High throughput approaches to studying virology: applications to the koala retrovirus KoRV and gibbon ape leukemia virus GALV

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Niccolò Alfano

from Milano, Italy

2016

This Dissertation was done in the Leibniz-Institute for Zoo and Wildlife Research in Berlin during the period 1/09/2012 - 30/06/2016 under the supervision of Prof. Alex D. Greenwood PhD and it is submitted to the Department of Biology, Chemistry and Pharmacy of Freie Universität Berlin

1st Reviewer: Prof. Alex D. Greenwood Ph.D
2nd Reviewer: Prof. Dr. Heribert Hofer D.phil

Date of defense: 20 September 2016.

This thesis is based on the following manuscripts:

1. Alfano N, Courtiol A, Vielgrader H, Timms P, Roca AL, Greenwood AD. 2015. Variation in koala microbiomes within and between individuals: effect of body region and captivity status. *Scientifc Reports*, **5**: 10189. http://dx.doi.org/10.1038/srep10189.

2. Alfano N, Kolokotronis SO, Tsangaras K, Roca AL, Xu W, Eiden MV, Greenwood AD. 2015. Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses. *Journal of virology*, **90**:1757-1772. http://dx.doi.org/10.1128/JVI.02745-15.

3. Alfano N, Michaux J, Morand S, Aplin K, Tsangaras K, Löber U, Fabre PH, Fitriana Y, Semiadi G, Ishida Y, Helgen KM, Roca AL, Eiden MV, Greenwood AD. 2016. An endogenous gibbon ape leukemia virus (GALV) identified in a rodent (*Melomys burtoni* subsp.) from Wallacea. *Journal of virology* (In review). http://dx.doi.org/10.1128/JVI.00723-16.

Contents

Zusammenfassung

Retroviren sind behüllte Viren, die in der Lage sind ihr RNA-Genom revers zu transkribieren, die erhaltene DNA in ein Wirtszellgenom zu integrieren und die Wirtszellmaschinerie zu verwenden um ihre retroviralen Gene zu exprimieren und schließlich neue Viruspartikel zu produzieren. Wenn ein infektiöser (exogener) Retrovirus eine Keimbahnzelle infiziert, kann das integrierte retrovirale Genom von den Eltern auf die Nachkommen übertragen werden und über Generationen hinweg weitervererbt werden, was dann zu einem endogenen Retrovirus (ERV) führt. Retroviren sind umfassend erforscht, da sie wichtige Pathogene von Wirbeltieren, einschließlich des Menschen (z.B. HIV), sind. Sie haben die Tendenz neue Arten zu infizieren mit der Gefahr, die Krankheit auch bei diesem neuen Wirt auszulösen (z.B. Aids), und sie sind wertvolle biomedizinische Werkzeuge, die in der Genübertragung und Gentherapie angewandt werden. Außerdem haben ERVs das Genom von den meisten Wirbeltierarten kolonisiert, wobei Endogenisierung von Retroviren eine Schlüsselrolle in der Evolution der Genome von Wirbeltieren gespielt hat.

Der Koala Retrovirus (KoRV) ist der einzige bekannte Retrovirus, der zurzeit in die Keimzellen seiner Wirte eindringt. KoRV wurde auch wegen seiner Relevanz in der Arterhaltung von Koalas erforscht. Es wird vermutet, dass KoRV Leukämie, Lymphome und Immunsuppression in Koalas verursacht, was letztendlich zu einer höheren Anfälligkeit von Koalas gegenüber der weit verbreiteten Infektion mit Chlamydien führen könnte. Die kombinierten Folgen der Infektionen mit KoRV und Chlamydien könnten zur lokalen Ausrottung von Koalas führen. Das erste Ziel dieser Dissertation war es, die Wirkung von KoRV auf die Gesundheit von Koalas durch die Studie der mikrobiotischen Zusammensetzung (Mikrobiome) zu untersuchen.

In **Kapitel II** habe ich mittels Hochdurchsatzsequenzierung von 16S ribosomaler RNA-Amplikons das Mikrobiom des Auges, des Mundes, des Afters und des Kots von zwei in Gefangenschaft lebenden Koalas charakterisiert. Das Mikrobiom des Auges wurde untersucht, da Chlamydien-Infektionen in Koalas häufig die Augen betreffen und zu Konjunktivitis, anderen Augenkrankheiten und im schlimmsten Fall zu Blindheit führen können. Ich habe auch die Bakteriengemeinschaften analysiert, die mit der Verdauung in Verbindung stehen (Mikrobiom des Mundes, des Afters und des Kots) um zu bestimmen, ob hier Auffälligkeiten auf Grund der speziellen Nahrung von Koalas, die fast ausschließlich aus Eukalyptusblättern besteht, zu finden sind. Diese Studie hat demonstriert, dass Koala Mikrobiome in der Zusammensetzung den Mikrobiomen der gleichen Körperstellen anderer Säugetiere allgemein ähnlich waren, auch wenn das

Mikrobiom des Auges einige einzigartige Eigenschaften gezeigt hat. Weiterhin hat sie den Normalzustand der Mikrobiome in gesunden Koalas etabliert, zu welchen Mikrobiome von kranken Koalas verglichen werden können. Außerdem weist die Ähnlichkeit des fäkalen Mikrobioms der gefangenen Koalas aus der Studie mit dem Mikrobiom wilder Koalas darauf hin, dass Gefangenschaft die Gesundheit des Mikrobioms der Koalas kaum verändert.

Eine andere Besonderheit von KoRV ist seine sehr nahe Verwandtschaft zum Leukämievirus des Gibbons (GALV), ein Retrovirus, der Gibbons in Südostasien befällt. KoRV und GALV sind vermutlich das Ergebnis einer Übertragung über die Artengrenzen hinweg, die wahrscheinlich über einen bisher unbekannten Zwischenwirt stattgefunden hat. Das zweite Ziel dieser Dissertation war es, die Entwicklungsgeschichte von KoRV und GALV zu untersuchen, und den Zwischenwirt zu identifizieren, der an der artübergreifenden Übertragung zwischen Gibbon und Koala beteiligt war. Dafür war es zunächst notwendig, die Wissenslücke der Genetik der GALV-Stämme zu schließen. Im Gegensatz zu KoRV, welcher seit seiner Entdeckung umfassend charakterisiert worden ist, sind nur zwei der fünf isolierten GALV-Stämme bis heute sequenziert worden. Im **Kapitel III** habe ich eine Methode zur gezielten Anreicherung durch Hybridisierung und Sequenzierung im Hochdurchsatzverfahren angewendet, um die gesamte genomische Sequenz aller GALV-Stämme aus GALV-infizierten Zelllinien zu rekonstruieren. Die phylogenetischen Analysen haben gezeigt, dass die GALVs eine monophyletische Gruppe bilden, eingeschlossen der Virusstämme die aus Gibbons und Wollaffen (WMV) isoliert wurden. WMV ist vermutlich das Produkt eines horizontalen GALV-Transfers von Gibbon auf Wollaffe. Die GALV-WMV Klade war eine Schwestergruppe der Koalaretroviren (KoRVs). Hinweise auf positive Selektion wurden überall im Genom der pathogeneren Stämme der GALV und KoRV gefunden, vor allem in dem Gen, das die Virushülle kodiert, welches dem Immunsystem am meisten exponiert ist. Dies weist darauf hin, dass der Selektionsdruck vom Immunsystem des Wirtes ausgeht und so die Evolution der Retrovirenklade geprägt hat.

Im **Kapitel IV** habe ich die in Kapitel III gesammelten genetischen Informationen verwendet, um mögliche Zwischenwirte von GALV und KoRV und deren potentiellen Vorgängervirus zu identifizieren. GALV ähnliche Viren sind in mehreren südostasiatischen Nagetieren entdeckt worden, und ein vor kurzem durchgeführtes Screening australischer Wildtiere hat einen mit dem GALV in Australien verwandten Retrovirus in der murinen Art *Melomys burtoni* entdeckt. Daher wurde ein groß angelegtes Screening von südostasiatischen Nagetierarten auf KoRV und GALV ähnliche Sequenzen durchgeführt mit Hilfe der Anreicherung durch eine Hybridisierungsmethode und Hochdurchsatzsequenzierung. Ein ERV, der sehr eng mit WMV und MbRV verwandt ist,

wurde in einer neulich entdeckten Unterart von *Melomys burtoni* in Indonesien entdeckt. M. Burtoni kommt nicht in Südostasien vor, weswegen Gibbons durch einen anderen Wirt infiziert worden sein müssen, der bisher noch unbekannt ist. Dennoch ist der neu identifizierte Virus der mit GALV am nächsten verwandte Nicht-Primaten Retrovirus, der bisher isoliert wurde. Auch ist er der GALV, der am nächsten zum asiatischen Kontinent und der Verbreitung von Gibbons entdeckt wurde. Dies legt nahe, dass M. burtoni und vermutlich verwandte murine Arten eine wichtige Rolle bei der artübergreifenden Übertragung zwischen Koalas und Gibbons gespielt haben.

Summary

Retroviruses are enveloped viruses which are able to reverse transcribe their RNA genome, incorporate the resulting DNA into a host cell genome and use host cellular machinery to express their retroviral genes and ultimately produce new viral particles. When an infectious (exogenous) retrovirus infects a germ line cell, the integrated retroviral genome can be transmitted from parent to offspring and be inherited across generations, giving rise to an endogenous retrovirus (ERV). Retroviruses are widely studied since they are important pathogens of vertebrates, including humans (e.g. HIV), they have the tendency to infect new species, with the associated risk of causing pathologies in the new hosts (e.g. AIDS) and they are valuable biomedical tools applied in gene transfer and gene therapy. Furthermore, ERVs have colonized the genome of most vertebrate species, with retroviral endogenization having played a key role in the evolution of vertebrate genomes.

The koala retrovirus (KoRV) is the only known retrovirus which is currently in the process of invading the germ line of its host species. KoRV has also been studied for its relevance in koala conservation. KoRV is believed to induce leukemia, lymphomas and immunosuppression in koalas, which may eventually cause higher susceptibility to the highly prevalent *Chlamydia* infection in this species. The combined effects of KoRV and *Chlamydia* infection may lead to local extinctions of koalas. The first aim of this thesis was to evaluate the effect of KoRV on koala health through the study of koala microbial communities (microbiomes).

In **Chapter II**, I characterized by 16S ribosomal RNA amplicon high-throughput sequencing the ocular, oral, rectal and fecal microbiomes of two healthy captive koalas. The ocular microbiome was examined since *Chlamydia* infection frequently affects this body site in koalas causing keratoconjunctivitis, ocular diseases and eventually blindness. I also analysed the digestion-associated bacterial communities (microbiomes of mouth, rectum and feces) to determine whether such communities may be unusual in koalas, given their special diet based almost exclusively on *Eucalyptus* leaves. This study demonstrated that koala microbiomes were generally similar in composition to the microbiomes from the same body regions of other mammals, even if ocular communities showed some unique features, and established the healthy baseline for koala microbiomes to which microbiomes of diseased states can be compared. Furthermore, the similarity of the fecal microbiomes of the captive koalas from this study to those reported for wild koalas, suggests that captivity unlikely affects koala bacterial health.

Another particular aspect of KoRV is its closest relationship to the gibbon ape leukemia virus (GALV), a retrovirus which infects gibbons in Southeast Asia. KoRV and GALV are supposed to be the results of a cross-species transmission which likely occurred via intermediate as yet unknown hosts. The second aim of this thesis was to investigate the evolutionary history of KoRV and GALV, trying to identify intermediate hosts involved in the cross-species transmission between gibbons and koalas. As a preliminary step, it was necessary to fill the gap of genetic knowledge for the GALV strains. In contrast to KoRV which has been extensively characterized since its discovery, only two of the five GALV strains isolated so far have been sequenced. In **Chapter III**, I applied hybridization capture targeted enrichment and high-throughput sequencing to generate the complete genomic sequences of each GALV strain from GALV-infected cell lines. The phylogenetic analyses showed that the GALVs formed a monophyletic clade including the strains isolated from gibbons and the woolly monkey virus (WMV), which is likely the product of a horizontal transfer of GALV from a gibbon to a woolly monkey. The GALV-WMV clade was sister group of the koala retroviruses (KoRVs). Signs of positive selection were detected across the genome of the more pathogenic strains of GALV and KoRV, particularly in the *envelope* gene, the most exposed portion of the virus to host immune system, suggesting that host immune pressure is shaping the evolution of this retroviral clade.

In **Chapter IV**, I used the genetic information gathered in Chapter III to identify potential intermediate hosts of GALV and KoRV and, possibly, the ancestral virus from which the two viruses originated. GALV-like viruses have been discovered in several Southeast Asian rodent species and a recent screening of Australian wildlife has identified a retrovirus related to GALV in the Australian murid species *Melomys burtoni*. Therefore, a wide range of Southeast Asian rodent species were screened for the presence of KoRV and GALV-like sequences using hybridization capture and high-throughput sequencing. An ERV very closely related to WMV and MbRV was found in a newly discovered subspecies of *Melomys burtoni* from Indonesia. *M. burtoni* is not present in Southeast Asia, therefore gibbons must have been infected by another host which is still unknown. However, the newly identified virus is the most closely related non-primate retrovirus to GALV isolated to date and the most proximate record of GALV to the Asian continent and to the distribution of gibbons, suggesting that *M. burtoni*, and possibly related murine lineages, may have played an important role in the cross-species transmission between koalas and gibbons.

Chapter I

General Introduction

General Introduction

1.1 Retroviruses

The *Retroviridae* are a large and diverse family of viruses that have been primarily isolated from a wide range of vertebrate species, although they are also found in mollusks (1) and insects (2). Retroviruses are enveloped positive-stranded diploid RNA viruses that replicate in a host cell through the process of reverse transcription, i.e. the transcription of retroviral RNA into DNA (3). The life cycle of a retrovirus starts with the specific binding of retroviral particles to a host cell membrane, through the interaction between cellular receptors and viral surface proteins. After attachment and following entry into the host cell, the retroviral RNA genome is copied into double-stranded DNA using the reverse transcriptase enzyme. The genome is then transported to the nucleus and incorporated by a viral integrase enzyme into the host genome becoming a provirus. The virus thereafter replicates and its viral genes are expressed as part of the host cell DNA. Newly created viral proteins and nucleic acids are assembled to form new viral particles, which are then released by budding through the plasma membrane of the cell completing the retroviral life cycle (3, 4).

If retroviral integration occurs in the germ line or early-stage embryos, the proviruses can be inherited across generations. This event gives rise to an endogenous retrovirus (ERV) (4) that is transmitted vertically from parent to offspring by Mendelian inheritance (5). Retroviruses can also exist as horizontally transmitted infectious viruses which are transmitted from one individual to the other and are called exogenous retroviruses (5).

Retroviral particles (virions) are enveloped, about 100 nm in diameter, and their morphology consists of two proteinaceous structures, a dense core and an envelope that surrounds the core (6). The viral core contains two copies of the retroviral RNA genome, which is protected from degradation by nucleocapsid (NC) proteins (Fig. 1). The core also encloses the viral enzymes protease (PR), reverse transcriptase (RT) and integrase (IN) (3), and is surrounded by a protein capsid (CA) (3) (Fig. 1). The viral particle is enclosed in a lipid bilayer envelope derived from the plasma membrane of the host cell during the budding process, studded with viral glycoproteins composed of two subunits, the transmembrane (TM) unit anchored in the virion envelope and the surface (SU) unit exposed on the virion surface (6) (Fig. 1).

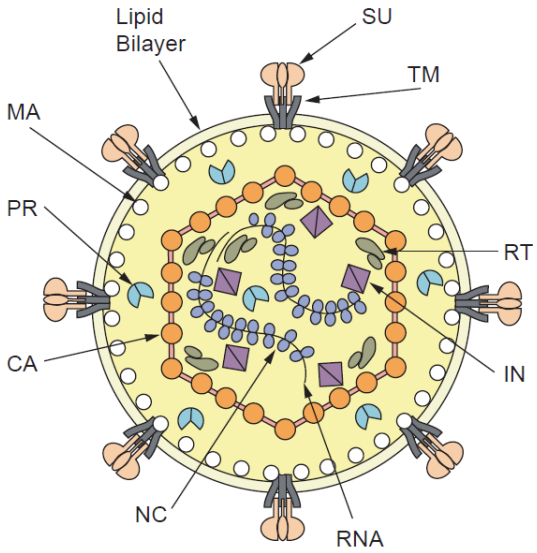


Figure 1: Schematic structure of a retroviral particle. Figure taken with permission from Voisset and Andrawiss (7) and Vogt (4).

Retroviral genomes consist of two, usually identical, molecules of single-stranded RNA, ranging from about 7 to 10 kilobases in length. They contain three major genes which code for essential proteins for viral structure and function: group specific antigen (encoded by *gag* gene), polymerase (*pol*) and envelope (*env*) (Fig. 2). These genes are flanked by 5' and 3' long terminal repeats (LTRs). Each LTR consists of three regions – untranslated 3' (U3), repeat (R) and untranslated 5' (U5) (8) – and contains many elements regulating transcription of the integrated retroviruses. Even if the same elements are present in each LTR, the majority of retroviruses use the 5' LTR for transcription initiation and the 3' LTR for termination (6).

The genome order 5' LTR-*gag*-*pol*-*env*-LTR 3' is conserved amongst known retroviruses (9). The *gag* gene encodes the CA, MA and NC proteins, which are involved in assembly and packaging of newly formed retroviral particles, as well as forming structural components of the virion. *Pol* encodes two enzymes, RT and IN, which are essential for retroviral transcription and integration. *Env* encodes the SU and TM glycoproteins of the retroviral envelope: SU is involved in receptor binding and TM in membrane fusion (3) (Fig. 2). An additional, smaller, coding domain present in all retroviruses, and located between *gag* and *pol* open reading frames, is *pro*, which encodes the virion protease. This protein is involved in cleaving viral polyproteins into their separate subunits. The genome of *complex* retroviruses can contain several other genes that regulate genome expression or replication and are not present in *simple* retroviruses.

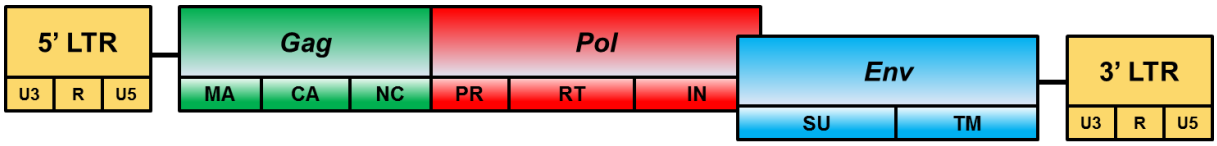


Figure 2: Schematic representation of the genomic structure of a simple retrovirus.

