_levipes:8.15971000)'191':0.09029000,(Paramelomys_moncktoni:7.98637667,Paramelomys_platyops:7.98637667)'175':0.26362333)'205':0.00000000)'204':0.00000000,(Solomys_salebrosus:6.88193000,Solomys_ponceleti:6.88193000)'203':1.36807000)'202':0.00000000,(Melomys_leucogaster:7.19197033,((Melomys_cervinipes:6.97828000,((Melomys_capensis:6.78578000,Melomys_rubicola:6.78578000)'201':0.19250000,((Melomys_sp._LMB-2011:0.49000000,Melomys_lutillus:0.49000000)'212':2.16411333,Melomys_burtoni:2.65411333)'200':4.32416667)'218':0.00000000)'229':0.12027767,Melomys_rufescens:7.09855767)'228':0.09341267)'227':1.05802967)'226':0.00000000,(Uromys_anak:7.53000000,(Uromys_hadrourus:5.66702333,Uromys_caudimaculatus:5.66702333)'225':1.86297667)'224':0.72000000)'223':0.00000000,(((Mesembriomys_gouldii:3.65000000,Mesembriomys_macrurus:3.65000000)'243':0.00000000,Conilurus_penicillatus:3.65000000)'242':0.40826429,Leporillus_conditor:4.05826429)'241':4.19173571)'240':0.55000000,(((((Pseudomys_fumeus:8.80000000,(((((((((Pseudomys_albocinereus:8.67061667,Pseudomys_apodemoides:8.67061667)'239':0.12938333,Pseudomys_australis:8.80000000)'251':0.00000000,(Pseudomys_oralis:8.80000000,Pseudomys_gracilicaudatus:8.80000000)'250':0.00000000)'238':0.00000000,((((Pseudomys_johnsoni:8.80000000,Pseudomys_patrius:8.80000000)'222':0.00000000,(Pseudomys_calabyi:3.28395000,Pseudomys_laborifex:3.28395000)'221':5.51605000)'217':0.00000000,Pseudomys_occidentalis:8.80000000)'216':0.00000000,(Pseudomys_fieldi:8.75061667,Pseudomys_higginsi:8.75061667)'265':0.04938333)'268':0.00000000)'264':0.00000000,((((Pseudomys_pilligaensis:8.66728333,Pseudomys_novaehollandiae:8.66728333)'274':0.07333333,Pseudomys_delicatulus:8.74061667)'273':0.05938333,Pseudomys_bolami:8.80000000)'272':0.00000000,Pseudomys_hermannsburgensis:8.80000000)'271':0.00000000)'263':0.00000000,(Pseudomys_desertor:8.80000000,Pseudomys_shortridgei:8.80000000)'262':0.00000000)'261':0.00000000,Pseudomys_chapmani:8.80000000)'260':0.00000000,Pseudomys_nanus:8.80000000)'259':0.00000000)'257':0.00000000,Mastacomys_fuscus:8.80000000)'215':0.00000000,((Notomys_cervinus:6.54928000,Notomys_aquilo:6.54928000)'199':0.00000000,((Notomys_fuscus:6.54928000,Notomys_alexis:6.54928000)'284':0.00000000,Notomys_mitchellii:6.54928000)'198':0.00000000)'174':2.25072000)'173':0.00000000,(Leggadina_lakedownensis:2.74404000,Leggadina_forresti:2.74404000)'298':6.05596000)'296':0.00000000,((Zyzomys_maini:7.81812000,Zyzomys_pedunculatus:7.81812000)'301':0.00000000,((Zyzomys_argurus:7.81812000,Zyzomys_palatilis:7.81812000)'295':0.00000000,Zyzomys_woodwardi:7.81812000)'305':0.00000000)'308':0.98188000)'304':0.00000000)'294':0.32764100,(((Leptomys_elegans:5.17705571,Pseudohydromys_ellermani:5.17705571)'313':1.66073429,Parahydromys_asper:6.83779000)'317':0.45714286,(Hydromys_chrysogaster:5.17587167,Xeromys_myoides:5.17587167)'316':2.11906119)'312':1.83270814)'327':1.77235900,(Mammelomys_rattoides:10.90000000,Mammelomys_lanosus:10.90000000)'326':0.00000000)'332':0.00000000,(((Chiruromys_vates:9.26512286,Anisomys_imitator:9.26512286)'331':0.46068857,Hyomys_goliath:9.72581143)'325':0.00000000,Lorentzimys_nouhuysi:9.72581143)'324':0.26363429,((Mallomys_rothschildi:8.89238500,Abeomelomys_sevia:8.89238500)'323':0.63829125,Macruromys_major:9.53067625)'321':0.45876946)'311':0.91055429)'293':0.00000000,(Pogonomys_loriae:5.47020833,(Pogonomys_sylvestris:5.12569400,Pogonomys_macrourus:5.12569400)'343':0.34451433)'346':5.42979167)'342':1.47174545,(((((Murinae_sp._SAJ-2012d:1.40399333,(Murinae_sp._SAJ-2012c:1.15420000,Murinae_sp._SAJ-2012b:1.15420000)'340':0.24979333)'350':1.44392667,(Chrotomys_silaceus:2.47486400,(((Chrotomys_whiteheadi:1.27801500,Chrotomys_mindorensis:1.27801500)'353':0.13042750,Chrotomys_gonzalesi:1.40844250)'349':0.20474750,Chrotomys_sibuyanensis:1.61319000)'339':0.86167400)'358':0.37305600)'361':3.73208000,(Rhynchomys_soricoides:6.58000000,Rhynchomys_isarogensis:6.58000000)'357':0.00000000)'365':0.00000000,(Archboldomys_sp._SAJ-2012:1.81671333,Archboldomys_luzonensis:1.81671333)'369':4.76328667)'368':0.52806222,(((((((Apomys_sp._SJS-2010c:0.40973000,Apomys_sp._SJS-2010b:0.40973000)'364':0.11585000,(Apomys_lubangensis:0.38693000,Apomys_sacobianus:0.38693000)'356':0.13865000)'338':0.66008000,Apomys_sp._SJS-2010e:1.18566000)'292':0.31442000,Apomys_datae:1.50008000)'377':0.51028000,Apomys_gracilirostris:2.01036000)'375':2.04644667,(((((Apomys_iridensis:0.55898000,Apomys_sp._SJS-2010g:0.55898000)'383':0.09485000,Apomys_sp._SJS-2010a:0.65383000)'391':0.19624000,Apomys_sp._SJS-2010d:0.85007000)'389':0.33470000,Apomys_sp._SJS-2010f:1.18477000)'399':0.52790000,Apomys_abrae:1.71267000)'398':2.34413667)'397':0.16860667,((Apomys_musculus:2.72403714,Apomys_microdon:2.72403714)'396':0.23624286,((Apomys_hylocoetes:0.07936000,Apomys_insignis:0.07936000)'394':0.90277000,Apomys_camiguinensis:0.98213000)'388':1.97815000)'408':1.26513333)'415':2.88264889)'414':5.26368323)'420':0.65857026,Chiropodomys_gliroides:13.03031571)'419':4.28050929,Hapalomys_delacouri:17.31082500)'413':0.70733089,((((((((((((((((Arvicanthis_somalicus:0.13923000,Arvicanthis_neumanni:0.13923000)'412':3.18912714,Arvicanthis_niloticus:3.32835714)'425':3.75261286,Arvicanthis_abyssinicus:7.08097000)'411':1.06435000,(Arvicanthis_ansorgei:8.14532000,Arvicanthis_nairobae:8.14532000)'407':0.00000000)'405':1.15468000,(((Lemniscomys_bellieri:7.38569000,Lemniscomys_macculus:7.38569000)'387':1.69639500,Lemniscomys_rosalia:9.08208500)'431':0.21791500,(((Lemniscomys_barbarus:4.65032500,Lemniscomys_zebra:4.65032500)'429':0.66621667,Lemniscomys_striatus:5.31654167)'436':3.98345833,Lemniscomys_griselda:9.30000000)'434':0.00000000)'428':0.00000000)'386':0.05337625,((Pelomys_campanae:8.59000000,Pelomys_fallax:8.59000000)'382':0.00000000,Mylomys_dybowskii:8.59000000)'380':0.76337625)'374':0.41720625,(Rhabdomys_dilectus:4.72781333,Rhabdomys_pumilio:4.72781333)'441':5.04276917)'439':0.73411321,Desmomys_harringtoni:10.50469571)'373':0.34887984,(Dasymys_incomtus:4.37696500,Dasymys_rufulus:4.37696500)'291':6.47661056)'457':0.35124778,((Grammomys_cometes:11.00000000,(((Grammomys_dolichurus:2.51297333,Grammomys_surdaster:2.51297333)'463':3.11889000,Grammomys_macmillani:5.63186333)'466':0.89584067,Grammomys_ibeanus:6.52770400)'462':4.47229600)'461':0.00000000,Micaelamys_namaquensis:11.00000000)'471':0.20482333)'470':0.29645111,(Stochomys_longicaudatus:10.09854375,(Hybomys_lunaris:7.36592000,Hybomys_univittatus:7.36592000)'460':2.73262375)'456':1.40273069)'477':0.89872556,(Golunda_ellioti:12.40000000,Oenomys_hypoxanthus:12.40000000)'475':0.00000000)'455':0.38083556,(((((((((Otomys_lacustris:3.46319000,Otomys_anchietae:3.46319000)'481':0.92376167,Otomys_denti:4.38695167)'480':1.01090262,Otomys_angoniensis:5.39785429)'454':2.27455905,Otomys_barbouri:7.67241333)'487':0.00000000,((((((Otomys_simiensis:1.50000000,Otomys_jacksoni:1.50000000)'485':3.44574667,Otomys_orestes:4.94574667)'453':0.03425333,Otomys_dartmouthi:4.98000000)'452':0.17908000,Otomys_typus:5.15908000)'451':1.96666667,(Otomys_burtoni:4.20241333,Otomys_tropicalis:4.20241333)'495':2.92333333)'493':0.18333333,(Otomys_occidentalis:6.95018000,Otomys_irroratus:6.95018000)'492':0.35890000)'450':0.36333333)'503':0.41758667,((Otomys_laminatus:7.73049500,Otomys_saundersiae:7.73049500)'501':0.35950500,Otomys_maximus:8.09000000)'507':0.00000000)'511':0.00000000,(Parotomys_brantsii:7.29748333,Parotomys_littledalei:7.29748333)'510':0.79251667)'506':0.85000000,(Myotomys_unisulcatus:8.94000000,Myotomys_sloggetti:8.94000000)'500':0.00000000)'499':3.84083556)'449':1.05570667,(Vandeleuria_sp._KCR-2008:10.62781500,(Aethomys_kaiseri:10.60027000,(Aethomys_ineptus:2.65035000,Aethomys_chrysophilus:2.65035000)'520':7.94992000)'518':0.02754500)'523':3.20872722)'517':1.21716778,((Millardia_meltada:5.93057143,Millardia_kathleenae:5.93057143)'528':5.62638607,Cremnomys_cutchicus:11.55695750)'535':3.49675250)'541':0.64487974,(((((Tokudaia_osimensis:9.14472667,Tokudaia_muenninki:9.14472667)'540':3.25527333,(Apodemus_gurkha:12.40000000,(((((((((Apodemus_ilex:2.68500000,Apodemus_draco:2.68500000)'539':1.38932500,Apodemus_semotus:4.07432500)'538':2.22191500,Apodemus_latronum:6.29624000)'534':1.74308571,Apodemus_peninsulae:8.03932571)'533':0.83637286,Apodemus_speciosus:8.87569857)'531':0.64348893,(Apodemus_agrarius:2.11075167,Apodemus_chevrieri:2.11075167)'527':7.40843583)'550':2.88081250,Apodemus_epimelas:12.40000000)'548':0.00000000,((((Apodemus_iconicus:1.45000000,Apodemus_hermonensis:1.45000000)'52