The family *Retroviridae* is currently divided into two subfamilies and seven genera (Fig. 3; Tab. 1), although retroviruses have been previously grouped largely based on virion morphology (type B, C, and D) (3). The *Orthoretrovirinae* subfamily consists of the genera *Alpha-*, *Beta-*, *Gamma-*, *Delta-*, *Epsilon-retrovirus* and *Lentivirus*, whereas the *Spumaretrovirinae* contains only the *Spumavirus* genus (International Commitee on Taxonomy of Viruses, 2002) (Fig. 3; Tab. 1). These classifications are based on morphological and structural characteristics, life cycle, accessory genes and genetic similarity. Except for lentiviruses and spumaviruses, the other five genera include retroviruses with oncogenic potential (formerly referred to as oncoviruses) and which can exist in both exogenous and endogenous forms.

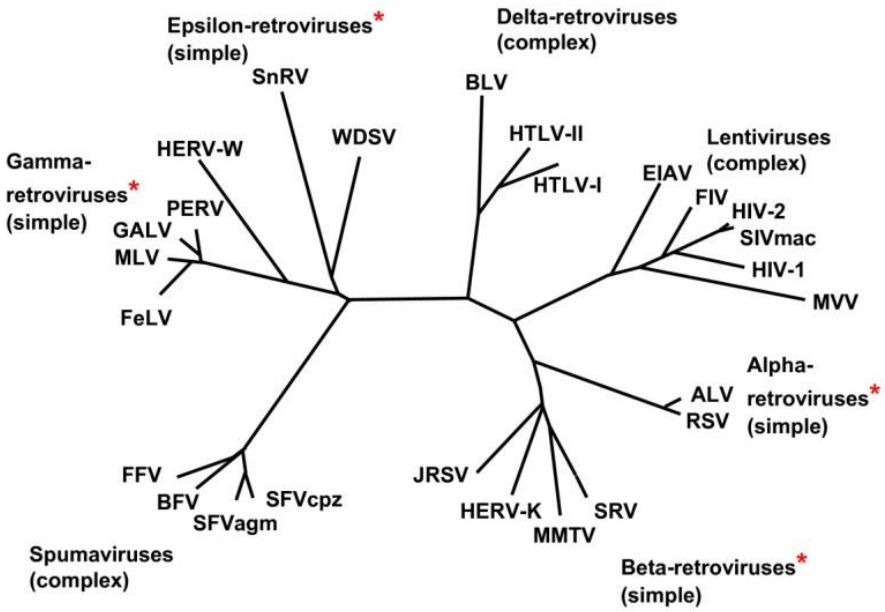


Figure 3: Phylogeny of retroviruses. Genera including endogenous retroviruses are marked with a red asterisk. Figure taken with permission from Weiss (10).

Table 1: Classification of retroviruses

Genus	Virion type	Genome	Example
Alpharetrovirus	C	Simple	Rous sarcoma virus (RSV)
Betaretrovirus	B and D	Simple	Mouse mammary tumour virus (MMTV)
Gammaretrovirus	C	Simple	Murine leukemia virus (MLV)
Deltaretrovirus	C	Complex	Human T-lymphotropic virus (HTLV)
Epsilonretrovirus	C	Simple	Walleye dermal sarcoma virus (WDSV)
Lentivirus	Lenti	Complex	Human immunodeficiency virus (HIV)
Spumavirus	Spuma	Complex	Simian foamy virus (SFV)

1.2 Importance of studying retroviruses

Interest in retroviruses relates to their importance as human and animal pathogens. Indeed, retroviral infections can cause malignant disease, as well as a range of other pathogenic states, in a broad range of species. Some retroviruses lead to disease through progressive immunodeficiency. This is the case for the human immunodeficiency virus types 1 and 2 (HIV-1, HIV-2), the causative agents of AIDS (acquired immunodeficiency syndrome), a disease affecting over 35 million people worldwide and causing approximately 1.2 million deaths per year (http://www.who.int/gho/hiv/). A similar virus, feline immunodeficiency virus (FIV), affects approximately 11% of cats worldwide and is responsible for a disease which is usually fatal due to the progression to feline acquired immunodeficiency syndrome (11). Some others, the oncogenic retroviruses, like MMTV (12), HTLV (12) and RSV (13), cause excessive cell proliferation through cellular transformations and tumours.

However, the attention raised by retroviruses extends beyond their importance as pathogens. Research in the area of retrovirology has led to the discovery of oncogenes, a major advance in cancer genetics: the first confirmed oncogene was indeed discovered in 1970 in a chicken retrovirus (RSV) (14), then followed by the discovery of other viral oncogenes in retroviruses of a wide range of mammals (rodents, cats and monkeys for example) (15). Studies of the viral oncogenes in turn led to the discovery of cellular proto-oncogenes in the host genomes. Cellular oncogenes have been shown to be activated in a variety of human cancers, including those with no viral involvement (16).

In addition, retroviruses are proving to be valuable research tools in molecular biology and have been the preferred gene transfer vectors in clinical gene therapy. Gene therapy consists in the delivery of nucleic acids into patient's cells for therapeutic purposes, and gammaretroviral and lentiviral vectors are used to mediate stable genetic modification of treated cells by chromosomal integration of the transferred vector genomes (17). Retroviral vectors have also been successfully used in cancer gene

therapy, a technique based on the use of the cytopathic effect of replication-selective oncolytic viruses to selectively target and kill tumor cells (18).

Interest in retroviruses has been further stimulated by the fact that retroviruses are known to infect new host species by horizontal transfer (19). There are several examples of naturally occurring cross-species transmissions involving retroviruses (19), some of which resulted in the emergence of novel fatal diseases. This is the case of the transmission of the simian precursors of HIV-1 and HIV-2 into humans which finally led to the AIDS epidemic (20).

The importance of retroviruses, especially in their endogenous forms, also relies on the fact that they represent a fundamental element constituting the genome of a wide range of vertebrate species.

1.3 Endogenous retroviruses

When a retrovirus integrates into a germ line, rather than a somatic cell, it has the potential to become an endogenous retrovirus (ERV) and be inherited as part of the host genome across generations (21). Once a retrovirus has endogenised, it is subject to selection, mutation and genetic drift like any other genetic element, and can spread through the host population or be eliminated from the population entirely (9). ERVs have been identified in all vertebrate genomes examined (9, 22), and they often occupy a substantial fraction of mammalian genomes, accounting for about 8% of human (23) and 10% of mouse nuclear genome sequences (24), greater than protein-coding sequences (1 to 2% in the human genome) (IHGS Consortium). Analysis of ERVs in host genomes indicates a long-standing association between retroviruses and vertebrates, probably dating back several hundreds of million years, during which retroviruses have repeatedly colonized host genomes (25). The integration and replication of ERVs in vertebrate genomes represent a source of genetic variation which is thought to have had a strong impact on the biology and evolution of host species (5, 9). There are several possible fates of an ERV: for example, it can remain replication competent and keep the ability to produce virus particles; it can proliferate within the genome by retrotransposition; it can become inactive through mutations and deletions and decay into noncoding DNA; or virus sequences can be co-opted into host genome function (26). Retrotransposition in particular represents an important factor in genome evolution and function through the incidental rearrangement of host DNA and the effects of retroelement insertion on gene expression and function. For example, the possible mutagenic effects of retrotransposition includes disrupting the function of a host gene by inserting into it or locating viral promoters near host genes which can alter gene expression, with either beneficial

(27) or negative (28) effects for the host. ERVs can also contribute adaptively to host genome evolution by providing sequences that can be utilized by the host (29). Among humans for example, *syncytin 1* and *2* are genes that were originally part of a retrovirus that became endogenous, but currently play a role in placental formation (30-32). Otherwise, some ERVs function to protect the host species by interfering with horizontal infection by exogenous retroviruses, e.g., by coding for intact envelope proteins that block the host cellular receptor used by exogenous viruses (33). Given its importance in shaping the genomes of most vertebrates, the process of retroviral endogenization – the invasion of the germ line by infectious retroviruses – has attracted much scientific interest. However, all ERVs identified until recently were found to be of ancient origin, derived from retroviruses that invaded the ancestral host genome many thousands or millions of years ago. Many ancient ERVs have been subject to extensive mutation and deletion, and in many cases the original exogenous viruses from which the ERVs are derived are now extinct (34). Thus, the evolutionary events that occurred during retroviral endogenization are obscured by time, making difficult an understanding of the mechanisms involved.

1.4 Koala retrovirus (KoRV)

The koala retrovirus (KoRV) is the only known case where an infectious exogenous retrovirus is currently in the process of invading the host germ line and becoming an endogenous part of its host species (35). KoRV indeed is still spreading both horizontally and vertically among koalas (*Phascolarctos cinereus*) (34) and therefore allows to study the process of retroviral endogenization as it is happening right now (35). KoRV was discovered in the late 1990s when gammaretrovirus particles were reported by electron microscopy in mitogen-stimulated peripheral blood mononuclear cell cultures from a wide range of koalas tested and koala lymphoma tissue (36). KoRV was originally thought to be an endogenous retrovirus: it was ubiquitous in South East Queensland koalas, present in koala sperm and inherited across generations (37). However, the associations of the virus with malignancies in koalas and the high level production of viral particles, together with the variation among individuals in proviral copy number and in number and pattern of proviral insertions (34), suggested the simultaneous exogenous nature of KoRV. Furthermore, KoRV is ubiquitous among northern Australian koalas, but is less common in southern Australian mainland and island populations (35, 38, 39), suggesting that KoRV initially affected koalas in northern Australia and is currently spreading to southern populations (34, 37). The analysis of museum specimens of koalas demonstrated that KoRV was already ubiquitous in northern Australian koalas by the late 19th century (40) and its genome conserved over the last 130 years of evolution (41).

KoRV variants with more limited distributions that are believed to have originated more recently and which are possibly exogenous have been discovered in the last years (42-44).

1.5 Effect of KoRV on koala health

KoRV has been associated with myeloid leukemia, lymphomas and immunodeficiencies in koalas (38). Even though a causative role for KoRV in development of these diseases has not been yet established, it is possible that KoRV is involved in inducing such pathologies since tumors and immunosuppression are common consequences of retroviral infections (45). There is also some evidence that KoRV infection may lead to higher susceptibility of koalas to *Chlamydia* infections (37, 38), as a consequence of the immunosuppressive effect which KoRV, like many other gammaretroviruses, produces (46). Infection with KoRV, and subsequent immunosuppresion, would thus provide opportunities for secondary infectious agents such as *Chlamydia* in less-resistant hosts (46). Immunosuppression has been demonstrated for HIV-1, HIV-2, FIV, MuLV, and FeLV, and is associated with opportunistic infections (47). For example, *Chlamydia* infections are associated with HIV and FIV infections (48, 49). The mechanism how retroviruses induce immunodeficiency is still not completely understood, but there is accumulating evidence that the viral transmembrane (TM) protein of all gammaretroviruses including KoRV is involved. TM proteins of HIV-1, HERV-K, PERV and KoRV (50) have been demonstrated to inhibit lymphocyte activation by mitogens and modulate cytokine expression in peripheral blood mononuclear cells (PBMCs). Moreover, all retroviral TM proteins contain a highly conserved sequence, the so-called immunosuppressive (isu) domain, and synthetic peptides corresponding to these domains have been shown to inhibit lymphocyte activation and to modulate gene expression (51-53). By these means, immunosuppression impairs antibacterial defenses and therefore creates a more permissive microbial environment where opportunistic pathogens can better survive, causing ultimately dysbiosis or imbalance of the microbiome. This is the likely scenario for how KoRV predisposes koalas to more severe chlamydial disease. Chlamydiosis indeed occurs at an extremely high incidence in koalas (70-98% of populations in south-east Queensland and Victoria) (54, 55), and is considered a major health threat to the species (38). Chlamydiosis in koalas is caused by *Chlamydia pecorum* and *C. pneumoniae* (56). Of the two, infection with *C. pecorum* is more severe, occur at ocular and urogenital sites and can result in impaired reproduction, infertility, or blindness (57). Likewise, leukemia and lymphoma are present in 3-5% of necropsies in the wild, and may cause up to 60-80% of koala mortality in some captive

colonies (36, 38). Even if koalas are not considered a threatened species, according to the World Wildlife Foundation, the combination of mortality due to leukemia and lymphoma with the effect of *Chlamydia* infection could lead to the local extinctions of koalas within the next 50 years (46).

KoRV has attracted attention in the last decades both as a model to study the process of retroviral endogenization and in the context of koala conservation. However, KoRV is also interesting for its peculiar phylogenetic relationships. KoRV is genetically most closely related to the gibbon ape leukemia virus (GALV), a retrovirus which infects apes from Southeast Asia.

1.6 Gibbon ape leukemia virus (GALV)

Gibbon ape leukemia virus (GALV) is an exogenous gammaretrovirus associated with hematopoietic neoplasms in captive colonies of white-handed gibbon (*Hylobates lar*). GALV was discovered in captive gibbons in the early 1970s following an outbreak of lymphocytic and myelogenous leukemia (58, 59). Investigations on captive gibbons housed in breeding facilities in Thailand, the United States and Bermuda revealed 11.3% antibody prevalence, 3% viremia, and 8% with neoplastic malignancies (60) and led in quick succession to the isolation of several different strains of GALV. The first was isolated from an animal with lymphocytic leukemia in a colony at the San Francisco Medical Center (strain SF) (61, 62). GALV was later isolated from a gibbon with granulocytic leukemia, at the South East Asia Treaty Organization Medical Research Laboratory in Bangkok, Thailand (strain SEATO) (63, 64), where several other individuals were diagnosed with the same disease, and from another gibbon with lymphocytic leukemia from a colony on Hall's Island, near Bermuda (strain GALV-H) (65, 66). The Brain strain, instead, was identified in frozen brain samples from two healthy gibbons injected with brain extracts from human patients with kuru and from an uninoculated cage mate, imported from Southeast Asia and stored at the Gulf South Primate Center, National Insitutes of Health, Louisiana (67). In 1971, while GALV was being isolated from gibbons, a closely related retrovirus was identified in a 3-year-old male woolly monkey (*Lagothrix lagotricha*) diagnosed with fibrosarcoma (68). The virus was originally designated SSAV (simian sarcoma-associated virus) and was found to exist as a mixture of a replication-defective transforming virus (SSV - simian sarcoma virus) and its associated replication-competent helper virus (SSAV) (69). SSAV has been recently renamed woolly monkey virus (WMV). WMV was found to be related to GALV as supported by immunological (70) and serological tests (68), antigenic similarities in some gene products (67, 71, 72), and high RNA sequence homology (65, 67). The woolly

monkey from which WMV was isolated was kept as a pet in an apartment in San Francisco, alongside a white-handed gibbon for the 3 months before its death (58). Therefore, WMV has been suggested to be the product of a horizontal transfer of GALV from the gibbon to the woolly monkey and is considered a member of the GALV lineage (68). GALV has also been isolated as a contaminant in various cell lines (73, 74), including an HIV-infected human cell line (75). This GALV strain was named GALV-X (76) and its origin remains unknown.

The SEATO strain has been shown to cause chronic myelogenous leukemia when injected into juvenile gibbons (77), while GALV-H and GALV-SF have been identified by seroepidemiology as the primary agent of lymphocytic leukemia in gibbon apes (78).

No GALV outbreaks have been described since the 1970s and in 2015 a survey of GALV infection in gibbons maintained in North American zoos has revealed that no animal was positive for GALV (59). However, the current prevalence of GALV infection in Asian free-ranging gibbons remains unknown.

GALV has been widely studied not only as pathogen of gibbons, but also in respect to its utility as biomedical tool. Because of its broad host range, GALV-based retroviral vectors have been developed for use in gene transfer (79). In particular, the cell surface receptor for GALV has been found to be expressed on human adult and fetal tissues. Thus, GALV has provided a useful source of envelopes for retroviral vectors frequently used in current gene therapy protocols (80). GALV has also been used in cancer gene therapy, with positive results in the treatment of certain types of tumors (81). The strong cytotoxic effect of GALV envelope fusogenic membrane glycoprotein can be used to efficiently kill tumor cells, after transduction into target cells (82).

Given the importance of GALV as an epizootic agent and clinical tool, the lack of genetic information on this virus has been surprising. Before the studies on GALV presented in this thesis, full-genome sequences had been determined for only two (GALV-SEATO and GALV-X) of the six strains of GALV which have been isolated to date. Furthermore, the genome sequence of SEATO available in GenBank (83) was chimeric, with part of the *pol* gene of the SF strain incorporated into the SEATO genome. Until this thesis, only *envelope* sequences had been determined for the remaining GALV strains (Brain, Hall's Island, and SF) (84).

1.7 Cross-species transmission

Cross-species transmission (CST) consists in the transfer of a viral infection from one species to another. Several factors play a key role in the occurrence and outcome of such events, such as rapid mutation which allows viruses to overcome host-specific

barriers, host contact rates, host evolutionary relationships and biological similarity in host defense systems which lowers the adaptation challenge the virus faces (85). CST is common among retroviruses, especially among gamma- and beta-retroviruses, which seem to have an inherent capacity to switch across diverse mammalian hosts (19, 86). In particular, gammaretroviruses show a precise evolutive pattern where interorder transmission (e.g. between primates and rodents) is much more frequent than interclass transmission (e.g. from birds to mammals) (86). The consequences of CSTs can be very different, depending on the species and the retrovirus involved: some are fatal in the new host; others are asymptomatic; some retroviruses cause diseases in the new host even if they were apathogenic in the original species (e.g. HIV-1 and HIV-2); others can be pathogenic or apathogenic in both species; or, finally, the virus may become endogenous in the new species (20). KoRV represents an excellent example of endogenization following a CST.

GALV is the most closely related retrovirus to KoRV among those sequences to date, so that GALV and KoRV are considered to be derived from a common ancestral virus (26, 36). Nevertheless, the host species of the two viruses, koalas and gibbons, are distant from an evolutionary point of view: placental and marsupial mammals split at least 160 million years ago (87) and KoRV-like viruses are absent from other marsupials (19, 36). Therefore, gibbons and koalas unlikely acquired the two viruses from a common ancestor. Furthermore, koalas are endemic to eastern and southern mainland Australia, while white-handed gibbons are distributed in mainland Southeast Asia, making direct natural transmission of the virus improbable. A similar pattern of closely related viruses existing in such diverse species is explainable with a CST event. Given the evolutionary and geographical isolation of koalas and gibbons, the natural transfer likely occurred via an intermediate host (19, 36, 88).