6':0.45000000,Apodemus_wardi:1.90000000)'516':4.23000000,Apodemus_hyrcanicus:6.13000000)'448':0.78000000,(Apodemus_mystacinus:6.91000000,((((Apodemus_pallipes:2.91397750,Apodemus_uralensis:2.91397750)'555':1.18522917,((Apodemus_flavicollis:2.19317400,Apodemus_ponticus:2.19317400)'553':0.93175100,Apodemus_sylvaticus:3.12492500)'447':0.97428167)'445':0.12784000,Apodemus_alpicola:4.2270466 7)'566':0.91975667,Apodemus_witherbyi:5.14680333)'564':1.76319667)'570':0.00000000)'573':5.49000000)'569':0.00000000,Apodemus_argenteus:12.40000000)'563':0.00000000)'576':0.00000000)'562':2.10000000,(Malacomys_cansdalei:6.94550667,(Malacomys_longipes:5.00264100,Malacomys_edwardsi:5.00264100)'587':1.94286567)'586':7.55449333)'591':0.00000000,(((Mus_tenellus:9.52000000,Mus_sorella:9.52000000)'585':0.00000000,(((Mus_bufo:4.76078500,Mus_setulosus:4.76078500)'583':0.73234100,(((Mus_minutoides:3.46872000,Mus_musculoides:3.46872000)'599':0.57848200,Mus_indutus:4.04720200)'598':0.47626467,(Mus_mattheyi:3.83495000,Mus_haussa:3.83495000)'597':0.68851667)'596':0.96965933)'594':4.02687400,Mus_baoulei:9.52000000)'582':0.00000000)'610':0.00000000,((((Mus_lepidoides:3.50155000,Mus_nitidulus:3.50155000)'608':4.38845000,(Mus_triton:7.89000000,((((Mus_booduga:3.15173500,Mus_fragilicauda:3.15173500)'613':0.00883167,Mus_terricolor:3.16056667)'607':3.20943333,(((Mus_spicilegus:2.82642500,(Mus_musculus:2.54224000,Mus_macedonicus:2.54224000)'620':0.28418500)'619':0.23905722,Mus_spretus:3.06548222)'618':3.30451778,Mus_famulus:6.37000000)'617':0.00000000)'626':1.04259800,((Mus_cookii:2.74136143,Mus_cervicolor:2.74136143)'616':1.37831857,Mus_caroli:4.11968000)'606':3.29291800)'630':0.47740200)'635':0.00000000)'639':0.19934600,(Mus_shortridgei:5.85537500,((Mus_platythrix:0.30000000,Mus_cf._saxicola:0.30000000)'638':0.00193333,Mus_saxicola:0.30193333)'634':5.55344167)'643':2.23397100)'633':0.20148036,(Mus_pahari:3.70023750,Mus_crociduroides:3.70023750)'629':4.59058886)'605':1.22917364)'581':4.98000000)'646':0.08907489,(((((((Stenocephalemys_albipes:6.94858667,Stenocephalemys_griseicauda:6.94858667)'580':0.30545000,Stenocephalemys_albocaudata:7.25403667)'651':1.27596333,(Myomyscus_verreauxii:8.53000000,(Myomyscus_brockmani:7.55063286,Myomyscus_yemeni:7.55063286)'649':0.97936714)'579':0.00000000)'561':0.78558800,(Colomys_goslingi:6.34589444,Zelotomys_hildegardeae:6.34589444)'559':2.96969356)'671':1.78441200,(Mastomys_huberti:11.10000000,(((Mastomys_coucha:7.49901000,((((Mastomys_hildebrandtii:0.66403667,Mastomys_erythroleucus:0.66403667)'675':1.73014833,Mastomys_awashensis:2.39418500)'674':1.31746750,Mastomys_natalensis:3.71165250)'670':1.60864036,Mastomys_kollmannspergeri:5.32029286)'669':2.17871714)'668':1.02500000,Mastomys_verheyeni:8.52401000)'667':0.84632375,Mastomys_pernanus:9.37033375)'666':1.72966625)'665':0.00000000)'664':0.70000000,(((((Praomys_misonnei:2.90198455,Praomys_tullbergi:2.90198455)'663':5.29178323,(Praomys_daltoni:1.29933000,Praomys_derooi:1.29933000)'662':6.89443778)'661':0.31795222,(Praomys_degraaffi:3.87937667,Praomys_jacksoni:3.87937667)'660':4.63234333)'659':1.67686143,(Praomys_verschureni:7.47964000,Praomys_lukolelae:7.47964000)'658':2.70894143)'693':1.61141857,(((Praomys_mutoni:11.75904500,Praomys_minor:11.75904500)'698':0.00000000,((((Praomys_obscurus:0.74542000,Praomys_hartwigi:0.74542000)'697':1.79914000,Praomys_petteri:2.54456000)'696':2.86948500,(Praomys_sp._ADM-2011:2.08000000,Praomys_morio:2.08000000)'692':3.33404500)'703':2.16297750,Praomys_rostratus:7.57702250)'691':4.18202250)'657':0.04095500,Praomys_delectorum:11.80000000)'655':0.00000000)'713':0.00000000)'719':0.00000000,(Heimyscus_fumosus:9.30851875,((((Hylomyscus_alleni:3.71760500,Hylomyscus_parvus:3.71760500)'718':0.12975875,Hylomyscus_stella:3.84736375)'717':1.55263625,((Hylomyscus_kerbispeterhansi:0.83000000,Hylomyscus_anselli:0.83000000)'716':1.01000000,Hylomyscus_arcimontensis:1.84000000)'712':3.56000000)'711':0.99259667,((((Hylomyscus_denniae:2.36000000,Hylomyscus_endorobae:2.36000000)'729':2.56510000,Hylomyscus_aeta:4.92510000)'728':0.56172000,Hylomyscus_baeri:5.48682000)'726':0.48682000,Hylomyscus_grandis:5.97364000)'710':0.41895667)'735':2.91592208)'733':2.49148125)'709':2.78907489)'739':1.10951485)'744':2.31956614)'743':2.86926152)'751':0.16848812,((((Batomys_granti:4.71510000,Crateromys_heaneyi:4.71510000)'754':0.45771667,Batomys_salomonseni:5.17281667)'750':4.07552905,Carpomys_phaeurus:9.24834571)'757':3.85787329,(Phloeomys_sp._RMA-2005:1.37824000,Phloeomys_cumingi:1.37824000)'749':11.72797900)'748':7.94968652)'742':0.94409448,((((Thallomys_loringi:6.68265000,(Thallomys_nigricauda:5.73713500,Thallomys_paedulcus:5.73713500)'738':0.94551500)'708':6.02030000,Lamottemys_okuensis:12.70295000)'767':8.34852500,Dephomys_defua:21.05147500)'766':0.94852500,((Coccymys_ruemmleri:19.11996000,Pogonomelomys_mayeri:19.11996000)'765':0.00000000,Crossomys_moncktoni:19.11996000)'772':2.88004000)'764':0.00000000)'775':0.00000000,(Malpaisomys_insularis:12.84282000,Muridae_sp._SAJ-2009a:12.84282000)'763':9.15718000)'781':6.57552087,((((((((((Acomys_wilsoni:7.02248667,Acomys_russatus:7.02248667)'780':0.68801083,(((((Acomys_chudeaui:2.04415250,Acomys_cahirinus:2.04415250)'778':0.80290750,((Acomys_minous:1.92818000,Acomys_nesiotes:1.92818000)'762':0.01558667,Acomys_cilicicus:1.94376667)'787':0.90329333)'786':0.77823250,Acomys_dimidiatus:3.62529250)'794':0.15157750,Acomys_johannis:3.77687000)'793':2.40738750,Acomys_ignitus:6.18425750)'791':1.52624000)'785':2.19478583,Acomys_percivali:9.90528333)'761':0.62272667,(Acomys_ngurui:3.96474000,Acomys_spinosissimus:3.96474000)'707':6.56327000)'798':0.43973667,Acomys_subspinosus:10.96774667)'706':0.78210333,(Acomys_cineraceus:11.74985000,Acomys_kempi:11.74985000)'654':0.00000000)'558':5.76102704,Deomys_ferrugineus:17.51087704)'444':1.63589963,(Uranomys_ruddi:13.91373429,(Lophuromys_woosnami:8.47892333,((((Lophuromys_melanonyx:3.20811667,Lophuromys_brevicaudus:3.20811667)'290':0.15814667,(Lophuromys_zena:0.45839250,Lophuromys_flavopunctatus:0.45839250)'805':2.90787083)'808':0.64525667,Lophuromys_chrysopus:4.01152000)'804':0.98704714,Lophuromys_sikapusi:4.99856714)'802':3.48035619)'813':5.43481095)'811':5.23304238)'801':5.74976939,((Desmodilliscus_braueri:16.18375600,Pachyuromys_duprasi:16.18375600)'289':2.17638400,((((((Dipodillus_dasyurus:10.40000000,((Dipodillus_campestris:0.51028500,Dipodillus_rupicola:0.51028500)'820':7.93607100,Dipodillus_simoni:8.44635600)'825':1.95364400)'823':0.00000000,((Gerbillus_poecilops:6.85868000,Gerbillus_henleyi:6.85868000)'819':3.34132000,((((((Gerbillus_sp._LG-2011:0.07704800,Gerbillus_perpallidus:0.07704800)'829':6.90567260,Gerbillus_pyramidum:6.98272060)'828':0.78098940,Gerbillus_occiduus:7.76371000)'818':0.24057600,(Gerbillus_hesperinus:0.59399000,Gerbillus_hoogstrali:0.59399000)'816':7.41029600)'288':1.35660800,((Gerbillus_nigeriae:2.33899500,Gerbillus_andersoni:2.33899500)'172':3.63437500,Gerbillus_nancillus:5.97337000)'98':3.38752400)'4':0.21595100,(((Gerbillus_latastei:6.07115667,Gerbillus_tarabuli:6.07115667)'2':1.22926000,Gerbillus_gerbillus:7.30041667)'841':0.11784889,Gerbillus_nanus:7.41826556)'839':2.15857944)'849':0.62315500)'852':0.20000000)'848':1.59566000,((Taterillus_petteri:6.92899500,Taterillus_emini:6.92899500)'846':3.64447500,((Taterillus_pygargus:4.22092000,Taterillus_arenarius:4.22092000)'844':3.01799250,(Taterillus_gracilis:2.42000000,Taterillus_sp._PC2002:2.42000000)'838':4.81891250)'861':3.33455750)'859':1.42219000)'866':0.20974222,(Meriones_tamariscinus:11.98771000,((((Meriones_meridianus:4.08925800,Meriones_chengi:4.08925800)'864':3.47745950,((Meriones_shawi:3.58297714,Meriones_unguiculatus:3.58297714)'858':0.08702286,Meriones_sp._Garat_An_Njila:3.67000000)'871':3.89671750)'869':1.31400917,Meriones_tristrami:8.88072667)'857':1.49137333,(((Meriones_persicus:4.88686000,Meriones_rex:4.88686000)'855':1.00014750,Meriones_crassus:5.88700750)'837':1.40575500,Meriones_libycus:7.29276250)'835':3.07933750)'880':1.61561000)'878':0.21769222)'885':2.39954206,(((Brachiones_przewalskii:10.65085000,Rhombomys_opimus:10.65085000)'883':0.30994000,(Psammomys_vexillaris:4.97610500,Psammomys_obesus:4.97610500)'877':5.98468500)'876':3.62052800,Sekeetamys_calurus:14.58131800)'874':0.02362629)'834':2.41440286,(((((((Gerbilliscus_gambianus:3.02564000,Gerbilliscus_kempi:3.02564000)'901':4.29568200,Gerbilliscus_guineae:7.32132200)'900':0.68725800,((Gerbilliscus_brantsii:5.53821600,Gerbilliscus_afra:5.53821600)'907':0.60483000,(Gerbilliscus_validus:4.41885000,Gerbilliscus_leucogaster:4.41885000)'910':1.72419600)'906':1.86553400)'915':2.99142000,(((Gerbilliscus_robustus:2.16349667,Gerbilliscus_vicinus:2.16349667)'918':6.30968333,Gerbilliscus_phillipsi:8.