It has been hypothesized that rodents, and in particular rats, serve as major reservoir for gammaretroviral spread among mammalian orders (53% of all known gamma-ERVs occur in rodent taxa) (86), in reason of their widespread distribution and of the commensal behavior which facilitates their dispersal. Furthermore, since several Southeast Asian rodent species of the family Muridae (*Mus caroli*, *Mus cervicolor* and *Vandeleuria oleracea*) (89-91) harbor ERVs known to cross-hybridize with GALV at high stringency and these ERVs have been suggested to be ancestral to GALV, it has been hypothesized that GALV may have originated from them (37) (Fig. 4). However, the sequences of these ERVs have never been determined. More recently, genomic sequences of ERVs found in the genome of Asian rodent species *Mus caroli* (McERV) (92), *Mus dunni* (MDEV) (93) and *Mus musculus* (MmERV) (94) have been reported, but, even though they are part of the same clade, they are not closely enough related to KoRV

and GALV to be considered their progenitors. However, the evidence that GALV is more closely related to several Southeast Asian rodent ERVs than to any ERV found in primates and that GALV is not endogenous in gibbons suggests that the precursor of GALV evolved in rodents, probably in Southeast Asia, and spread secondarily to gibbons and koalas (95). In particular, the outbreak of GALV in gibbons in the 1970s may have originated from a single spillover event from rodents in Southeast Asia. Such a CST case represents a good example of how a long-term ERV resident in one species (likely rodent) can infect an unrelated species causing epidemics (gibbon), and eventually become endogenized in a newly adopted host (koala) (96). However, the precise species giving rise to GALV and KoRV remains to be determined.

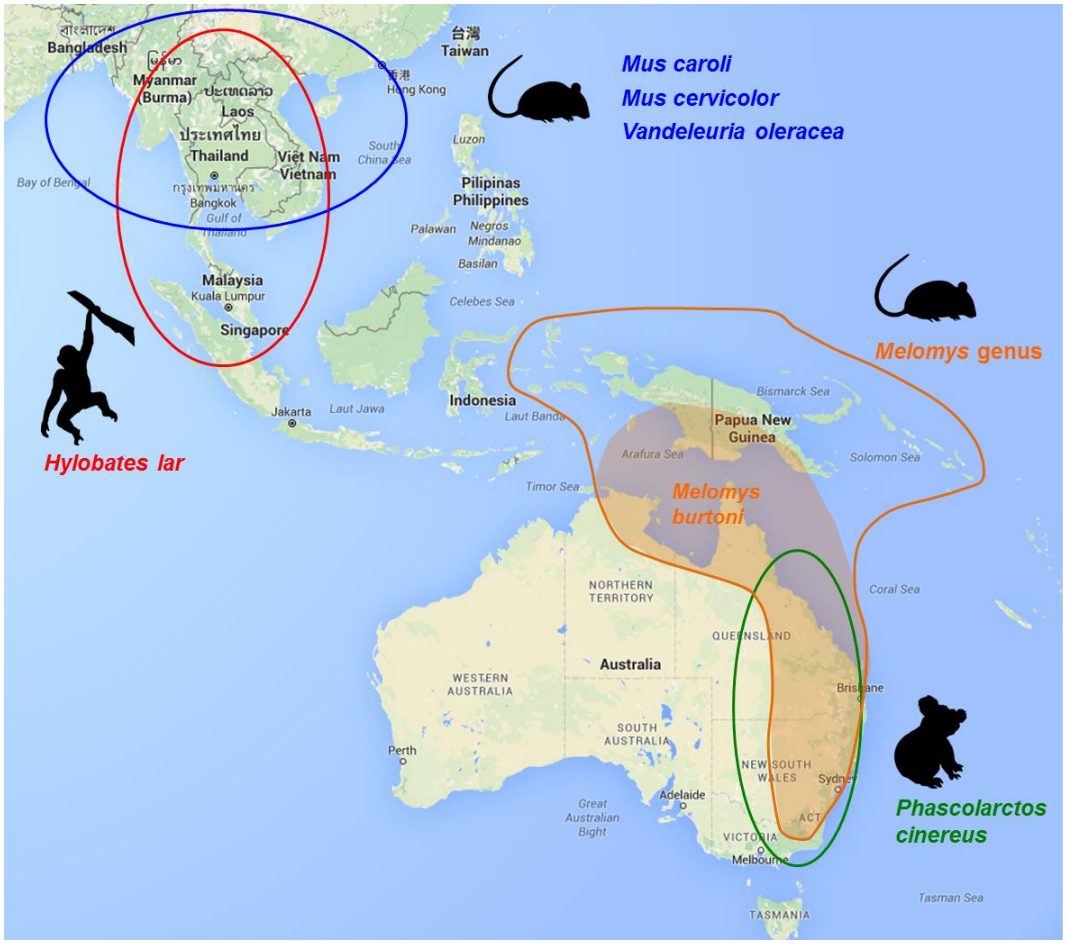


Figure 4: Geographic distribution of the species harboring KoRV, GALV or related retroviruses. The distribution of *Phascolarctos cinereus* (koala), *Hylobates lar* (white-handed gibbon), *Mus caroli*, *Mus cervicolor*, *Vandeleuria oleracea* and genus *Melomys* is shown. The known distribution of *Melomys burtoni* until this thesis is represented as a shaded area. The map was extracted from Google Maps 2016.

In the attempt to identify the reservoir host which carries the progenitor of KoRV and GALV, a recent study showed the results of a screen of a wide range of native or introduced vertebrate species in Australia for the presence of viruses related to KoRV and GALV (97). Partial proviral sequences sharing close identity with KoRV and especially with GALV were obtained from a native Australian rodent, the grassland mosaic-tailed rat *Melomys burtoni*. The new virus was named *Melomys burtoni* retrovirus (MbRV) and could be considered another strain of GALV (97). The geographic overlap between *Melomys burtoni* and koalas and the high identity between MbRV and KoRV suggests there may have been a CST between koalas and grassland *Melomys* at some time in the past. By contrast, the genus *Melomys*, which has a wide Australo-Papuan distribution ranging from eastern Australia to the Melanesian islands (98), is not present in mainland Southeast Asia, where gibbons are distributed, even though MbRV is very closely related to GALV (Fig. 4). Therefore, it is improbable that MbRV is the direct progenitor of GALV and GALV´s source remains unknown. However, it is possible that several intermediate hosts may have been involved the CST among koalas and gibbons (99), in a stepwise process which finally led to the outbreak of GALV in gibbons and to the emergence of KoRV in koalas.

1.8 Receptors involved in KoRV/GALV cross-species transmission

In order for a CST to happen, a virus must be able to efficiently infect the appropriate cells of the new host. This process can be restricted at several different levels, including receptor binding, entry or fusion, trafficking within the cell, genome replication, and gene expression (100). Other significant impediments to infection include intracellular mechanisms, like the APOBEC and TRIM5-α proteins systems, which restrict cell infection by retroviruses (100). Virus entry is largely dependent on the interactions between virus particles and their receptors at the host cell surface. In retroviruses, it is the surface unit (SU) of the retroviral envelope protein which initiates entry by binding to a specific cell surface receptor. Studies of gammaretroviruses have suggested that the major determinant for receptor specificity resides in the variable regions A and B (VRA/VRB) of the SU, which are often collectively referred to as the receptor binding domain (RBD) (84, 101, 102). For example, KoRV-B and KoRV-J, the variants of KoRV which are believed to be exogenous and more pathogenic, have been shown to encode an envelope characterized by a significantly different RBD, specifically VRA, compared to the endogenous KoRV (KoRV-A), resulting in the use of an alternative receptor to that of KoRV-A (42, 43).

Retroviral receptors are usually surface transporter proteins which do not play any active role in receiving the virus, but simply provide an attachment point to a target cell and a signal to start viral entry into the cell (103). Therefore, the function of these proteins as retroviral receptors *per se* is not dependent on their function as transporters (104). The virus-receptor interaction is highly specific, and a single amino acid change in the receptor can completely abrogate viral binding. For example, HIV uses the immune signaling protein CD4 as a receptor and, despite the widespread distribution of CD4 in mammals, HIV-1 can use only the homolog found in primates for entry (103). In particular, gammaretroviruses exhibit a propensity to use as receptors multiple membrane-spanning carrier facilitator transporter proteins found on the surface of a wide variety of cell types (101, 105). GALV, WMV and KoRV-A have been demonstrated to use the same type III sodium-dependent phosphate transporter membrane protein (SLC20A1, also called PiT1) to infect human cells (106, 107). PiT1 is a multiple membrane-spanning protein which is ubiquitously expressed in mammalian cells (108, 109). Despite their genetic similarities, GALV and KoRV have overlapping but distinct host ranges. GALV can infect *in vitro* several mammalian cells, such as those derived from felids, canids, bovids, rats, hamsters, bats, minks, monkeys, and humans, but fails to infect most mice cells (84, 107). KoRV also has a broad *in vitro* host range, but is able to infect mice cells, together with rat, bovine, human and hamster cells (107). WMV has a similar host range to the other GALVs but cannot infect hamster cells. Besides PiT1, GALV can use also PiT2, a paralog of PiT1, as receptor to infect Chinese hamster and Japanese feral mouse cells, which are the only mice cells GALV has been shown to be able to infect so far (110, 111). In contrast to KoRV-A, KoRV-B and KoRV-J use the thiamine transport protein 1 (THTR1) as a receptor (42, 43). KoRV-B, similarto KoRV-A, infects a wide range of cells from different species including human (43). Using pseudotyped KoRV-J, infection of human and cat cells was observed, but not of rat and mouse cells (42). The diversification in the KoRV envelope gene, and subsequent receptor usage, observed in the different KoRV variants, has been proposed as a mechanism used by KoRV to overcome superinfection blocks and broaden host tropism (112). Superinfection resistance consists in the capacity of cells to prevent a second infection by the same virus or a virus which is closely related (and use the same receptor) to the one that has already established an infection.

1.9 Virus-host “arms race” and positive selection

Superinfection resistance, together with other cellular restriction factors and host immune responses (e.g. APOBEC and TRIM5-α proteins), are used by the host to counteract viral infection and replication. Viruses in turn can switch receptor usage or

encode proteins that antagonize these systems (e.g. Vif or Vpu in HIV) (103). Host genomes are continuously under selective pressure to encode protective systems to better recognize and avoid viral infections, while viruses are continuously selected to circumvent these blocks. The resulting evolutionary conflict, often referred to as an ‘‘arms race’’ (113, 114), deeply influences the evolution of both viral and host proteins involved in virus-host interactions, driving continuous rounds of selection for advantageous mutations for either the virus or the host (positive selection). Signatures of positive selection, and therefore of these evolutionary struggles, can be detected and quantified analyzing the rates of nucleotide substitutions among orthologous genes (115). Nucleotide changes within protein coding sequences can be synonymous, i.e. they do not result in an amino acid change, or non-synonymous, i.e. they cause an amino acid change. Generally non-synonymous substitutions are deleterious and therefore are eliminated by purifying (negative) selection. In this case we expect synonymous changes to occur more frequently than non-synonymous ones, and the ratio of non-synonymous substitutions per non-synonymous sites (dN) to synonymous substitutions per synonymous sites (dS) to be below one (dN/dS ≤1) (116). Sometimes though, non-synonymous changes can be beneficial and become fixed in a population under diversifying (positive) selection. In this case we expect a deviations of the dN/dS ratio towards positive values (dN/dS >1) (117).

In the original models to estimate selection, the dN/dS ratio was measured as shared by all the sites of the gene under consideration and signatures of positive selection were hard to detect since substitutions in most sites of a gene are expected to be neutral or deleterious (118). Indeed, in host-virus arms race context, patterns of dN/dS>1 are not be expected across the entire length of a gene, but rather in the codons corresponding to the residues located at the critical interaction interfaces between host and viral proteins (113, 114). In the old models it was also assumed that positive selection remains constant throughout time and across lineages (118, 119). In this way, purifying selection acting on some lineages or sites of a gene would have masked the signal of positive selection on others, preventing signatures of positive selection to be detected. More recent algorithms allow the distribution of the dN/dS ratio to vary from site to site and also from branch to branch at a site, making it possible to identify situations when positive selection has acted only on a small proportion of sites or lineages (118, 120). Indeed, adaptive evolution frequently occurs in episodic bursts, localized to a few sites in a gene, and to a small number of lineages in a phylogenetic tree. This phenomenon is called episodic diversifying selection (118, 120).

1.10 High-throughput approaches to microbiology

Recent advances in nucleic acid high-throughput sequencing (HTS) technologies have dramatically changed almost every field of biological research, thanks to the capability of the HTS systems of rapidly sequencing and analyzing complex mixtures of nucleic acid templates, in a massively parallel fashion and for relatively little cost (121). This makes such technologies ideal tools for metagenomics, the study of total genetic content of a given sample, without the need of culturing the organisms present in it. Until recently, the study of both bacterial and viral communities have primarily relied upon culture-based methods, which are known to have extreme biases since most of bacteria and viruses cannot be grown in culture (122). This problem has now been largely overcome with the advent of HTS technologies. Several different HTS platforms have been commercialized to date, each one differing in sequencing technology, throughput, runtime, costs, read lengths, and error patterns (123). At the moment, the most widely used HTS technology is the Illumina sequencing (122, 124), which is able to produce the greatest throughput but with the drawback of short reads (up to 6 billions reads 300 base pairs long per run, with the HiSeq X Series) (http://www.illumina.com/). With constantly decreasing sequencing costs and increasing throughput, the development of the HTS technologies has posed a new challenge in terms of space and computational power needed to store and analyze the huge volumes of data generated (123). When only particular portions of a whole genome need to be analyzed, costs, data storage space and computational efforts can be reduced significantly by selective recovery and subsequent sequencing of genomic loci of interest, compared with shotgun sequencing and metagenomics, where the whole genetic content of a sample is sequenced (125).

1.11 Targeted enrichment methods

Targeted enrichment (or targeted resequencing) consists in selecting and then sequencing only defined regions of interest of a genome. Amplicon sequencing and hybrid capture are two of the possible approaches that can be used for targeted enrichment (125, 126). In amplicon sequencing a polymerase chain reaction (PCR) is directed toward a specific genomic region of interest and the following ultra-deep sequencing of the PCR products (amplicons) allows efficient variant identification and characterization in the targeted region. For example, amplicon sequencing is widely used to sequence the bacterial 16S ribosomal RNA gene across multiple species to study the composition of microbial communities, or microbiomes. The 16S rRNA is an optimal molecular marker for assessing microbial diversity since it is highly conserved across bacteria and includes

nine hyper-variable regions (V1-V9) flanked by relatively conserved regions which allows the design of universal primers to amplify and analyse the variation in such hypervariable regions (127). 16S rRNA genes from hundreds of thousands of organisms have been fully sequenced and classified, therefore very large and comprehensive ribosomal databases are available to allow bacterial identification, generally up to the genus level (128). Microbiome diversity can be inferred using clustering methods (127). This approach consists in clustering sequences into Operational Taxomonic Units (OTUs) based on a certain threshold of sequence similarity. The most abundant sequence within each OTU is chosen as the OTU's representative sequence and is aligned and taxonomically classified against one of the 16S rRNA databases available (129). In this way it is possible to produce a taxonomic profile of a microbial community and estimate population richness and diversity. Bioinformatics pipelines and software packages such as mothur (130) and QIIME (129) have been developed to perform microbial community analyses using HTS data. These techniques have allowed to study microbiome composition in complex bacterial environments such as soil (131), ocean (132), biofilms (133), groundwater (134), cow rumen (135) and the gut of humans and other animals (136).

Hybrid capture represents an alternative to amplicon sequencing to perform targeted enrichment. It consists in the hybridization of the nucleic acids from an input sample, in the form of genomic sequencing libraries, to specific immobilized probes (PCR products or synthesized oligonucleotides), which are complementary to the targeted regions of interest, so that the sequences of interest of the input sample can be physically captured, eluted and finally sequenced (125, 137). Hybrid capture reactions can be performed either in solution or on a solid support (on-array capture). Hybrid capture approaches give the possibility to capture large target regions (several kilobases) in a single experiment, even though they can suffer from off-target capture and achieve suboptimal enrichment over the complete region of interest (125, 138). In contrast, amplicon sequencing is preferable when smaller regions need to be analyzed, since it allows a deep and even coverage across the target region (125). However, being based on PCR, amplicon sequencing has the disadvantage to be affected by primer-target mismatches and amplification biases. Hybrid capture, in contrast, can tolerate bait-target mismatches well over 15% (139), even though significant insertion/deletion mutations are not easy to be identified by this approach (138).

1.12 Applications of high-throughput sequencing to virology

High-throughput sequencing techniques have enabled significant contributions to multiples areas of virology. HTS methods have been applied in:

1) reconstruction of full-length viral genomes, even in the case of unknown or poorly characterized viruses;
2) viral discovery, and more specifically, virus candidate pathogen discovery in human and animal diseases;
3) characterization of the virome of environmental or animal samples (viral metagenomics);
4) investigation of viral variability within the host (i.e., quasispecies);
5) detection of antiviral drug resistance (122, 140-143).

One of the main problems in using high-throughput sequencing methods to study viruses is the usually very small proportion of viral nucleic acids in an animal or environmental sample as compared to host-derived or environmental genetic material (144). Therefore, in most cases viral material is hard to detect using pure shotgun sequencing, despite the high coverage depth these techniques guarantee, since no more than a few viral sequence reads can be expected per million reads from host or environmental DNA (144). This can be especially true for viruses integrating into the host genome, for example retroviruses. Retroviral genomes rarely exceed 10 kb, and hence constitute a minor fraction of the genome of the infected host cell. Moreover, the infected cell type may constitute only a small fraction of the sample, and finally, the infected cells may contain a relatively low number of viral genome copies (144). In general, the proportion of viral nucleic acids can be considerably enriched using samples low in contaminating host nucleic acid, such as feces (145) or serum (146), or by mechanical and enzymatic procedures that reduce the host genetic material combined with (random) amplification of the capsid-protected nuclease-resistant viral nucleic acids (143, 144). However, these techniques cannot be applied to the study of integrated proviral DNA, for which, instead, target enrichment by hybrid capture can be performed. This approach enables to enrich viral nucleic acid prior to deep sequencing, allowing the removal of contaminating host genetic materials and maximizing sensitivity for viral detection. For example, hybrid capture has been used in clinical virology to enrich clinical samples for HIV-1 (144), herpesviruses (147) and Merkel cell polyomavirus (148) sequences. The same technique has been also recently used to recover KoRV genome from modern and museum DNA samples of koala in order to investigate the evolution of this virus (41).

2. Study aims

The two primary aims of this study were (i) to evaluate the effect of KoRV on koala health through the study of koala microbiome and (ii) to investigate the evolutionary

history of KoRV and of its closest relative GALV, trying to identify the intermediate hosts involved in the cross-species transmission between gibbons and koalas.