47318000)'914':1.03673833,Gerbilliscus_nigricaudus:9.50991833)'913':1.49008167)'905':0.00000000,((Gerbillurus_vallinus:7.71000000,(Gerbillurus_tytonis:5.64069000,Gerbillurus_setzeri:5.64069000)'899':2.06931000)'925':0.00000000,Gerbillurus_paeba:7.71000000)'924':3.29000000)'922':2.26709125,Desmodillus_auricularis:13.26709125)'898':1.09796589,(Tatera_sp._KIK1704:7.32000000,Tatera_sp._KE102:7.32000000)'896':7.04505714)'935':2.65429000)'934':1.34079286)'933':6.53640606)'942':2.07380823,Lophiomys_imhausi:26.97035429)'941':1.60516658);

Text file S3.2

Constructed coding sequences (based on the longest uninterrupted ORF) for *M. leucogaster* restriction factors that were used as input query sequences in DGINN pipeline.

>M. leucogaster_CDS_APOBEC3

ATGCAACCCCAGGGTCTGGGGCCCAACGCTGGGATGGGACCAGTGTGCCTGGGATGCAGCCATCGCAGACCCTATTCACCGATCAGAAACCCGCTAAAGAAGTTA
TATCAACAAACATTCTACTTTCATTTTAAGAACGTACGCTATGCCTGGGGTCGAAAGAATAACTTCTTGTGCTATGAAGTGAATGGGATGGACTGCGCTTTACCTGTCC
CCCTTCGCCAAGGGGTCTTCAGGAAACAGGGCCACATCCACGCCGAACTCTGCTTCATATACTGGTTCCACGACAAAGTCCTGAGAGTGCTGTCCCCGATGGAAGA
GTTCAAGGTCACGTGGTACATGTCCTGGAGCCCCTGCAGCAAGTGCGCGGAGCAGGTAGCCAGGTTCCTGGCCGCACACCGCAACCTAAGCCTGGCCATCTTCAG
CTCCCGCTTGTACTACTACTTAAGGAACCCGAACTACCAGCAGAAGCTCTGCAGGCTGATTCAGGAAGGAGTCCACGTGGCTGCCATGGACCTACCAGAATTTAAA
AAGTGTTGGAACAAGTTTGTGGACAATGACGGCCAACCATTCAGGCCTTGGATGAGACTGAGAATAAATTTTAGTTTCTATGATTGCAAGCTTCAGGAGATTTTCAG
CCGAATGAATCTGCTAAGGGAAGATGTATTTTACTTGCAATTTAACAACAGCCACCGGGTCAAGCCAGTCCAGAATCGCTACTATCGCAGGAAGTCCTATCTGTGCTA
CCAACTGGAGCGGGCCAATGGCCAAGAGCCACTCAAAGGCTACCTGCTATACAAGAAAGGTGAACAGCATGTAGAAATCCTCTTCCTTGAGAAGATGCGGTCCAT
GGAGCTGAGCCAAGTGCGAATTACCTGCTACCTCACCTGGAGCCCCTGCCCAAACTGTGCCCGGCAACTCGCTGCATTCAAAAAGGATCACCCAGACCTAATTCTG
CGGATCTATACCTCCCGCCTGTATTTCTACTGGAGGAAGAAGTTCCAGAAGGGGCTGTGTACTCTGTGGCGATCAGGGATCCACGTGGACGTCATGGACCTCCCTCA
GTTTACTGACTGCTGGACAAACTTTGTGAACTCACAAAGGCCATTTAGGCCATGGAATGAACTGAAGAAAAACAGCTGGCGCATACAAAGGCGGCTTCAGAGGAT
CAAGGAGTCCTGGGGCCTG

>M. leucogaster_CDS_BST2

ATGGCACCCTCTTTCTACCACTATCTGCCCGTGGCGATGGACGAGAGGTGGGAGCCAAAAGGATGGAGCATCCGCCGGTGGTGGCTGGTGGCCGCAATCTTGGTG
GTCCTGATCGGGGTTGTCTTAGTCTGCCTGATAGTCTACTTCGCCAACGCAGCGCACAGCGAGGCCTGTAAGAACGGGTTGCGGTTGCAGGATGAGTGCCGAAAC
ACCACGCACCTGTTGAAGCACCAGCTCACCCGCGCCCAGGACAGCCTGCTGCAGACGGAGATGCAGGCAAACTCCTGCAACCAGACCGTGATGGACCTTCGGGA
TTCCCTGAAGAAGAAGGTGTCTCAAACCCAGGAGCAGCAGGCCCGCATCAAGGAACTTGAGNNNNNNAGGACCCAAAAGGAAATTTCTACCACAGTGCAGGTG
AACTCAGGCGGCTCCGTGGTGGTCTCCAGCCTACTGGTGCTTGTGGCGGTACTGTTCCTGCACTTT

>M. leucogaster_CDS_SAMHD1

ATGCAGAGACCAGACTCGGAGCAACCTGCTAAGCGGTCCCGCTGCGATGGCAGCCCAAGGACGCCACCGAGCACCCGTCCTGCAGCAGCGAACGAGTCTGCAGA
TCCCGAAAGCAGCGACCTCCGAACCTGGGGTCCCGAGGACGTGTGCTCCTTCTTAGAGAAACGTGGTTTCCGAGAGAAAAAAGTGCTGGACATCTTCAGAGAAA
ATAAAATCACTGGTTCATTTCTGCCCTTTTTGGATGAGAATCGTCTTGAGAATCTGGGAGTAAGTTCCTTGGAGCAGAGGAAGAAGATGATAGAATGTATCCAGAAG
CTGAATCAGTCTCGGATTGATCTAATGAAGGTATTTAACGATCCTATTCATGGCCATATTGAATTCCACCCCCTCCTCATCCGAATCATTGACACACCTCAGTTCCAGCG
GCTTCGCTACATCAAGCAGCTGGGAGGCGGTTACTACGTCTTCCCTGGCGCGTCACACAATCGGTTTGAACACAGTCTCGGAGTGGGGTACCTAGCAGGCTGCCTC
GTGCGAGCGCTTGCCGAAAAACAGCCAGAGCTGCAGATCAGTGAACGAGATATGCTCTGTGTTCAGATCGCGGGGCTCTGCCATGACCTTGGTCATGGGCCATTCT
CCCATATGTTTGATGGAAGATTTATTCCACTTGCTCGCCCAGACATAAAGTGGAGGCATGAACAGGGCTCAGTTGAGATGTTTGAGCATCTGGTTAACTCCAATGAA
CTCAAACTTGTCATGAAGAACTATGGTCTCGTCCCTGAAGAAGACATTACCTTTATCAAGGAACAAATTATGGGACCACCTGTATCACCAGTCAAAGACTGCTTGTGG
CCGTATAAGGGGCGCCCCGCCAAGAAGAGCTTCCTGTACGAGATAGTGGCTAACAAAAGAAATGGCATTGATGTGGACAAATGGGACTATTTTGCCAGAGACTGTC
ACCATCTTGGAATCCAAAATAATTTTGATTACAAGCGCTTCATTAAGTTTGCCCGTATCTGTGAAGTGGACGACGAGACCCGGGCACATAAGGTGAAGCACATTTGT
ACCAGAGAAAAGGAGGTTGGAAATCTGTATGACATGTTTCACACACGGAACTGCTTACACCGAAGAGCTTATCAGCACAAAATCGGCAACCTCATCGATATCATGAT
TACCGAAGCTTTCCTCAAAGCAGACCCCCACGTGGAGATTACCGGGACCGAAGGGAAGAAGTTTCGAATTTCCACAGCCATTCATGACATGGAAGCCTTCAGTAA
GCTGACAGATAACATCTTTCTGGAGATTTTATACTCCACGGATCCACAGTTGTCTGAGGCCCGGAATATTTTAAGGAACATTGAATGCCGTAATCTGTACAAGTATTTG
GGTGAGACCCAGCCGAAGCGTGAGAAGATTAAAAAGGCAGAGTATGACAAGCTTGCCCAAGAAGTTGCTAATGCCAAACCTGAAATTGCCCCGGATGTTGAACTA
AAGGCTGAAGATTTCATCGTTGATGTTATCAATATGGATTACGGGATGGAAGACAAGAACCCAATTGATAATGTTCACTTCTATTGTAAGAGTGACAGCAGGCAAGC
GGTCACGATCACTAAAGACCAGGTGTCACAGCTGCTGCCGGAGAAATTTGCAGAGCAGCTGATTCGAGTGTACTGTAAGAAGAAAGAC