Regarding the first objective, KoRV is believed to cause immunosuppression in koalas, which may lead to higher susceptibility to secondary pathologies like the highly prevalent *Chlamydia* infection. It is likely that KoRV-induced immunodeficiency lowers koala antimicrobial defenses allowing opportunistic pathogens, such as *Chlamydia*, to colonize and eventually perturb koala bacterial communities (microbiomes). In **chapter II**, I characterized the microbiome of two healthy KoRV-positive koalas in different body regions (eye, mouth, rectum and feces) in order to create a baseline for koala microbiome. This will be useful for future comparisons with KoRV-negative or *Chlamydia*-infected koalas to assess the effect of KoRV and *Chlamydia* infection on koala microbiome. The analysis of digestion-associated organs microbiomes tried also to address the question whether koalas, since they have an unique diet based almost exclusively on *Eucalyptus* leaves, have unusual bacterial communities compared to other mammalian species.

In order to pursue the second aim of this thesis, the study of the evolutionary history of KoRV and GALV, it was necessary to have the genomes of the two viruses characterized. If the genetics of KoRV and of its variants has been widely studied in the last decade, there is a surprising lack of genetic information about GALV. Therefore, in **chapter III**, I recovered the full genome sequences of all GALV strains, described their genomic structure and analysed the phylogenetic relationships within the GALVs and with the other gammaretroviruses, examining the selection pressures acting on the GALV/KoRV clade. After retrieving such information, in **chapter IV**, I investigated the origin of GALV and KoRV. A wide range of rodent species from Southeast Asia were screened for the presence of GALV and KoRV-like sequences, in the attempt to identify potential GALV and KoRV intermediate hosts and the viral progenitor from which the two viruses originated. I describe a new retrovirus which was discovered in an Indonesian *Melomys burtoni* subspecies and which is an endogenous GALV.

In order to achieve these aims we used next-generation sequencing coupled with target enrichments techniques. In particular, we used two different enrichment approaches. Amplicon sequencing of the 16S ribosomal RNA was used in **chapter II** to identify and compare the bacteria present in the different koala body regions. Hybrid capture, instead, was used to recover the genomes of the GALV strains from GALV-infected cell lines in **chapter III**, and to search for KoRV and GALV sequences in Southeast Asian rodent samples in **chapter IV.**

3. References

1. **Poulet FM, Bowser PR, Casey JW.** 1994. Retroviruses of Fish, Reptiles, and Molluscs, p. 1-38. *In* Levy JA (ed.), The *Retroviridae*. Springer US, Boston, MA.
2. **Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, Dastugue B, Vaury C.** 2000. Life cycle of an endogenous retrovirus, ZAM, in *Drosophila melanogaster*. J Virol **74:**10658-10669.
3. **Goff SP.** 2007. Retroviridae: The Retroviruses and Their Replication. *In* Knipe DM, Howley PM (ed.), Fields Virology. Lippincott Williams & Wilkins, Philadelphia, PA.
4. **Vogt PK.** 1997. Historical Introduction to the General Properties of Retroviruses. *In* Coffin JM, Hughes SH, Varmus HE (ed.), Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
5. **Denner J.** 2010. Endogenous retroviruses. *In* Kurth R, Bannert N (ed.), Retroviruses: Molecular Biology, Genomics and Pathogenesis. Caister Academic Press, Norfolk, UK.
6. **Bannert N, Fiebig U, Hohn O.** 2010. Retroviral particles, proteins & genomes. *In* Kurth R, Bannert N (ed.), Retroviruses: Molecular Biology, Genomics and Pathogenesis. Caister Academic Press, Norfolk, UK.
7. **Voisset C, Andrawiss M.** 2000. Retroviruses at a glance. Genome Biology **1:**1-4.
8. **Lenasi T, Contreras X, Peterlin M.** 2010. Transcription, splicing & transport of retroviral RNA. *In* Kurth R, Bannert N (ed.), Retroviruses: Molecular Biology, Genomics and Pathogenesis. Caister Academic Press, Norfolk, UK.
9. **Jern P, Coffin JM.** 2008. Effects of retroviruses on host genome function. Annu Rev Genet **42:**709-732.
10. **Weiss RA.** 2006. The discovery of endogenous retroviruses. Retrovirology **3:**67.
11. **Richards JR.** 2005. Feline immunodeficiency virus vaccine: implications for diagnostic testing and disease management. Biologicals **33:**215-217.
12. **Burmeister T.** 2001. Oncogenic retroviruses in animals and humans. Rev Med Virol **11:**369-380.
13. **Weiss RA, Vogt PK.** 2011. 100 years of Rous sarcoma virus. J Exp Med **208:**2351-2355.
14. **Martin GS.** 2001. The hunting of the Src. Nat Rev Mol Cell Biol **2:**467-475.
15. **Cooper GM.** 1995. Oncogenes (2nd ed.). Jones and Bartlett Publishers, Boston, MA.
16. **Maeda N, Fan H, Yoshikai Y.** 2008. Oncogenesis by retroviruses: old and new paradigms. Rev Med Virol **18:**387-405.

17. **Muhlebach M, Schule S, Gerlach N, Schweizer M, Buchholz C, Hohenadi C, Cichutek K.** 2010. Gammaretroviral & lentiviral vectors for gene delivery. *In* Kurth R, Bannert N (ed.), Retroviruses: Molecular Biology, Genomics and Pathogenesis. Caister Academic Press, Norfolk, UK.
18. **Fu X, Tao L, Jin A, Vile R, Brenner MK, Zhang X.** 2003. Expression of a fusogenic membrane glycoprotein by an oncolytic herpes simplex virus potentiates the viral antitumor effect. Mol Ther **7:**748-754.
19. **Martin J, Herniou E, Cook J, O'Neill RW, Tristem M.** 1999. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. J Virol **73:**2442-2449.
20. **Denner J.** 2007. Transspecies transmissions of retroviruses: new cases. Virology **369:**229-233.
21. **Bannert N, Kurth R.** 2006. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet **7:**149-173.
22. **Gifford R, Tristem M.** 2003. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes **26:**291-315.
23. **Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.** 2001. Initial sequencing and analysis of the human genome. Nature **409:**860-921.
24. **Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al.** 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420:**520-562.
25. **Gifford R, Kabat P, Martin J, Lynch C, Tristem M.** 2005. Evolution and distribution of class II-related endogenous retroviruses. J Virol **79:**6478-6486.
26. **Bromham L.** 2002. The human zoo: endogenous retroviruses in the human genome. Trends in Ecology & Evolution **17:**91-97.
27. **Meisler MH, Ting CN.** 1993. The remarkable evolutionary history of the human amylase genes. Crit Rev Oral Biol Med **4:**503-509.
28. **Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, Mager DL.** 2012. Epigenetic interplay between mouse endogenous retroviruses and host genes. Genome Biol **13:**R89.
29. **Kidwell MG, Lisch DR.** 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution **55:**1-24.
30. **Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, Sapin V, Heidmann T.** 2005. Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A **102:**725-730.

31. **Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B.** 2004. The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci U S A **101:**1731-1736.

32. **Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, et al.** 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature **403:**785-789.

33. **McDougall AS, Terry A, Tzavaras T, Cheney C, Rojko J, Neil JC.** 1994. Defective endogenous proviruses are expressed in feline lymphoid cells: evidence for a role in natural resistance to subgroup B feline leukemia viruses. J Virol **68:**2151-2160.

34. **Tarlinton RE, Meers J, Young PR.** 2006. Retroviral invasion of the koala genome. Nature **442:**79-81.

35. **Stoye JP.** 2006. Koala retrovirus: a genome invasion in real time. Genome Biol **7:**241.

36. **Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF.** 2000. The nucleotide sequence of koala (*Phascolarctos cinereus*) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus. J Virol **74:**4264-4272.

37. **Tarlinton R, Meers J, Young P.** 2008. Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci **65:**3413-3421.

38. **Tarlinton R, Meers J, Hanger J, Young P.** 2005. Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. J Gen Virol **86:**783-787.

39. **Simmons GS, Young PR, Hanger JJ, Jones K, Clarke D, McKee JJ, Meers J.** 2012. Prevalence of koala retrovirus in geographically diverse populations in Australia. Aust Vet J **90:**404-409.

40. **Avila-Arcos MC, Ho SY, Ishida Y, Nikolaidis N, Tsangaras K, Honig K, Medina R, Rasmussen M, Fordyce SL, Calvignac-Spencer S, et al.** 2013. One hundred twenty years of koala retrovirus evolution determined from museum skins. Mol Biol Evol **30:**299-304.

41. **Tsangaras K, Siracusa MC, Nikolaidis N, Ishida Y, Cui P, Vielgrader H, Helgen KM, Roca AL, Greenwood AD.** 2014. Hybridization capture reveals evolution and conservation across the entire Koala retrovirus genome. PLoS One **9:**e95633.

42. **Shojima T, Yoshikawa R, Hoshino S, Shimode S, Nakagawa S, Ohata T, Nakaoka R, Miyazawa T.** 2013. Identification of a novel subgroup of Koala retrovirus from Koalas in Japanese zoos. J Virol **87:**9943-9948.

43. **Xu W, Stadler CK, Gorman K, Jensen N, Kim D, Zheng H, Tang S, Switzer WM, Pye GW, Eiden MV.** 2013. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proc Natl Acad Sci U S A **110:**11547-11552.

44. **Shimode S, Nakagawa S, Yoshikawa R, Shojima T, Miyazawa T.** 2014. Heterogeneity of koala retrovirus isolates. FEBS Lett **588:**41-46.

45. **Rosenberg N, Jolicoeur P.** 1997. Retroviral Pathogenesis. *In* Coffin JM, Hughes SH, Varmus HE (ed.), Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

46. **Macphee RD, Greenwood AD.** 2013. Infectious disease, endangerment, and extinction. Int J Evol Biol **2013:**571939.

47. **Denner J.** 2014. Immunization with envelope proteins of the KoRV as a basis for a preventive vaccine. *In* Pye GW, Johnson RN, Greenwood AD (ed.), The Koala and its Retroviruses: Implications for Sustainability and Survival. Technical Reports of the Australian Museum (online), The Australian Museum, Sydney, Australia.

48. **Contini C, Fainardi E, Seraceni S, Granieri E, Castellazzi M, Cultrera R.** 2003. Molecular identification and antibody testing of Chlamydophila pneumoniae in a subgroup of patients with HIV-associated dementia complex. Preliminary results. J Neuroimmunol **136:**172-177.

49. **O'Dair HA, Hopper CD, Gruffydd-Jones TJ, Harbour DA, Waters L.** 1994. Clinical aspects of *Chlamydia psittaci* infection in cats infected with feline immunodeficiency virus. Vet Rec **134:**365-368.

50. **Denner J, Young PR.** 2013. Koala retroviruses: characterization and impact on the life of koalas. Retrovirology **10:**108.

51. **Cianciolo GJ, Copeland TD, Oroszlan S, Snyderman R.** 1985. Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope proteins. Science **230:**453-455.

52. **Denner J.** 1998. Immunosuppression by retroviruses: implications for xenotransplantation. Ann N Y Acad Sci **862:**75-86.

53. **Ruegg CL, Monell CR, Strand M.** 1989. Inhibition of lymphoproliferation by a synthetic peptide with sequence identity to gp41 of human immunodeficiency virus type 1. J Virol **63:**3257-3260.

54. **Weigler BJ, Girjes AA, White NA, Kunst ND, Carrick FN, Lavin MF.** 1988. Aspects of the epidemiology of *Chlamydia psittaci* infection in a population of koalas (*Phascolarctos cinereus*) in southeastern Queensland, Australia. J Wildl Dis **24:**282-291.

55. **Lee AK, Martin RW, Handasyde KA.** 1988. Research into chlamydiosis in natural populations of the koala in Victoria, Phases 1-3.

56. **Bodetti TJ, Jacobson E, Wan C, Hafner L, Pospischil A, Rose K, Timms P.** 2002. Molecular evidence to support the expansion of the hostrange of *Chlamydophila pneumoniae* to include reptiles as well as humans, horses, koalas and amphibians. Syst Appl Microbiol **25:**146-152.

57. **Jackson M, White N, Giffard P, Timms P.** 1999. Epizootiology of *Chlamydia* infections in two free-range koala populations. Vet Microbiol **65:**255-264.

58. **Eiden MV, Taliaferro DL.** 2011. Emerging retroviruses and cancer. *In* Dudley J (ed.), Retroviruses and insights into cancer. Springer Press, New York.

59. **Siegal-Willott JL, Jensen N, Kimi D, Taliaferro D, Blankenship T, Malinsky B, Murray S, Eiden MV, Xu W.** 2015. 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 **46:**27-33.

60. **Kawakami TG, Sun L, McDowell TS.** 1977. Infectious primate type-C virus shed by healthy gibbons. Nature **268:**448-450.

61. **Kawakami TG, Huff SD, Buckley PM, Dungworth DL, Synder SP, Gilden RV.** 1972. C-type virus associated with gibbon lymphosarcoma. Nat New Biol **235:**170-171.

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

63. **DePaoli A, Johnsen DO, Noll MD.** 1973. Granulocytic leukemia in white handed gibbons. J Am Vet Med Assoc **163:**624-628.

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

65. **Gallo RC, Gallagher RE, Wong-Staal F, Aoki T, Markham PD, Schetters H, Ruscetti F, Valerio M, Walling MJ, O'Keeffe RT, et al.** 1978. Isolation and tissue distribution of type-C virus and viral components from a gibbon ape (*Hylobates lar*) with lymphocytic leukemia. Virology **84:**359-373.

66. **Reitz MS, Jr., wong-Staal F, Haseltine WA, Kleid DG, Trainor CD, Gallagher RE, Gallo RC.** 1979. Gibbon ape leukemia virus-Hall's Island: new strain of gibbon ape leukemia virus. J Virol **29:**395-400.

67. **Todaro GJ, Lieber MM, Benveniste RE, Sherr CJ.** 1975. Infectious primate type C viruses: Three isolates belonging to a new subgroup from the brains of normal gibbons. Virology **67:**335-343.

68. **Theilen GH, Gould D, Fowler M, Dungworth DL.** 1971. C-type virus in tumor tissue of a woolly monkey (*Lagothrix* spp.) with fibrosarcoma. J Natl Cancer Inst **47:**881-889.

69. **Wolfe LG, Smith RK, Deinhardt F.** 1972. Simian sarcoma virus, type 1 (*Lagothrix*): focus assay and demonstration of nontransforming associated virus. J Natl Cancer Inst **48:**1905-1908.

70. **Hino S, Stephenson JR, Aaronson SA.** 1975. Antigenic determinants of the 70,000 molecular weight glycoprotein of woolly monkey type C RNA virus. J Immunol **115:**922-927.

71. **Rangan SR.** 1974. Antigenic relatedness of simian C-type viruses. Int J Cancer **13:**64-70.

72. **Reitz MS, Jr., Luczak JC, Gallo RC.** 1979. Mapping of related and nonrelated sequences of RNA from woolly monkey virus and gibbon ape leukemia virus. . Virology **93:**48-56. .

73. **Okabe H, Gilden RV, Hatanaka M, Stephenson JR, Gallagher RE, Aaronson SA, Gallo RC, Tronick SR.** 1976. Immunological and biochemical characterisation of type C viruses isolated from cultured human AML cells. Nature **260:**264-266.

74. **Chan E, Peters WP, Sweet RW, Ohno T, Kufe DW, Spiegelman S, Gallo RC, Gallagher RE.** 1976. Characterisation of a virus (HL23V) isolated from cultured acute myelogenous leukaemic cells. Nature **260:**266-268.

75. **Burtonboy G, Delferriere N, Mousset B, Heusterspreute M.** 1993. Isolation of a C-type retrovirus from an HIV infected cell line. Arch Virol **130:**289-300.

76. **Parent I, Qin Y, Vandenbroucke AT, Walon C, Delferriere N, Godfroid E, Burtonboy G.** 1998. Characterization of a C-type retrovirus isolated from an HIV infected cell line: complete nucleotide sequence. Arch Virol **143:**1077-1092.

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

78. **Eiden M, Trainor CD, Reitz MS.** 1986. Gibbon ape leukaemia virus RNA in leukaemic T-lymphoid cell lines: expression of a novel RNA transcript. J Gen Virol **67 (Pt 7):**1455-1460.

79. **Miller AD, Garcia JV, von Suhr N, Lynch CM, Wilson C, Eiden MV.** 1991. Construction and properties of retrovirus packaging cells based on gibbon ape leukemia virus. J Virol **65:**2220-2224.

80. **Wormser G.** 2004. AIDS and other manifestations of HIV infection, Fourth Edition. Academic Press.

81. **Zhu B, Yang JR, Jiang YQ, Chen SF, Fu XP.** 2014. Gene therapy of lung adenocarcinoma using herpes virus expressing a fusogenic membrane glycoprotein. Cell Biochem Biophys **69:**583-587.

82. **Higuchi H, Bronk SF, Bateman A, Harrington K, Vile RG, Gores GJ.** 2000. Viral fusogenic membrane glycoprotein expression causes syncytia formation with bioenergetic cell death: implications for gene therapy. Cancer Res **60:**6396-6402.

83. **Delassus S, Sonigo P, Wain-Hobson S.** 1989. Genetic organization of gibbon ape leukemia virus. Virology **173:**205-213.

84. **Ting YT, Wilson CA, Farrell KB, Chaudry GJ, Eiden MV.** 1998. Simian sarcoma-associated virus fails to infect Chinese hamster cells despite the presence of functional gibbon ape leukemia virus receptors. J Virol **72:**9453-9458.

85. **Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE.** 2010. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science **329:**676-679.

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

87. **Luo ZX, Yuan CX, Meng QJ, Ji Q.** 2011. A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature **476:**442-445.

88. **Fiebig U, Hartmann MG, Bannert N, Kurth R, Denner J.** 2006. Transspecies transmission of the endogenous koala retrovirus. J Virol **80:**5651-5654.

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

90. **Benveniste RE, Callahan R, Sherr CJ, Chapman V, Todaro GJ.** 1977. Two distinct endogenous type C viruses isolated from the asian rodent *Mus cervicolor*: conservation of virogene sequences in related rodent species. J Virol **21:**849-862.

91. **Callahan R, Meade C, Todaro GJ.** 1979. Isolation of an endogenous type C virus related to the infectious primate type C viruses from the Asian rodent *Vandeleuria oleracea*. J Virol **30:**124-131.

92. **Miller AD, Bergholz U, Ziegler M, Stocking C.** 2008. Identification of the myelin protein plasmolipin as the cell entry receptor for *Mus caroli* endogenous retrovirus. J Virol **82:**6862-6868.

93. **Wolgamot G, Bonham L, Miller AD.** 1998. Sequence analysis of Mus dunni endogenous virus reveals a hybrid VL30/gibbon ape leukemia virus-like structure and a distinct envelope. J Virol **72:**7459-7466.

94. **Bromham L, Clark F, McKee JJ.** 2001. Discovery of a novel murine type C retrovirus by data mining. J Virol **75:**3053-3057.