>M. leucogaster_CDS_TRIM5alpha

ATGGCTTCAGAATTCGTGATGAATTTAAAAGAGGAGGTGACCTGTCCTATCTGCCTGGACCTGATGGTAGAACCTGTGAGTGGAGATTGTGGTCACAGCTTCTGCCA
AGCCTGCATCACGCTGAACTATGAATCCAGCAAATGCAATCAGGATGAGTTCATTTGCCCTGTGTGCCGAGTTAGTTACCTGTTTAAGAACCTGAGGCCCAATCGAC
ATGTGGCCAACATAGTGCAGAGGCTCAAAGAGTTCAAGTCCAGCCCAGAAGAGGAGCCGAAGGTGCTTTCTTGTGCAAGGCATGGAGAGAAACTCCAGCTCTTC
TGTAAGAAGGACATGATGCCCATCTGCTGGCTTTGTGAGCGATCTCAGGAGCACCGTGGACACCAAACAGTTCTCATTGAAGAGGTGGTCCAGGAGTATAAGGAG
AAGCTGCAGGCAGCTCTGGAAAAGCTGATGGCAGACAAGAAAGAATTTGAGAACTGGAATGATGAACTTCAAAAGGAGAGAACTTACTGGGAGAATCAAATACA
GAAAGATGTGGAAAATGTTCAGTCAGAGTTTCAACGAATGAGGGGTATCATGGACTCTGAGGAGAAGAATGAGTTGCAGAAGCTGATGCAAGAGAAGGAGGGC

GTCATCAACAGCCTGGCCGGGTCTGAGAATGAGCATGCTCAGCAGAGCAAGTTGCTAGGAGACCTCATCTTAAGTGTGGAACATCAGTTACAGTGCTCAGCCATGG
AAATGCTGCAGGGGAGTCAGACTTTTTCAATGAGGAAGCCCAAAACCATCCCCAGGGAACAAAGAAGAGTGTTCCGAGCCCCTGATCTGCAAGACATGCTGCAAA
TGTTGCAAGTTCAAGTGACGCTGGTTGAAAGCAACAATCCAAACATTTTCATTACCGCCGACAAAAGACAGATACGATATGAAGACCACCAAGCAAGACATTTTGC
CCGTCCGACTGAAAACTGTCATGCAGGTGTCCTGGGATACCCAGCTATCCAATCAGGAAAACACTACTGGGAAGTAGATGTGTCTGGAAAAGGTTCTTGGGTTCTG
GGATTAAGTGATGGAAGCTACCTCTTTAATCCAATATTTCGTTCAAATGCAGAAAGACCCCCAAACCCCCCCTTATTTCGTTTGAATCTTAGTAATGATTCACATTCTCG
TTTGAGTCTTAGTAATGATTCACATTATCAACCTAAATATGGCTTCTGGGTTATAGGGCTGTGGGGAAATTCTGTGTATAATGCTTTTGAGGAGTGTACGTTCACAGGC
AAGCCCAGTGTGTTGACCCTCTCTCTGATGGTCCGACCCTGTCGTGTCGGTATTTTCCTCGACTGTGCAGCTGGCACCCTCTCGTTTTACAATATTTCCAACCATGGCA
CTCTTATCTACAGATTCTGTGCAGGTTCCTTTCCTGATAGGGTTTTTCCATATTTTAACCCCATGGGAAGTTCAGAGCCATTGACAATATGCTGGCCAGACTCT