95. **Weiss RA.** 2015. What's the host and what's the microbe? The Marjory Stephenson Prize Lecture 2015. J Gen Virol **96:**2501-2510.

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

97. **Simmons G, Clarke D, McKee J, Young P, Meers J.** 2014. Discovery of a novel retrovirus sequence in an Australian native rodent (*Melomys burtoni*): a putative link between gibbon ape leukemia virus and koala retrovirus. PLoS One **9:**e106954.

98. **Bryant LM, Donnellan SC, Hurwood DA, Fuller SJ.** 2011. Phylogenetic relationships and divergence date estimates among Australo-Papuan mosaic-tailed rats from the Uromys division (Rodentia: Muridae). Zoologica Scripta **40:**433-447.

99. **Simmons G, Meers J, Clarke DTW, Young PR, Jones K, Hanger JJ, Loader J, McKee JJ.** 2014. The origins and ecological impact of koala retrovirus. *In* Pye GW, Johnson RN, Greenwood AD (ed.), The Koala and its Retroviruses: Implications for Sustainability and Survival. Technical Reports of the Australian Museum (online), The Australian Museum, Sydney, Australia.

100. **Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P.** 2008. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev **72:**457-470.

101. **Overbaugh J, Miller AD, Eiden MV.** 2001. 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 **65:**371-389, table of contents.

102. **Battini JL, Danos O, Heard JM.** 1995. Receptor-binding domain of murine leukemia virus envelope glycoproteins. J Virol **69:**713-719.

103. **Coffin JM.** 2013. Virions at the gates: receptors and the host-virus arms race. PLoS Biol **11:**e1001574.

104. **Bottger P, Pedersen L.** 2002. Two highly conserved glutamate residues critical for type III sodium-dependent phosphate transport revealed by uncoupling transport function from retroviral receptor function. J Biol Chem **277:**42741-42747.

105. **Prassolov V, Hein S, Ziegler M, Ivanov D, Munk C, Lohler J, Stocking C.** 2001. *Mus cervicolor* murine leukemia virus isolate M813 belongs to a unique receptor interference group. J Virol **75:**4490-4498.

106. **O'Hara B, Johann SV, Klinger HP, Blair DG, Rubinson H, Dunn KJ, Sass P, Vitek SM, Robins T.** 1990. Characterization of a human gene conferring sensitivity to infection by gibbon ape leukemia virus. Cell Growth Differ **1:**119-127.

107. **Oliveira NM, Farrell KB, Eiden MV.** 2006. In vitro characterization of a koala retrovirus. J Virol **80:**3104-3107.

108. **Johann SV, Gibbons JJ, O'Hara B.** 1992. GLVR1, a receptor for gibbon ape leukemia virus, is homologous to a phosphate permease of *Neurospora crassa* and is expressed at high levels in the brain and thymus. J Virol **66:**1635-1640.

109. **Kavanaugh MP, Miller DG, Zhang W, Law W, Kozak SL, Kabat D, Miller AD.** 1994. Cell-surface receptors for gibbon ape leukemia virus and amphotropic murine retrovirus are inducible sodium-dependent phosphate symporters. Proc Natl Acad Sci U S A **91:**7071-7075.

110. **Schneiderman RD, Farrell KB, Wilson CA, Eiden MV.** 1996. The Japanese feral mouse Pit1 and Pit2 homologs lack an acidic residue at position 550 but still function as gibbon ape leukemia virus receptors: implications for virus binding motif. J Virol **70:**6982-6986.

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

112. **Xu W, Gorman K, Santiago JC, Kluska K, Eiden MV.** 2015. Genetic diversity of koala retroviral envelopes. Viruses **7:**1258-1270.

113. **Daugherty MD, Malik HS.** 2012. Rules of engagement: molecular insights from host-virus arms races. Annu Rev Genet **46:**677-700.

114. **Duggal NK, Emerman M.** 2012. Evolutionary conflicts between viruses and restriction factors shape immunity. Nat Rev Immunol **12:**687-695.

115. **Carpentier KS, Geballe AP.** 2016. An Evolutionary View of the Arms Race between Protein Kinase R and Large DNA Viruses. J Virol **90:**3280-3283.

116. **Nei M, Jin L.** 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol Biol Evol **6:**290-300.

117. **Pond SK, Muse SV.** 2005. Site-to-site variation of synonymous substitution rates. Mol Biol Evol **22:**2375-2385.

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

119. **Nielsen R, Yang Z.** 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:**929-936.

120. **Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K.** 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol **28:**3033-3043.

121. **Mardis ER.** 2008. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet **9:**387-402.

122. **Datta S, Budhauliya R, Das B, Chatterjee S, Vanlalhmuaka, Veer V.** 2015. Next-generation sequencing in clinical virology: Discovery of new viruses. World J Virol **4:**265-276.

123. **Metzker ML.** 2010. Sequencing technologies - the next generation. Nat Rev Genet **11:**31-46.

124. **Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al.** 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456:**53-59.

125. **Mertes F, Elsharawy A, Sauer S, van Helvoort JM, van der Zaag PJ, Franke A, Nilsson M, Lehrach H, Brookes AJ.** 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics **10:**374-386.

126. **Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ.** 2010. Target-enrichment strategies for next-generation sequencing. Nat Methods **7:**111-118.

127. **Foster JA, Bunge J, Gilbert JA, Moore JH.** 2012. Measuring the microbiome: perspectives on advances in DNA-based techniques for exploring microbial life. Brief Bioinform **13:**420-429.

128. **Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G, Licciulli F, Liuni S, Marzano M, Alonso-Alemany D, Valiente G, et al.** 2012. Reference databases for taxonomic assignment in metagenomics. Brief Bioinform **13:**682-695.

129. **Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, et al.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods **7:**335-336.

130. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol **75:**7537-7541.

131. **Panke-Buisse K, Poole AC, Goodrich JK, Ley RE, Kao-Kniffin J.** 2015. Selection on soil microbiomes reveals reproducible impacts on plant function. Isme j **9:**980-989.

132. **Hamdan LJ, Coffin RB, Sikaroodi M, Greinert J, Treude T, Gillevet PM.** 2013. Ocean currents shape the microbiome of Arctic marine sediments. Isme j **7:**685-696.

133. **Thomas AM, Gleber-Netto FO, Fernandes GR, Amorim M, Barbosa LF, Francisco AL, de Andrade AG, Setubal JC, Kowalski LP, Nunes DN, et al.**

2014. Alcohol and tobacco consumption affects bacterial richness in oral cavity mucosa biofilms. BMC Microbiol **14:**250.

134. **Hemme CL, Tu Q, Shi Z, Qin Y, Gao W, Deng Y, Nostrand JD, Wu L, He Z, Chain PS, et al.** 2015. Comparative metagenomics reveals impact of contaminants on groundwater microbiomes. Front Microbiol **6:**1205.
135. **Henderson G, Cox F, Ganesh S, Jonker A, Young W, Janssen PH.** 2015. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep **5:**14567.
136. **Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, et al.** 2008. Evolution of mammals and their gut microbes. Science **320:**1647-1651.
137. **Maricic T, Whitten M, Paabo S.** 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One **5:**e14004.
138. **Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P.** 2014. Library construction for next-generation sequencing: overviews and challenges. Biotechniques **56:**61-64, 66, 68, passim.
139. **Mason VC, Li G, Helgen KM, Murphy WJ.** 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Res **21:**1695-1704.
140. **Barzon L, Lavezzo E, Militello V, Toppo S, Palu G.** 2011. Applications of next-generation sequencing technologies to diagnostic virology. Int J Mol Sci **12:**7861-7884.
141. **Capobianchi MR, Giombini E, Rozera G.** 2013. Next-generation sequencing technology in clinical virology. Clin Microbiol Infect **19:**15-22.
142. **Quinones-Mateu ME, Avila S, Reyes-Teran G, Martinez MA.** 2014. Deep sequencing: becoming a critical tool in clinical virology. J Clin Virol **61:**9-19.
143. **Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N.** 2012. Application of next-generation sequencing technologies in virology. J Gen Virol **93:**1853-1868.
144. **Vinner L, Mourier T, Friis-Nielsen J, Gniadecki R, Dybkaer K, Rosenberg J, Langhoff JL, Cruz DF, Fonager J, Izarzugaza JM, et al.** 2015. Investigation of Human Cancers for Retrovirus by Low-Stringency Target Enrichment and High-Throughput Sequencing. Sci Rep **5:**13201.
145. **Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, Naeem A, Zaidi S, Delwart E.** 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J Virol **83:**4642-4651.

146. **Towner JS, Sealy TK, Khristova ML, Albarino CG, Conlan S, Reeder SA, Quan PL, Lipkin WI, Downing R, Tappero JW, et al.** 2008. Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog **4:**e1000212.

147. **Depledge DP, Palser AL, Watson SJ, Lai IY, Gray ER, Grant P, Kanda RK, Leproust E, Kellam P, Breuer J.** 2011. Specific capture and whole-genome sequencing of viruses from clinical samples. PLoS One **6:**e27805.

148. **Duncavage EJ, Magrini V, Becker N, Armstrong JR, Demeter RT, Wylie T, Abel HJ, Pfeifer JD.** 2011. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. J Mol Diagn **13:**325-333.

Chapter II

Variation in koala microbiomes within and between individuals: effect of body region and captivity status

Published in *Scientific Reports*

http://dx.doi.org/10.1038/srep10189

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Variation in koala microbiomes within and between individuals: effect of body region and captivity status

Received: 02 December 2014
Accepted: 25 March 2015
Published: 11 May 2015

Niccoló Alfano[1], Alexandre Courtiol[1], Hanna Vielgrader[2], Peter Timms[3], Alfred L. Roca[4] & Alex D. Greenwood[1]

Metagenomic analysis of 16S ribosomal RNA has been used to profile microbial communities at high resolution, and to examine their association with host diet or diseases. We examined the oral and gut microbiome composition of two captive koalas to determine whether bacterial communities are unusual in this species, given that their diet consists almost exclusively of *Eucalyptus* leaves. Despite a highly specialized diet, koala oral and gut microbiomes were similar in composition to the microbiomes from the same body regions of other mammals. Rectal swabs contained all of the diversity present in faecal samples, along with additional taxa, suggesting that faecal bacterial communities may merely subsample the gut bacterial diversity. Furthermore, the faecal microbiomes of the captive koalas were similar to those reported for wild koalas, suggesting that captivity may not compromise koala microbial health. Since koalas frequently suffer from ocular diseases caused by *Chlamydia* infection, we also examined the eye microbiome composition of two captive koalas, establishing the healthy baseline for this body part. The eye microbial community was very diverse, similar to other mammalian ocular microbiomes but with an unusually high representation of bacteria from the family Phyllobacteriaceae.

The koala, *Phascolarctos cinereus*, is an arboreal marsupial that has a unique diet consisting almost exclusively of *Eucalyptus* sp. leaves. Eucalyptus foliage has been described as an "unpromising" dietary source, low in nutrients and proteins but at the same time rich in oils and secondary plant compounds, such as lignin, cellulose and tannins, which are toxic to most animals[1,2]. Koalas have evolved a set of behavioral, physiological, morphological and metabolic adaptations to such a diet[3]. For example, they have a specialized digestive tract with an extremely enlarged caecum[4] and very long retention times of food within the gut[5]. Koalas can thus break down plant material by fermentation and enzymatic degradation, and finally extract sufficient nutrients to maintain active metabolism. Bacteria are thought to play an important role in this process. Several different microorganisms that are able to degrade lignin and tannins have been isolated from the koala gastrointestinal tract[6,7]. However, whether such an exclusive diet influences the composition of koala bacterial communities, or microbiomes is unknown.

Recent developments in culture-independent methods based on large-scale comparative analyses of 16S ribosomal RNA and metagenomics have the potential to profile microbial communities at high resolution even in complex environments like the intestinal microbiota[8]. Such methods have therefore been employed to study how the composition of bacterial communities relates to the diet in several species[9–11]. For certain organisms such as humans and mice[12,13], the relationship between diet and microbiome can

[1]Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany. [2]Tiergarten Schönbrunn, Vienna, Austria. [3]University of the Sunshine Coast, Sippy Downs, Queensland, Australia. [4]Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. Correspondence and requests for materials should be addressed to A.D.G. (email: greenwood@izw-berlin.de)

be directly studied by modifying the diet of some individuals and assessing how the microbiome is being influenced by such change. This experimental approach has the benefit of effectively isolating the influence that diet exerts upon the microbiome from the influence of many other factors known to impact the microbiome. Unfortunately, this approach cannot be applied to koalas because of their extreme dietary specialization. Instead, fully assessing the extent to which koala's microbiome is specific to its unique diet requires profiling microbial gut communities in a representative sample of koalas and comparing the profiles to those of other animals.

Because wild koala samples can be difficult to acquire and invasive sampling of captive koalas is discouraged, defining an effective sampling strategy is essential. A recent study[14] employed high-throughput GS FLX pyrosequencing to describe the composition of the koala microbiome across the hindgut in two wild koalas. This demonstrated that the koala hindgut microbiome is a complex and diverse environment and that the bacterial communities vary considerably in different regions of the intestine. However it is unclear whether the samples are representative of the entire gut, and whether or not widely used non-invasive samples such as faeces would provide an accurate representation of host microbiome. To the best of our knowledge, there have been no comparative studies based on high-throughput sequencing addressing whether rectal swabs and faecal samples yield consistent results in wild mammal gut microbiome research. Therefore, whether faecal samples are a good proxy to profile the gut microbiome in mammals in general, and in koalas in particular, remains to be determined.

The microbiome is known to vary both among individuals and among populations living in different environments. For example, shifts in gut microbiome composition between wild and captive individuals have been highlighted in several mammalian species, such as primates[15,16], goats[17], red pandas[18] and giant pandas[8]. The microbiome differences may be a consequence of the artificial nature of the zoo environment, particularly dietary changes. Thus, whether or not captive koalas can be used to study the diet specialization of the microbiome remains to be established.

For this study, rectal swabs and faeces were sampled from two captive koalas from the Tiergarten Schönbrunn in Vienna (Austria): a 14 year old male (SN241) and a 12 year old female (SN265). Although previous studies focused on the gut microbiome, initial digestion takes place in the mouth and thus koalas could be expected to have a unique microbiome in this compartment. Therefore, we also obtained oral swabs from the two koalas. Moreover, we sampled the eye microbiome of our two captive koalas. This body region was included to obtain a comparison point independent to digestion associated organs and to establish a baseline for the microbiome of healthy koala eyes, since wild and captive koalas frequently suffer from ocular infections caused by the highly prevalent *Chlamydia*, which is regarded as the primary disease threat to the species[19,20]. Such high incidence of chlamydial infection has been correlated with the presence of the recently described koala retrovirus (KoRV)[21]. All samples were characterized by 16S rRNA high-throughput Illumina sequencing. From a total of 1,956,592 quality-filtered reads, we identified 7,843 operational taxonomic units (OTUs) defined at the 97% similarity level. We first described the bacterial communities from the different body parts and compared our results with microbiome studies on other mammalian species. Second, we compared the gut microbiome profiles obtained by high-throughput sequencing from rectal swabs and faecal samples and discuss the reliability of using faeces as non-invasive sampling method in microbiome studies. Finally, we assessed whether captivity plays a role in shaping the koala faecal microbiome by comparing results from captive animals with existing data on wild koalas.

Results & Discussion

General microbiome characteristics. The composition of microbial communities was similar between samples when analyzed at low taxonomic resolution. Bacteroidetes (6.08–87.64%), Firmicutes (0.81–63.61%) and Proteobacteria (0.40–76.56%) were the most abundant phyla across most of the samples, followed by Actinobacteria and Fusobacteria (Fig. 1; Suppl. Fig. 1; Suppl. Tab. 1). At a higher taxonomic resolution, the different parts of the koala gastrointestinal tract and the koala eye were characterized by distinct bacterial communities (Fig. 2A). Indeed, principal coordinate analysis of unweighted UniFrac distances, which measure the similarity between bacterial communities based on phylogenetic distances, showed that koala microbial communities clustered by body region (Fig. 2A; Suppl. Fig. 2A). No clustering pattern was evident based on the weighted PCoA (Suppl. Fig. 2B), but the unweighted PCoA was consistent with the UPGMA tree (Fig. 2B). Irrespective of the method, no clustering pattern based on koala individual was observed (Fig. 2A; Suppl. Fig. 2B). These finding were further validated by a permutational MANOVA analysis on the unweighted UniFrac distance matrix, which showed that the body region significantly influenced the similarity among our samples ($F = 2.7$; 10,000 permutations; $p = 0.02$), while the individual did not ($F = 0.75$; 10,000 permutations; $p = 0.60$). Furthermore we estimated the robustness of the UPGMA tree by jackknife and confirmed the clustering of the samples according to body region, with the faecal and the eye samples clustering with maximal jackknife support (100%), while the nodes regarding the rectal and the oral samples being less resolved (Fig. 2B). Thus, the microbiomes from the same body region were more similar across individuals than microbiomes from different body regions of the same koala, which is consistent with human microbiome studies[22]. Replication of this study on a larger number of koalas would be needed to explore variation in microbiome driven by factors other than body location. Nonetheless, as our samples included one male and one

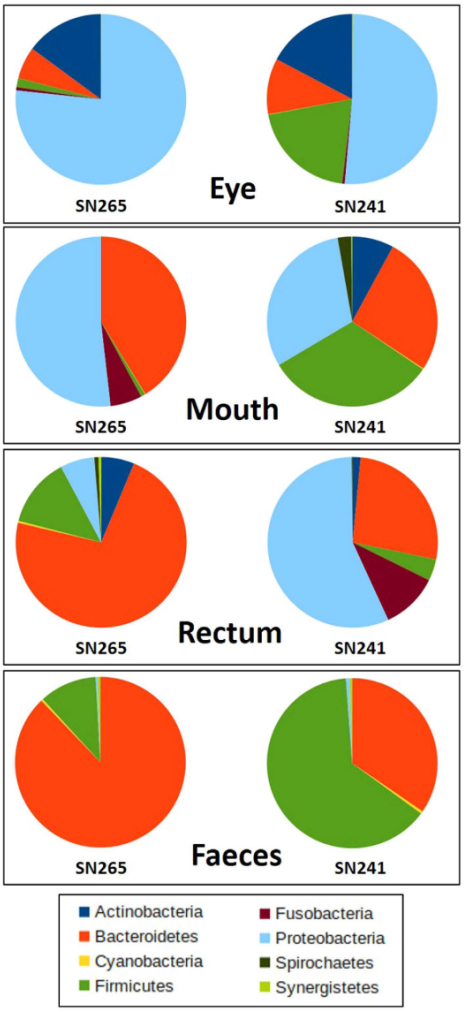


Figure 1. Relative abundance of bacterial phyla. Pie chart representation of the relative abundances of the most common phyla found in the eye, mouth, rectum and faeces of the two koalas. The relative abundance values of each phylum for each sample are reported in Suppl. Tab. 1.

female koala, the lack of significant difference between individuals suggests that sex does not strongly influence the koala microbiome.