>M. leucogaster_CDS_ZAP

ATGGCAGATCCCGGGGTATGCTGTTTCATCACCAAGATCCTGTGCGCCCACGGGGGCCGTATGACCCTGGAGGAACTGCTGGGTGAGATCAGGCTCCCCGAGGCG
CAGCTCTACGAGCTGCTGGAGACGGCGGGGCCCGATCGCTTCGTGCTATTGGAGACTGGAGGCCAGGCCGGGATCACTCGGTCTGTAGTGGCTACTACTCGAGCC
CGCGTCTGCCGTCGGAAGTACTGCCAGAGACCCTGCGACAGCCTGCACCTCTGCAAGCTTAATCTGCTCGGCCGGTGCCACTATGCACAGTCTCAGCGGAACCTCT
GCAAATACTCTCACGAGGTACTCTCGGAACAGAATTTCCAGGTCCTGAAGAATCATGAGCTCTCTGGGCTTAATCAAGAGGAGCTGGCGGTCCTCCTGATCCAAAA
CGACCCTTTTTTCATGCCTGAGATATGCAAGAGTTACAAAGGAGAGGGCCGAAAACAGACCTGCGGGCAGCCACAGCCATGCGAGAGACTCCACATCTGTGAGCA
CTTCACCCGGGGCAACTGCAGTTACCTCAACTGTCTCAGGTCTCACAACCTGATGGACAGAAAGGTGTTGACCATCATGAGGGAGCACGGGCTGAGTCCTGATGT
GGTCCAGAACATCCAGGACATCTGCAACAACAAACACGCCAGGAGGAACCCGCCTGGCACGAGAGCTGCCCATCCACACCGCAGAGGCGGCGCACACAGAGAC
AGAAGCAAAAGCAGAGACCGCTTCCTTCACAACAGTCTAGAATTTCTCTCACCTGTTGTCTCACCTCTGGGATCTGGTCCGCCTAGCCCAGATGTCACCAGCTGTAA
AGATTCCCTGGAGGATGTGTCTGTGGATGTCACCCAGAAGTTCAAGTACTTGGGGACGCACGACCGTGCGCAGCTCTCCCCAGTCTCATCTAAGGCTGCTGGTGTT
CAAGGACCCAGTCAAATGAGAGCAAGCCAGGAGTTTTCAGAGGATGGGAATCTAGATGACATATTTTCTAGGAATCGTTCTGATTCATCATCAAGTCGAGCCTCCGC
TGCCAAGGTGGCACAAAGAAATGAAGCTGTGGCCATGAAAATGGGCATGGAGGTCAAGGGCAAGAAGGAGGCTCCAGACATCGATCGGGTCCCATTTTTAAATA
GTTATATTGATGGGGTGACCATGGAAAAAGCATCGGTCTCAGGAATTCCAGGCAAAAAGTTCACAGCCAATGATCTGGAAAATTTGCTATTACTTAACGACACTTGG
AAGAATGTGGCTAAGCCCCAGGATCTGCAGACCACAGGCAGAATCACTGACAGTGGCCAAGACAAGGCATTCCTGCAGAGTAAATATGGAGGAAACCCAGTGTG
GGCAAGTGCATCCACCCATAATGCCCCAAATGGCTCTAGTCAAATTATGGATGAAACTCCTAATGTCTCTAAAAGTAGTACCAGTGGTTTTGCCATAAAACCAGCAATT
GCTGGAGGAAAAGAAGCAGTCTATTCTGGAGTTCAGAGTCCGAGAAGCCAGGTCCTAGCTATGCCTGGGGAGACTACTACCCCTGTACAGAGCAACAGGCTGCCT
CAGTCGCCTCTGTCTTCCTCAAGCCACAGAGCTGCAGCCTCTGGGAGCCCTGGCAAGAACTCCACCCATACCTCTGTGAGCCCAGCCATCGAGTCTTCAAGGATGA
CATCAGACCCCGATGAGTNTCTCCTACGCTACATCATAAGTCCTACTTCTCCAAAGATGGACAATCATGGCCCGAAGGAAATCTGTCAGGACCATCTGTACAAGGGCT
GTCAGCAGAGCCACTGCGACAGGAGTCACTTCCATCTGCCCTACCGGTGGCAGATGTTCGTATATACCACTTGGAGGGACTTCCAGGACATGGAGTCTATCGAACA
GGCCTATTGTGATCCCCACGTTGAACTCATTTTGATAGAAAACCATCAGATCAATTTCCAGAAAATGACCTGTGACTCCTACCCCATCCGACGCCTCTCCACTCCCTCA
TATGAGGAAAAGCCACTTAGTGCTGTCTTCGCCACCAAGTGGATTTGGTATTGGAAGAATGAATTTAATGAATATATCCAGTATGGGAATGAGAGCCCAGGCCACGC
CAACTCCAACGTCAGCTCTAGGTACCTGGAGTCTTTCTTCCAGTCCTGTCCCAGGGGAGTTTTGCCATTCCAGGCCGGTTCACAGAAGTACGAGTTAAGCTTTCAN
GGGATGATTCAGACAAATATAGCTTCCAAGACTCAAAGGCATGTTGTCAGAAGGCCAGTCTTTGTTTCTTCGAAGGATGTGGAGCAGAAGAGAAGAGGTCTAGAT
CATCAGCCAGTGACGCCCAAGGCAGATGCTCTGAATTCTATTCCCCAGAAGAATGCTAGCACTGTTTTGCCCAACGAATATGAGTTTCTAGAACTCAATGACCAGGA
TGTGGAGTATGTCACAGTAAGCGAACAGTTTAAAGCATCCATGAAACCTTTCAAGATTGTGAGAATAAAGCGGATATGGAACCAAAAACTCTGGGACGCTTTTGAA
AGGAAGAAGCAAAAGATGACAAATAAAAACGAGATGCTCTTGTTTCACGTGGCGAGCCGTGCTCACGTGGATTACATCTGTAAGAATAATTTCGAGTGGATCCTAC
ATGGAAATCGGGACACCAGATACGGAAAAGGAAATTATTTTGCAAAAGAAGCCATCTATTCCCACAAGAATTGTTCATATGATATCAGAAACATTGTTATGTTCGTAG
CCCGAGTCCTGGTTGGAAACGTCATTGAAGGGAATATGACGTTCAGTAGCCCTCCTGCACTCTATGACAGCTGCGTGGACACCAGGCTGAATCCCTCCGTCTTTGTC
ATTTTCCGGAAAGAACAGATTTACCCAGAGTATCTGATTGAGTATGTGGAATCAGAGAAAGAGAAAGGTTGCATAATTAGT