Eye microbiome. This is the first study describing the composition of the eye microbiome of a non-human mammal by high-throughput sequencing. We found that the koala eye microbiome was generally similar to that of humans. The eye community had the highest biodiversity among our samples as assessed by the number of OTUs, Phylogenetic Distance (PD) and Shannon index, and low Evenness (Suppl. table 2; Supp Fig. 3; Suppl. table 3). This implies that koala eyes are characterized by a diverse microbial community with a relatively small number of very abundant genera, similar to humans[23]. Furthermore, the community profile at the phylum level was similar between the two koala ocular samples, with representatives of Proteobacteria (76.6 and 51.1%, for SN265 and SN241, respectively) and Actinobacteria (14.9–17.2%) reaching high abundance (Fig. 1; Supp Fig. 1; Suppl. Tab. 1),

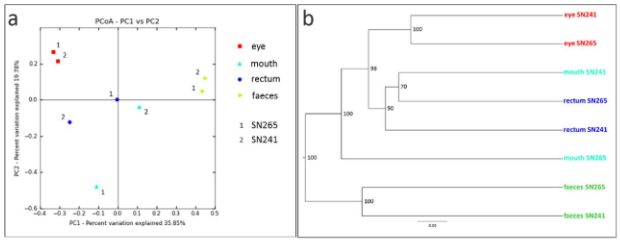


Figure 2. Principal Coordinate Analysis (PCoA) and UPGMA tree of Unifrac distances of the eye, mouth, rectal and faecal samples of two captive koalas. Panel A represents a 2 dimensional unweighted PCoA plot by sample type. Panel B represents the UPGMA clustering on unweighted Unifrac distances. Nodal supports were assessed by jackknife using 1,000 replicates. The two plots highlight that the samples tended to cluster by body region but not by koala specimen.

which is consistent with the few existing human studies[23,24]. At the genus level, ocular communities were rich in *Corynebacterium* and *Bradyrhizobium* (Fig. 3; Supp Fig. 4; Suppl. Tab. 4), which were found as common ocular bacteria in humans[23,24]. Nevertheless, 35 to 55% of all sequences from the koala eye were represented by uncultured bacteria from the family Phyllobacteriaceae, a group never described before in the eye. In humans it has been reported that the cultivable microbiota of the ocular surface is at a lower proportion than at many other mucosal sites (e.g. the oral cavity) suggesting that ocular communities harbor a hidden microbial diversity[25]. Our findings demonstrated that high-throughput culture-independent analysis of the ocular microbiome has the potential to unravel such diversity. Taking advantage of this methodology, our study sets a baseline for the koala eye microbiome to which microbiomes of diseased states can be compared. Keratoconjunctivitis indeed is one of the main consequences of the highly prevalent *Chlamydia* infection in koalas[19,20] and is likely to be the result of *Chlamydia* interplay with the resident bacteria constituting this complex and diverse microbiota.

Oral microbiome. In contrast to the eye, the oral microbiome has been well characterized in several other mammalian species and our results show that the composition of the oral microbial community in koalas shares several common features with other mammalian species, including herbivores (wallabies), omnivores (pigs, apes and humans) and carnivores (dogs). Together, Proteobacteria and Bacteroidetes accounted for over 90% of the detected bacteria in SN265 and 56% in SN241, with the remaining belonging mainly to the phylum Firmicutes (31.64%) (Fig. 1; Suppl. Fig. 1; Suppl. Tab. 1). These three phyla were also the main components of the oral microbiome in the majority of other mammalian species. The high abundance of Proteobacteria (30.4–50.9%) that we detected is consistent with previous reports in tammar wallabies[26], pigs[27], great apes and humans[16]. Bacteroidetes were also abundant (26.1–40.5%) as found in canine[28,29] and human studies[30,31]. Firmicutes were only abundant in SN241 (31.64%), consistently with human, pig, dog and wallaby microbiomes (17.8–52.3%)[26–28,30–33]. The koala oral samples presented low microbial diversity according to alpha diversity measures (Suppl. Fig. 3; Suppl. Tab. 2; Suppl. Tab. 3). In this respect, the data contrasts with the human oral cavity which was found to have the highest OTU richness and phylogenetic diversity within the gastrointestinal tract[34]. At the genus level, qualitative differences were noticeable between the two koala individuals (Fig. 3; Suppl. Fig. 4; Suppl. Tab. 4). However, the main genera present in each individual have been described in other mammalian species. For example, the majority of microbes were members of *Actinobacillus* and *Moraxella* in individual SN265, which are common oral bacteria in wallabies[26], pigs[27] and dogs[29]. *Flavobacterium*, which also had high abundance in the same koala (SN265), belongs to the family *Flavobacteriaceae* which includes typical inhabitants of the mammalian oropharyngeal flora (e.g. *Capnocytophaga*)[35]. However, in individual SN241, *Campylobacter*, which is a signature of the human oral microbiota[34], was the most dominant genus. SN241 exhibited a high abundance of *Porphyromonas*, a resident oral bacteria in dogs[28] and humans[33], *Lactobacillus* and *Clostridiales*, also found in pigs and humans[32]. Therefore, the koala oral microbiota does not appear to exhibit unique microbial community structure, despite the diet of Eucalyptus foliage unique to the species.

Rectal microbiome. The rectal swabs exhibited major differences between the two koalas with SN265 yielding a profile consistent at the phylum level with the few other gut microbiome studies based on the same sample type. The profile of SN265 was dominated by Bacteroidetes (72.0%) and Firmicutes (13.1%), followed by Proteobacteria (6.46%) and Actinobacteria (6.25%) (Fig. 1; Suppl. Fig. 1; Suppl. Tab. 1), similarly to wallabies[36] and humans[34,37]. In contrast, Bacteroidetes (26.6%) and Firmicutes (4.1%) were less common in SN241 than in SN265, while Proteobacteria (56.5%) and Fusobacteria (10.79%) were more common (Fig. 1; Suppl. Fig. 1). The most abundant genera in SN265 were *Bacteroides*, *Parabacteroides*

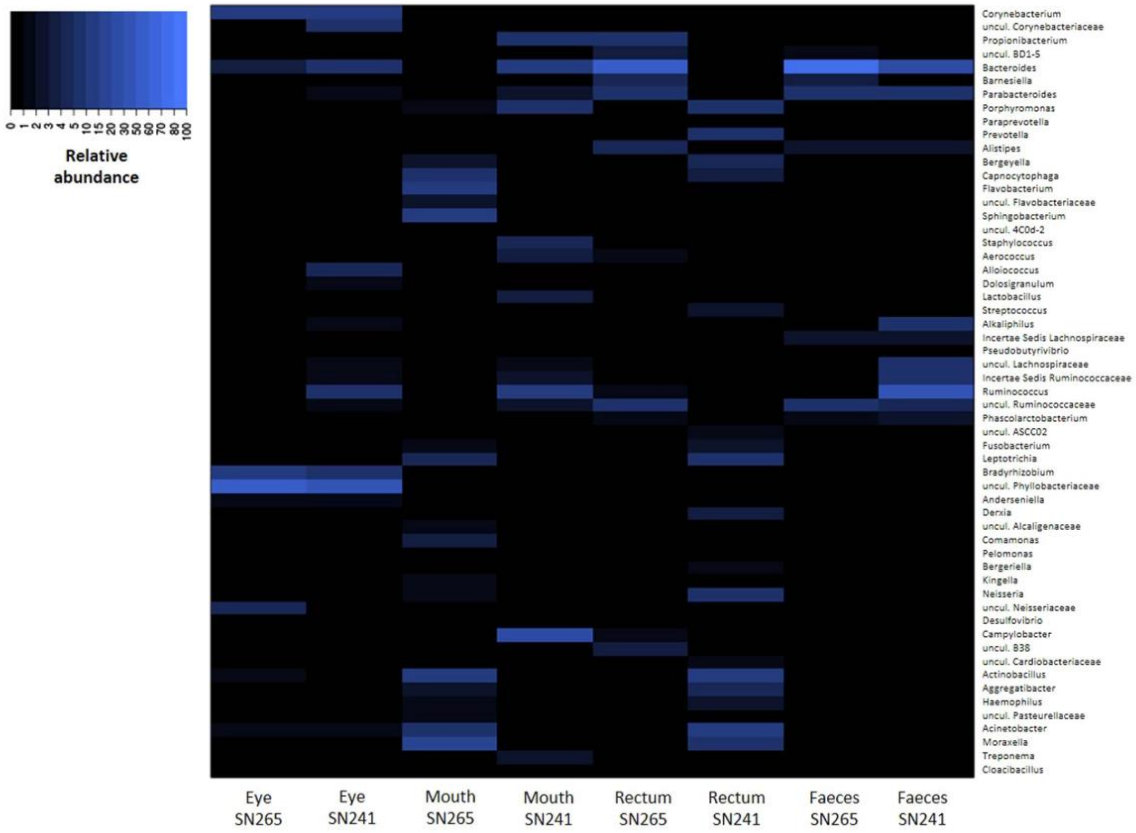


Figure 3. Heatmap analysis of the most abundant bacterial genera detected across all samples. The heatmap depicts the relative percentage of 16S rRNA gene sequences assigned to each bacterial genus (y axis) across the 8 samples analysed (x axis). The heatmap colors represent the relative percentage of the microbial genus assignments within each sample. Square colors shifted towards bright blue indicate higher abundance. The relative abundance values of each genus for each sample are reported in Suppl. Tab. 4.

and *Ruminococcaceae* (Fig. 3; Suppl. Fig. 4; Suppl. Tab. 4). *Bacteroides* accounts for approximately 25% of the bacterial population of the human gastrointestinal tract[38] and has been detected, together with *Parabacteroides* and *Ruminococcus*, in the colon of wild koalas as well[14]. The profile of SN241 was dominated by *Prevotella*, *Porphyromonas*, *Acinetobacter* and *Actinobacillus* (Fig. 3; Suppl. Fig. 4; Suppl. Tab. 4). *Prevotella* has also been associated with the human gut microbiome as the dominant group of type 2 enterotype replacing *Bacteroides* or *Ruminococcus* in some individuals[39]. *Porphyromonas* has been detected in wallaby anal opening[36], while *Acinetobacter* and *Actinobacillus* are human infectious agents. Despite the individual differences, the rectal microbial profiles of the two koalas overlapped in profiles consistent with those observed in other mammals.

Faecal microbiome. The faecal bacterial communities were dominated by Bacteroidetes (SN265 = 87.6%; SN241 = 34.5%) and Firmicutes (10.9%; 63.6%, respectively) in both koalas (Fig. 1; Suppl. Fig. 1; Suppl. Tab. 1), which is consistent with previously published results on faecal microbiomes of other mammalian species (Fig. 4; Suppl. Tab. 5). A similar composition has been observed in marsupials, including wallabies, kangaroos and also wild koalas[14,36,40]. Firmicutes are the most predominant phylum in the faecal microbiome of a wide range of mammalian species ranging from 9.4 to 95.4% relative abundance. Bacteriodetes usually occupy the second largest portion of gut microbial communities with abundances varying from 76.2% to 0.6% in bisons. Proteobacteria were detected at very low abundance as previously reported in wild koalas, kangaroos and wallabies, primates, dogs and cats, but in contrast to pandas, pygmy lorises, cows, bisons and chimpanzees, where this phylum represents 15.8 to 30.6% of the total. The relative abundance of Bacteroidetes and Firmicutes varied strongly among the two captive koalas (Fig. 1; Suppl. Fig. 1), but this variation has also been documented within and across other

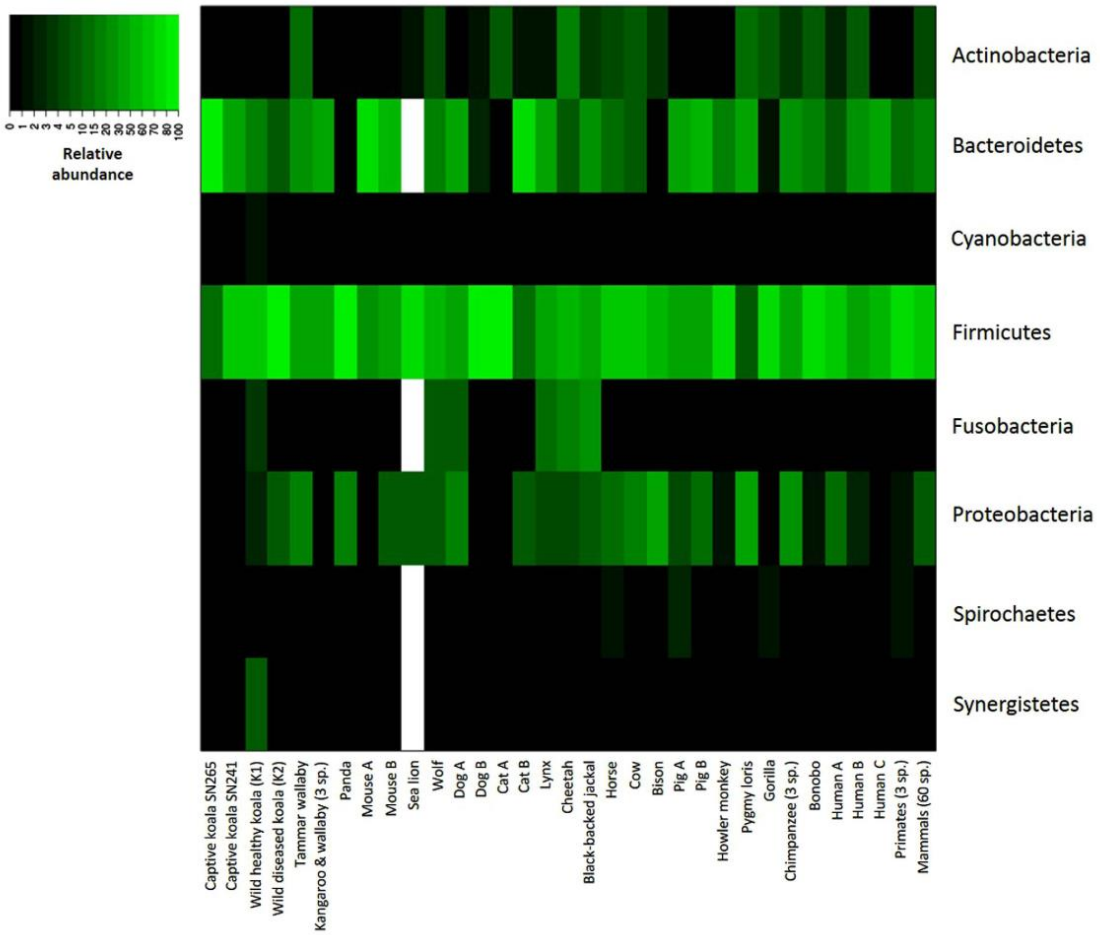


Figure 4. Heatmap analysis of the main bacterial phyla detected across mammalian species. The heatmap depicts the relative percentage of 16S rRNA gene sequences assigned to each bacterial phylum (y axis) across different mammalian species (x axis). The heatmap colors represent the relative percentage of the microbial phylum assignments within each species. Square colors shifted towards bright green indicate higher abundance. When a study reports average abundance values for more species, it is indicated how many species where used in the study. When more than one study per species is available, each study is indicated with a different letter. Blank squares correspond to NA, i.e. not available data, for those phyla for which the abundance values were not reported in the corresponding publication. The relative abundance values of each phylum for each species are reported in Suppl. Tab. 5.

mammalian species (Suppl. Tab. 5). Indeed, the Firmicutes to Bacteroidetes ratio (FB ratio) is generally close to 3:1 in mammals, but can change according to different variables such as host species, diet, age or sample type[9]. Here, individual SN265 presented the lowest FB ratio (FB = 0.12) (Suppl. Tab. 5), but was not very different from cats (FB = 0.17)[41]. In contrast, the FB ratio of individual SN241 (FB = 1.84) was well within the range of FB documented for other species and particularly close to the one found in tammar wallaby (FB = 1.48)[36].

Inter-individual differences in microbiomes. The origin of the inter-individual differences in microbiome composition observed in mouth, rectal and faecal samples is unclear, but such differences are also detected in other species[42]. For example, 70% of the phylotypes existing in the human gastro-intestinal microbiota have been shown to be subject-specific, with no phylotype being present at an abundance higher than 0.5% in all subjects[43]. This variation may result from competitive exclusion of phylotypes belonging to the same functional group which may select taxa differently depending on the

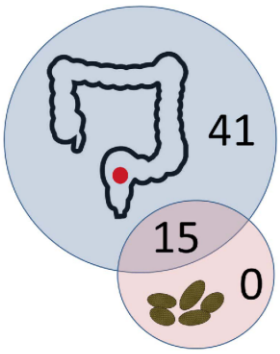


Figure 5. A Venn diagram showing the overlap between the rectal swab and the faeces of koala SN265 in the most abundant genera detected. While 15 genera were shared between rectal swab and faeces, and 41 genera were detected only in the rectal swab, there were no genera unique to the faeces.

internal environment[44]. Accordingly, human gut microbial communities exhibit a functional redundancy, such that even very different bacterial populations can achieve the same function[39]. Another possibility is an underlying but unobserved gastrointestinal pathology that may have altered the microbiome of one or both koalas although both were clinically healthy at the time of sampling. As the same extraction methods and identical independent triplicate PCR protocols were employed for all samples, the hypothesis of a purely methodological origin for the observed variation is unlikely.

Are faeces a good proxy for the gastrointestinal microbiome?. Faeces are the most commonly used sample type for gut microbiome investigation in mammals. This is the first study to directly compare the gut microbiome profiles obtained by high-throughput sequencing from faecal and rectal samples. We found that rectal and faecal bacterial communities were phylogenetically distinct based on weighted PCoA (Suppl. Fig. 2B), unweighted PCoA (Fig. 2A) and UPGMA (Fig. 2B). This result is consistent with a human study in which samples from the oral and digestive tract clustered strongly by gastrointestinal site and the multiple colonic samples (including rectal samples) were distinct from stool[34]. Our results contrast with those of Barker *et al.* 2013[14] where stool samples and colon content from a healthy wild koala were not distinct, but the higher sequencing depth of our study increases the sensitivity of the analysis.

For SN265 the rectal swab and the faeces exhibited similar microbiomes at the phylum level, which were dominated by Bacteroidetes (72 & 87.6%, for the abundance in rectum and faeces respectively) followed in abundance by Firmicutes (10.9 & 13.1%) (Fig. 1). In this individual, the rectal swab and faecal sample were thus correlated at the phylum (Spearman´s correlation test, $\rho = 0.66$, $p = 0.027$) and genus ($\rho = 0.61$, $p < 0.0001$) level when comparing the relative abundances of the most abundant phyla and genera (see methods for details). However, the microbial profiles exhibited by SN241 strongly differed between rectal and faecal samples: Proteobacteria were highly abundant (56.5%) in the rectal swab, but were almost absent (0.7%) in the faecal sample, which conversely was dominated by Firmicutes (63.6%). Firmicutes represented only a minor component in the rectal swab (4.1%). Bacteroidetes were similarly abundant between the two samples (26.6 & 34.5%) (Fig. 1). Accordingly, the profiles of SN241 showed no significant correlation at the phylum level ($\rho = 0.24$, $p = 0.48$) and a negative correlation was present at the genus level ($\rho = -0.40$, $p = 0.001$). These results are consistent with a previous study which observed that the composition of the microbiota changes as the gut contents are moved through the colon to the rectum and then excreted, with these changes attributed to differences in substrates, pH and water content[45].

The findings here suggest that the microbiota does not simply change, but may lose microbial diversity as it moves from the gut to faeces. Indeed, the rectal swabs had higher diversity than faecal samples according to alpha diversity (Suppl. Fig. 3; Suppl. Tab. 3). Furthermore, when each sample was examined for the presence/absence of the most abundant genera, only 27-34% of the genera that were found in the rectal swabs were also detected in the faecal samples (Suppl. Tab. 6). Accordingly, rectal swabs and faecal samples showed no significant similarity in both koalas when compared both at the phylum and genus level (Jaccard´s index = 0.27–0.6, $p > 0.05$) (Suppl. Table. 6). A majority (66–73%) of genera found in the rectal swabs were not found in the faeces. Conversely, all the genera that were found in the faeces were found in the rectal swabs, and thus there were no unique genera in the faecal samples (Suppl. Table. 6; Fig. 5). The pattern of presence/absence was exactly the same in both koalas, which was confirmed both at the genus and phylum level (Suppl. Tab. 6) without significant differences between the two koalas (Fisher´s exact test: $p = 0.477$ for genus level; $p \sim 1$ for phylum level). Therefore, according to our results, faeces represent only a subsample of the complex bacterial communities inhabiting the gut environment and caution should be used when faecal samples are used to investigate gut microbial diversity.

Are captive koalas a good proxy for wild microbiome?. The use of captive animals as representative of wild individuals has practical and logistical benefits, particularly for microbiome research where access to animal samples is facilitated in captivity. However, several lines of evidence suggest that different factors associated with captivity may interfere with gut microbiome composition. For example, obesity is known to cause shifts in gut microbiome composition in humans[46], mice[47] and is a possible consequence of captivity in zoos, where food is generally of high-quality and easily available, as reported in lemurs[48]. The artificial nature of the zoo environment can cause dietary and behavioural changes, for example in wide-ranging carnivores for which captivity constrains natural activities such as hunting and ranging, obesity can be a consequence.

In general, differences in gut microbiome composition of wild and captive individuals can be expected for species for which diet and activity patterns in captivity are markedly different than in nature. Differences between the microbiome of captive and wild animals are however less likely to happen for herbivorous species[49], as reported in studies comparing the gut microbiomes of wild and captive pandas[8] and of domestic and feral goats[17]. Accordingly, our results show that captivity does not appear to strongly influence the koala faecal microbiome at the phylum level. Wild koala gastrointestinal samples exhibited the same dominance of Firmicutes and Bacteroidetes detected in captive koalas with a FB ratio changing considerably across different areas of the hindgut but close to 3:1 in the faeces of the healthy individual[14]. In the diseased wild koala, Firmicutes were even more dominant. At the genus level the profiles of wild and captive koalas were also very similar. In both the present study and Barker *et al.* 2013[14], Bacteroidetes were mainly represented by *Bacteroides*, *Parabacteroides* and *Alistipes* (Fig. 3; Suppl. Tab. 4), which are common members of the microbiota of mammalian distal intestines. The percentages varied, especially for *Bacteroides*, in accordance with the higher levels of Bacteroidetes detected in one of the captive koalas. The majority of Firmicutes were identified as unknown (*Incertae sedis*) and uncultured Clostridiales. Many Firmicutes in the present study were assigned to the family Lachnospiraceae which is abundant in the digestive tracts of many mammals[50]. Except for the captive koala individual where Firmicutes had low abundances (SN265), *Ruminococcus* dominated. Not surprisingly this genus includes important cellulose-degrading species[51]. *Phascolarctobacterium*, which was originally isolated from koala faeces, but is also broadly distributed in human gastrointestinal tract as subdominant member[52], was also identified in both this study and Barker *et al.* 2013[14]. Consistently the faecal microbial profiles of the two wild and the two captive koalas were significantly correlated ($\rho = 0.64–0.94$, $p = 0.0001–0.033$) at the phylum level (Suppl. Tab. 7). The phyla presence/absence profiles were almost identical in the four different koalas with only 10 differences among the 66 possible pairwise comparisons between the four koalas (Suppl. Tab. 8) showing a very consistent bacterial community composition across the four different samples. Accordingly each pair of samples compared showed high significant similarity (Jaccard´s index = 0.8–1, $p < 0.05$) (Suppl. Tab. 9). Therefore we can conclude that captivity does not result in major alterations of koala gut microbiome compared to wild conditions when determined from faecal samples.

Our findings therefore suggest that koalas do not face diet related microbiome alterations in captivity. In zoos they are fed a diet almost identical to their natural one, which is based almost exclusively on Eucalyptus leaves. Koalas have evolved an adaptive flexibility that enables them to exploit various Eucalyptus species with a preference of about 50 different varieties out of over 800 existing ones. This diet is easily reproducible in zoos and koalas from this study, for example, were regularly fed 54 different species of Eucalyptus (personal communication of the zoo curator). We conclude that its unique diet, combined with a sedentary lifestyle facilitates koala nutritional management in captivity compared to other mammals and this is reflected in the similarity between wild and captive koala microbiomes.

Conclusion

The current study compared the microbial communities from multiple body regions of two captive koalas, including the eye and rectum, which are rarely described in the literature of mammal microbiome research. Therefore, getting a wider range of sample types per koala rather than a single sample type from multiple individuals was the priority. The results suggest that the koala eye microbiome is similar to that of other mammals though with some unique aspects observed. The oral and rectal microbiomes do not indicate any major shift in bacterial content that might be attributable to strong adaptive pressure from the Eucalyptus diet. However, the faecal microbiomes represented a subset of rectal microbial diversity suggesting the benefits of non-invasive samples such as faeces may be outweighed by the mixed gastrointestinal compartment origin of faecal bacteria. Nevertheless, captivity did not shift microbiome communities in koalas. Overall, we recommend future microbiome studies in koalas to be based on high-throughput sequencing applied to non-faecal samples. Due to the variation among individuals and the difficulty of obtaining wild koala samples, we suggest that analysis of captive individuals may be more appropriate for clarifying the numerous sources of koala microbial variation.

Materials and Methods

Koala Samples. Conjunctival, oral, rectal swabs and faecal samples were obtained from two captive koalas from the Tiergarten Schönbrunn in Vienna, Austria: Bilyarra (Pci-SN241; where "SN" is the studbook number), a 14 year old male and Mirra Li (Pci-SN265), a 12 year old female. The sample size was constrained by the small size of the captive koala population in Europe, by the fact that that most zoos have few koalas in their collections, and by the difficulty of collecting invasive samples, such as rectal

and conjunctival swabs. The two koalas were healthy with no pathological condition observed after blood tests, serological, parasitological and bacteriological examination, and had not received any antibiotic treatment for at least the previous 12 years. The faecal samples were collected immediately after defecation. Samples were stored in RNA*later*® solution at room temperature until processed.

DNA Extraction. Conjunctival, oral and rectal swabs were processed for DNA extraction using a QIAamp® DNA Mini kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. Genomic DNA was extracted from the faecal samples using a NucleoSpin® Tissue kit (Macherey-Nagel, Düren, Germany) following the protocol provided by the supplier. 50 mg of material were used of each faecal sample. DNA concentration was determined with a NanoDrop® (ND-1000) spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA).

Polymerase Chain Reaction. Universal primers 27F (5′-AGAGTTTGATCCTGGCTCAG-3′) and 338R (5′-TGCTGCCTCCCGTAGGAGT-3′)[53] were used for PCR amplification of the V1–V2 hypervariable regions of the bacterial 16S rRNA gene. Each sample was amplified in three replicate reactions to minimize stochastic PCR bias. Each 25 µl PCR reaction contained approximately 200 ng of DNA template, 12.5 µl MyTaq HS Mix (2x), 1.5 µl each primer (10 µM) and sterile distilled water to volume. The amplification conditions were as follows: 4 min of initial denaturation at 94 °C, followed by 16–26 cycles (according to the sample) of denaturation at 94 °C for 15 sec, annealing at 55 °C for 20 sec, and extension at 72 °C for 10 sec, with the last cycle followed by a 10 sec extension step at 72 °C. Water was used in the place of a DNA template as a negative control. After being visualized on a 1.5% agarose gel, the three replicate PCR products for each sample were pooled and purified using the MSB® Spin PCRapace kit (STRATEC Molecular GmbH, Berlin, Germany) according to manufacturer's protocol and eluted in 25 µl elution buffer. The PCR negative samples were also pooled and treated as a sample to monitor possible PCR contamination.

Illumina Library Preparation and Sequencing. Illumina sequencing libraries were generated as described in Meyer and Kircher 2010[54] with some modifications as described in Supplementary Methods. The libraries were first amplified in a 50 µl volume reaction using 5 µl of DNA library, 0.5 µl Herculase II Fusion DNA Polymerase (Agilent Technologies Inc.), 10 µl Herculase II Reaction Buffer (5x), 0.5 µl dNTPs (25 mM), 1 µl Single Index Primer P5 (10 µM), 1 µl Illumina Index Primer P7 (10 µM) and sterile distilled water to volume. A unique P7 Index Primer was used for each library to allow for subsequent sample discrimination after the sequencing of pooled libraries. Each library was amplified in three replicate reactions to minimize amplification bias in individual PCRs. PCR cycling conditions consisted of initial denaturation for 5 min at 95 °C, followed by 10 cycles of 30 sec denaturation at 95 °C, 30 sec annealing at 60 °C and 40 sec elongation at 72 °C. A final 7 min elongation step at 72 °C completed the reaction. The three replicate PCR products for each sample were pooled and purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) and eluted in 40 µl elution buffer. A negative control extraction library was also prepared and indexed separately to monitor any contamination introduced during the experiment.

Amplified libraries were quantified using the 2200 TapeStation (Agilent Technologies Inc.) on D1K ScreenTapes. The indexed DNA libraries were then pooled at equimolar concentrations for paired-end sequencing (2×250) on an Illumina MiSeq v2 platform at the Danish National High-Throughput DNA Sequencing Centre in Copenhagen, Denmark.

Ethics Statement. All experiments involving koala tissues were approved by the Internal Ethics Committee of the Leibniz Institute for Zoo and Wildlife Research, approval number 2012-09-06. All koala samples were obtained in accordance with the approved guidelines of the Leibniz Institute for Zoo and Wildlife Research and of Tiergarten Schönbrunn.

Bioinformatics and Statistical Analysis. A total of 2,584,237 paired-end sequence reads 250 bp long were generated (Suppl. Tab. 10) and then sorted by index sequences. 87.5% of paired-end reads were successfully merged reads into single reads. After primer and quality trimming, overall 2,064,872 sequences (91.2%) were retained. Quality trimmed reads were analysed using the Quantitative Insights Into Microbial Ecology (QIIME) pipeline software (version 1.6.0) (http://qiime.org). A further quality filtering step was performed using split_libraries_fastq.py command within the QIIME package to remove reads containing ambiguous bases.

Sequences were clustered into Operational Taxomonic Units (OTUs) based on 97% sequence similarity and the most abundant sequence within an OTU was chosen as the OTU's representative sequence. The representative sequences were then aligned and taxonomically classified against the SILVA reference database, release108 (SILVA 108; http://www.arb-silva.de). Chimeras were removed from the representative set, together with singletons and chloroplast sequences. Removal of these sequences left a total of 7,843 OTUs (Suppl. Tab. 10). OTUs with an abundance <0.1% of the total read count were removed from the OTU table to simplify the visualization of the results. This way lists of the "most abundant" phyla and genera were generated. Taxonomy summaries with relative abundance data at the phylum and genus

level were subsequently generated. More details about bioinformatic software and parameters used are available in the Supplementary Methods.

We calculated alpha and beta diversity metrics along with rarefaction plots from the complete OTU table using QIIME. The rarefaction curves tended to level off after approximately 100,000 reads demonstrating high coverage depth (Suppl. Fig. 5). Alpha diversity indices (within sample diversity) - Phylogenetic Distance, Shannon diversity index and Evenness - were calculated at a sequence depth of 161,378 reads/sample for 10,000 times and then averaged. The selected maximum sampling depth corresponded to the minimum number of quality reads obtained from any individual sample in the dataset. Phylogenetic distance (PD) is a measure of biodiversity that considers phylogenetic difference between species. Evenness measure how equally a community is numerically distributed among the species. Shannon diversity index (H) takes into account both abundance and evenness of species present in a community. Beta diversity (between samples diversity) was estimated by computing from the phylogenetic tree the unweighted and weighted UniFrac distances[55] between samples at the same sequence depth. UniFrac distances describe the dissimilarity among samples by assessing the evolutionary distances of bacterial phylotypes observed. Unweighted UniFrac only considers the presence/absence of taxa, while weighted UniFrac takes into account the differences in taxa abundance. UniFrac distance matrices were visualised using principal coordinates analysis (PCoA). The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) for clustering of samples was performed as an alternative hierarchical clustering method to interpret the Unifrac distance matrix. The robustness of the UPGMA tree was estimated using jackknife based on 1,000 replicates running the jackknifed_beta_diversity.py workflow in QIIME.

The results from QIIME were further analysed using the statistical software R version 3.1.0 (http://www.R-project.org). To test whether the body region or the individual had a significant influence on the similarity among our samples measured by unweighted Unifrac distance, we performed a permutational MANOVA using the function adonis of the package "vegan". We decided to focus only on the unweighted UniFrac because we considered it a safer measurement of similarity. Indeed, this distance is independent of abundance data and therefore less susceptible to variation due to methodology (e.g. PCR). Heatmaps were generated using the heatmap.2 function from the package "gplots". We calculated the mean and the standard deviation of the three alpha diversity metrics over the 10,000 iterations for each sample. We also measured the mean of the differences of the indices values among the four sample types for each koala across the 10,000 iterations and the 95% confidence intervals of the differences. To assess the similarity between rectal and faecal communities for each koala, the relative abundances of the most abundant phyla and genera detected in each sample were compared. For the comparison between the faecal communities of the captive koalas and of the wild koalas from Barker *et al.* 2013[14], the relative abundances of the eleven phyla detected in both studies (calculated from complete OTU lists) were used. The comparisons were performed using the Spearman's rank correlation coefficients (ρ) using the cor.test function in R. Unequal sampling depth could bias the correlation value because of the presence of many low abundance taxa in the samples. Indeed, those taxa may have been detected in samples characterized by higher number of reads set but not in those with less reads. We therefore subsampled (randomly, without replacement) the data sets to match the minimum number of reads in one of the samples so that each sample was represented by the same number of reads. Contingency tables with presence/absence data of the most abundant phyla and genera from the rectal swabs and faecal samples, and the phyla detected in the captive and wild koalas were created in R. Fisher´s exact test was performed in R in order to test if there was any significant difference between the two koalas and between phylum and genus level in the pattern of the distribution of the taxa among rectal swabs and faecal samples. We decided to compare only the phyla detected in our study with the ones detected in Barker *et al.* 2013[14], but not the genera because the different methods (extraction, PCR primers, sequencing) used in the two studies may not allow a comparison at such fine taxonomic resolution. Jaccard's coefficient[56] was also calculated to measure the similarity between the bacterial communities for each pair of samples compared. Jaccard's index was chosen since we decided not to count *double-zeros*, i.e. the absence of a taxon from two samples, to compute similarity. It ranges from 0 to 1, where 0 means no similarity between the two analysed samples and 1 complete similarity. To determine if the values for the index differed from what would be expected at random, we compared the observed similarity values with the table of statistical significance at $P = 0.05$ of lower and upper critical values[57], for the total number of taxa present in either of the two samples being compared.

References

1. Eberhard, I. H., McNamara, J., Pearse, R. J. & Southwell, I. A. Ingestion and excretion of *Eucalyptus punctata* DC and its essential oil by the Koala, *Phascolarctos cinereus* (Goldfuss). *Aust. J. Zool.* **23,** 169–179 (1975).
2. Cork, S. J., Hume, I. D. & Dawson, T. J. Digestion and metabolism of a natural foliar diet (*Eucalyptus punctata*) by an arboreal marsupial, the koala (*Phascolarctos cinereus*). *J. Comp. Physiol. B* **153,** 181–190 (1983).
3. Higgins, A. L., Bercovitch, F. B., Tobey, J. R. & Andrus, C. H. Dietary specialization and Eucalyptus species preferences in Queensland koalas (*Phascolarctos cinereus*). *Zoo Biol.* **30,** 52–58, doi:10.1002/zoo.20312 (2011).
4. McKenzie, R. A. The caecum of the Koala, *Phascolarctos cinereus*: Light, scanning and transmission electron microscopic observations on its epithelium and flora. *Aust. J. Zool.* **26,** 249–256 (1978).
5. Krockenberger, A. K. & Hume, I. D. A flexible digestive strategy accommodates the nutritional demands of reproduction in a free-living folivore, the Koala (*Phascolarctos cinereus*). *Funct. Ecol.* **21,** 748–756 (2007).
6. Osawa, R. Tannin-protein complex-degrading enterobacteria isolated from the alimentary tracts of koalas and a selective medium for their enumeration. *Appl. Environ. Microbiol.* **58,** 1754–1759 (1992).

7. Osawa, R. *et al.* E. Lonepinella koalarum gen nov., sp nov., a new tannin–protein complex degrading bacterium. *Syst. Appl. Microbiol.* **18,** 368–373 (1995).
8. Zhu, L., Wu, Q., Dai, J., Zhang, S. & Wei, F. Evidence of cellulose metabolism by the giant panda gut microbiome. *Proc. Natl. Acad. Sci. USA.* **108,** 17714–17719, doi:10.1073/pnas.1017956108 (2011).
9. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320,** 1647–1651, doi:10.1126/science.1155725 (2008).
10. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023, doi:10.1038/4441022a (2006).
11. Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332,** 970–974, doi:10.1126/science.1198719 (2011).
12. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. USA* **107,** 14691–14696, doi:10.1073/pnas.1005963107 (2010).
13. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1,** 6ra14, doi:10.1126/scitranslmed.3000322 (2009).
14. Barker, C. J., Gillett, A., Polkinghorne, A. & Timms, P. Investigation of the koala (*Phascolarctos cinereus*) hindgut microbiome via 16S pyrosequencing. *Vet. Microbiol.* **167,** 554–564, doi:10.1016/j.vetmic.2013.08.025 (2013).
15. Amato, K. R. *et al.* Habitat degradation impacts black howler monkey (*Alouatta pigra*) gastrointestinal microbiomes. *Isme J.* **7,** 1344–1353, doi:10.1038/ismej.2013.16 (2013).
16. Li, J. *et al.* The saliva microbiome of *Pan* and *Homo*. *BMC Microbiol.* **13,** 204, doi:10.1186/1471-2180-13-204 (2013).
17. De Jesus-Laboy, K. M. *et al.* Comparison of the fecal microbiota in feral and domestic goats. *Genes (Basel)* **3,** 1–18, doi:10.3390/genes3010001 (2011).
18. Kong, F. *et al.* Characterization of the gut microbiota in the red panda (*Ailurus fulgens*). *PLoS One* **9,** e87885, doi:10.1371/journal.pone.0087885 (2014).
19. Cockram, F. A. & Jackson, A. R. Keratoconjunctivitis of the koala, *Phascolarctos cinereus*, caused by Chlamydia psittaci. *J. Wildl. Dis.* **17,** 497–504 (1981).
20. Polkinghorne, A., Hanger, J. & Timms, P. Recent advances in understanding the biology, epidemiology and control of chlamydial infections in koalas. *Vet. Microbiol.* **165,** 214–223, doi:10.1016/j.vetmic.2013.02.026 (2013).
21. Tarlinton, R., Meers, J., Hanger, J. & Young, P. Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. *J. Gen. Virol.* **86,** 783–787, doi:10.1099/vir.0.80547-0 (2005).
22. Sonnenburg, J. L. & Fischbach, M. A. Community health care: therapeutic opportunities in the human microbiome. *Sci. Transl. Med.* **3,** 78ps12, doi:10.1126/scitranslmed.3001626 (2011).
23. Dong, Q. *et al.* Diversity of bacteria at healthy human conjunctiva. *Invest. Ophthalmol. Vis. Sci.* **52,** 5408–5413, doi:10.1167/iovs.10-6939 (2011).
24. Lee, S. H., Oh, D. H., Jung, J. Y., Kim, J. C. & Jeon, C. O. Comparative ocular microbial communities in humans with and without blepharitis. *Invest. Ophthalmol. Vis. Sci.* **53,** 5585–5593, doi:10.1167/iovs.12-9922 (2012).
25. Willcox, M. D. Characterization of the normal microbiota of the ocular surface. *Exp. Eye. Res.* **117,** 99–105, doi:10.1016/j.exer.2013.06.003 (2013).
26. Chhour, K. L., Hinds, L. A., Jacques, N. A. & Deane, E. M. An observational study of the microbiome of the maternal pouch and saliva of the tammar wallaby, *Macropus eugenii*, and of the gastrointestinal tract of the pouch young. *Microbiol.* **156,** 798–808, doi:10.1099/mic.0.031997-0 (2010).
27. Lowe, B. A. *et al.* Defining the "core microbiome" of the microbial communities in the tonsils of healthy pigs. *BMC Microbiol.* **12,** 20, doi:10.1186/1471-2180-12-20 (2012).
28. Dewhirst, F. E. *et al.* The canine oral microbiome. *PLoS One* **7,** e36067, doi:10.1371/journal.pone.0036067 (2012).
29. Sturgeon, A., Stull, J. W., Costa, M. C. & Weese, J. S. Metagenomic analysis of the canine oral cavity as revealed by high-throughput pyrosequencing of the 16S rRNA gene. *Vet. Microbiol.* **162,** 891–898, doi:10.1016/j.vetmic.2012.11.018 (2013).
30. Bik, E. M. *et al.* Bacterial diversity in the oral cavity of 10 healthy individuals. *Isme J.* **4,** 962–974, doi:10.1038/ismej.2010.30 (2010).
31. Keijser, B. J. *et al.* Pyrosequencing analysis of the oral microflora of healthy adults. *J. Dent. Res.* **87,** 1016–1020 (2008).
32. Ahn, J. *et al.* Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison. *PLoS One* **6,** e22788, doi:10.1371/journal.pone.0022788 (2011).
33. Zaura, E., Keijser, B. J., Huse, S. M. & Crielaard, W. Defining the healthy "core microbiome" of oral microbial communities. *BMC Microbiol.* **9,** 259, doi:10.1186/1471-2180-9-259 (2009).
34. Stearns, J. C. *et al.* Bacterial biogeography of the human digestive tract. *Sci. Rep.* **1,** 170, doi:10.1038/srep00170 (2011).
35. Jolivet-Gougeon, A., Sixou, J. L., Tamanai-Shacoori, Z. & Bonnaure-Mallet, M. Antimicrobial treatment of Capnocytophaga infections. *Int. J. Antimicrob. Agents* **29,** 367–373, doi:10.1016/j.ijantimicag.2006.10.005 (2007).
36. Chhour, K. L., Hinds, L. A., Deane, E. M. & Jacques, N. A. The microbiome of the cloacal openings of the urogenital and anal tracts of the tammar wallaby, *Macropus eugenii*. *Microbiol.* **154,** 1535–1543, doi:10.1099/mic.0.2007/014803-0 (2008).
37. Yu, G., Fadrosh, D., Ma, B., Ravel, J. & Goedert, J. J. Anal microbiota profiles in HIV-positive and HIV-negative MSM. *Aids* **28,** 753–760, doi:10.1097/qad.0000000000000154 (2014).
38. Xu, J., Chiang, H. C., Bjursell, M. K. & Gordon, J. I. Message from a human gut symbiont: sensitivity is a prerequisite for sharing. *Trends Microbiol.* **12,** 21–28 (2004).
39. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180, doi:10.1038/nature09944 (2011).
40. Gulino, L. M. *et al.* Shedding light on the microbial community of the macropod foregut using 454-amplicon pyrosequencing. *PLoS One* **8,** e61463, doi:10.1371/journal.pone.0061463 (2013).
41. Tun, H. M. *et al.* Gene-centric metagenomics analysis of feline intestinal microbiome using 454 junior pyrosequencing. *J. Microbiol. Methods* **88,** 369–376, doi:10.1016/j.mimet.2012.01.001 (2012).
42. Human Microbiome Project Consortium *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214, doi:10.1038/nature11234 (2012).
43. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484, doi:10.1038/nature07540 (2009).
44. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8,** e1002606, doi:10.1371/journal.pcbi.1002606 (2012).
45. Mai, V., Ukhanova, M. & Baer, D. J. Understanding the Extent and Sources of Variation in Gut Microbiota Studies; a Prerequisite for Establishing Associations with Disease. *Diversity* **2,** 1085–1096 (2010).
46. Ley, R. E. Obesity and the human microbiome. *Curr. Opin. Gastroenterol.* **26,** 5–11, doi:10.1097/MOG.0b013e328333d751 (2010).
47. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. USA.* **102,** 11070–11075, doi:10.1073/pnas.0504978102 (2005).
48. Schwitzer, C. & Kaumanns, W. Body weights of ruffed lemurs (*Varecia variegata*) in European zoos with reference to the problem of obesity. *Zoo Biol.* **20,** 261–269 (2001).

49. Delsuc, F. *et al.* Convergence of gut microbiomes in myrmecophagous mammals. *Mol. Ecol.* **23,** 1301–1317, doi:10.1111/mec.12501 (2014).
50. Meehan, C. J. & Beiko, R. G. A phylogenomic view of ecological specialization in the Lachnospiraceae, a family of digestive tract-associated bacteria. *Genome Biol. Evol.* **6,** 703–713, doi:10.1093/gbe/evu050 (2014).
51. Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R. & White, B. A. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat. Rev. Microbiol.* **6,** 121–131, doi:10.1038/nrmicro1817 (2008).
52. Watanabe, Y., Nagai, F. & Morotomi, M. Characterization of *Phascolarctobacterium succinatutens* sp. nov., an asaccharolytic, succinate-utilizing bacterium isolated from human feces. *Appl. Environ. Microbiol.* **78,** 511–518, doi:10.1128/aem.06035-11 (2012).
53. Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62,** 625–630 (1996).
54. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5448, doi:10.1101/pdb.prot5448 (2010).
55. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235, doi:10.1128/aem.71.12.8228-8235.2005 (2005).
56. Jaccard, P. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudoise Sci. nat.* **44,** 223–270 (1908).
57. Real, R. Tables of significant values of Jaccard's index of similarity. *Misc. Zoologica* **22,** 29–40 (1999).

Acknowledgments

The authors wish to thank Karin Hönig for technical support. This study was supported in part by funds from Grant Number R01GM092706 from the National Institute of General Medical Sciences (NIGMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the National Institutes of Health.

Author Contributions

N.A. and A.D.G. designed the project; H.V. collected and provided the samples; N.A. performed all laboratory experiments; N.A. and A.C analyzed the data; N.A., P.T., A.L.R. and A.D.G. discussed the results and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Alfano, N. *et al.* Variation in koala microbiomes within and between individuals: effect of body region and captivity status. *Sci. Rep.* **5**, 10189; doi: 10.1038/srep10189 (2015).

Supplementary material

Variation in koala microbiomes within and between individuals: effect of body region and captivity status

Published in *Scientific Reports*

http://dx.doi.org/10.1038/srep10189

Variation in koala microbiomes within and between individuals: effect of body region and captivity status.

Niccoló Alfano[1], Alexandre Courtiol[1], Hanna Vielgrader[2], Peter Timms[3], Alfred L. Roca[4], Alex D. Greenwood[1]*

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

[2] Tiergarten Schönbrunn, Vienna, Austria

[3] University of the Sunshine Coast, Sippy Downs, Queensland, Australia

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

* Correspondence to: Alex D. Greenwood, Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Str. 17, 10315 Berlin, Germany, +49 (0)30 5168255, greenwood@izw-berlin.de

Supplementary Methods

Illumina Library Preparation

Illumina sequencing libraries were generated as described in Meyer and Kircher 2010 with the following modifications. Blunt ending reactions were set up in a 50 µl volume containing 1 µg of pooled purified PCR product, 6 µl NEBuffer 2 (10x), 0.6 µl dNTPs (25mM), 7 µl ATP (10mM), 5 µl BSA (10mg/ml), 3 µl T4 Polynucleotide Kinase (10U/µl), 2 µl T4 DNA Polymerase (3U/µl) and sterile distilled water to volume. The reaction was incubated at 25°C for 15 min, followed by 15 min at 12°C. The blunt-ended DNA was then purified using the MinElute PCR Purification Kit (Qiagen, Hilden, Germany) and eluted in 20 µl elution buffer. Blunt-ended DNA was then ligated to Illumina multiplex adapters (Illumina Inc.)(5´-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT and 5´-AGA TCG GAA GAG C for one adapter, 5´-GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T and 5´-AGA TCG GAA GAG C for the other) in a 40 µl reaction containing 20 µl Quick Ligase Buffer (2x), 1 µl Quick Ligase (5U/µl) and 1 µl Illumina adapter mix (10 µM). The reaction was incubated at room temperature for 20 min. The adapter-ligated DNA was then purified using the MinElute PCR Purification Kit (Qiagen, Hilden, Germany) and eluted in 35 µl elution buffer. Finally the blunt-ended adapter-ligated DNA went through a 40 µl adapter fill-in reaction containing 4 µl Thermopol Buffer (10x), 1 µl dNTPs (25 mM) and 2 µl Bst Polymerase (8U/µl). The reaction was incubated at 65°C for 20 min, followed by 20 min at 80°C. All reagents used in library preparation were from New Englands Biolabs® Inc., Ipswich, MA, USA.

Bioinformatics

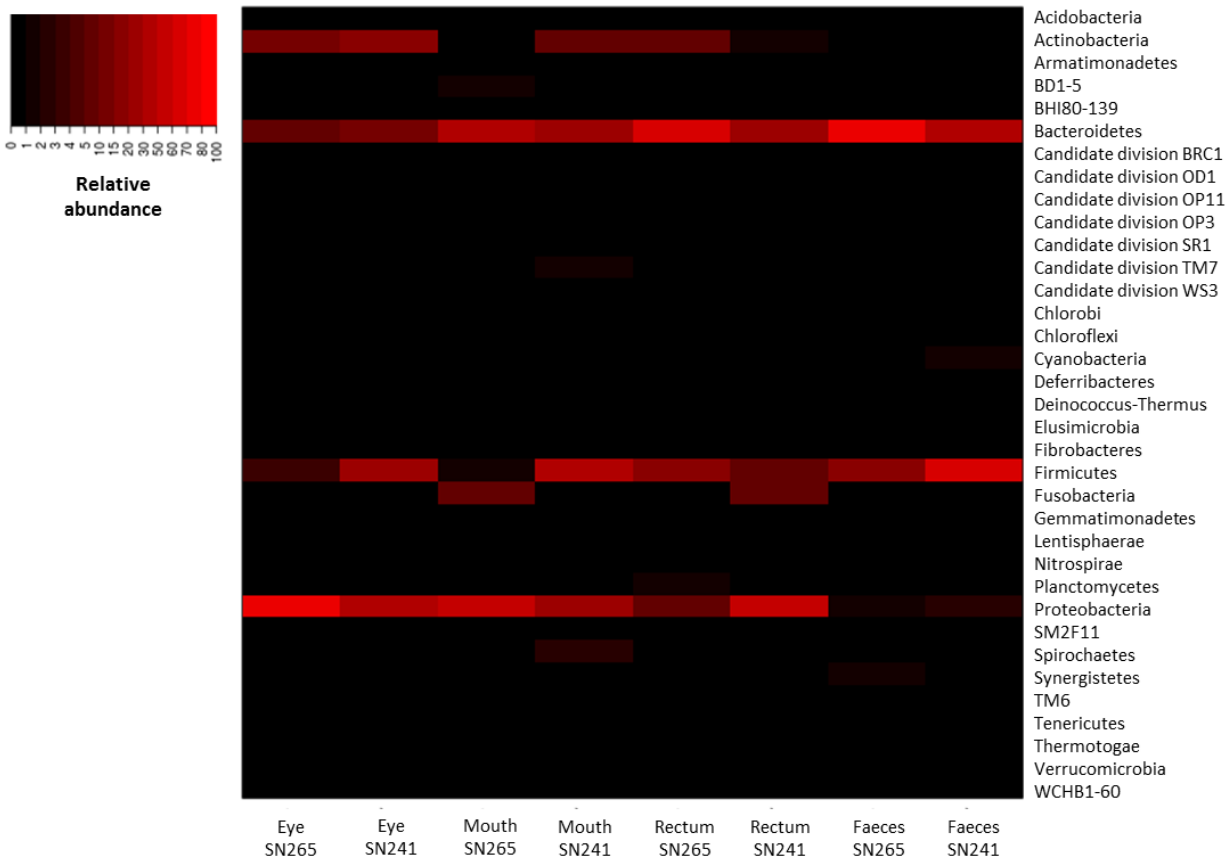
Paired-end reads were merged reads into single reads using the FLASH software tool (http://ccb.jhu.edu/software/FLASH/) using 10 bp as minimum overlap and 0.1 as maximum allowed ratio between the number of mismatched base pairs and the overlap length. Primers were removed from the reads using Cutadapt (https://code.google.com/p/cutadapt/): reads that did not contain the primers or with more than a mismatch in the primer sequence were discarded and reads

between 250–370 bp in length were retained. Low quality bases were trimmed using FASTX-Toolkit (FASTQ Quality Filter tool) (http://hannonlab.cshl.edu/fastx_toolkit/): only reads with at least 75% of read length with quality score above 30 were kept.

Within QIIME package, sequences were clustered using UCLUST into Operational Taxomonic Units (OTUs) based on 97% sequence similarity through open-reference OTU picking against the Greengenes database (version 12_10) (http://greengenes.lbl.gov/). The most abundant sequence within an OTU was chosen as the OTU's representative sequence. The representative sequences were then aligned against the 16S rRNA Greengenes core set using PyNAST with a minimum identity of 75%. Representative sequences were taxonomically classified using BLAST against the SILVA reference database, release108 (SILVA 108; http://www.arbsilva.de). The alignment was then filtered to remove gaps and a maximum-likelihood phylogenetic tree was constructed from the filtered alignment using FastTree. Chimeras were removed from the representative set using Chimera Slayer.

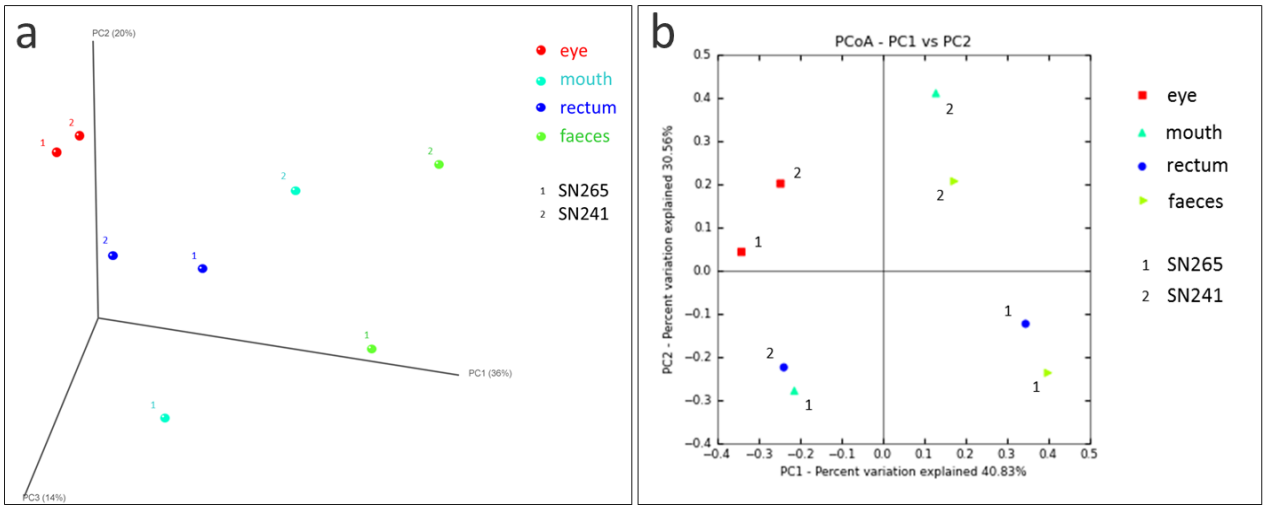
Supplementary Figures

Supplemental Figure 1



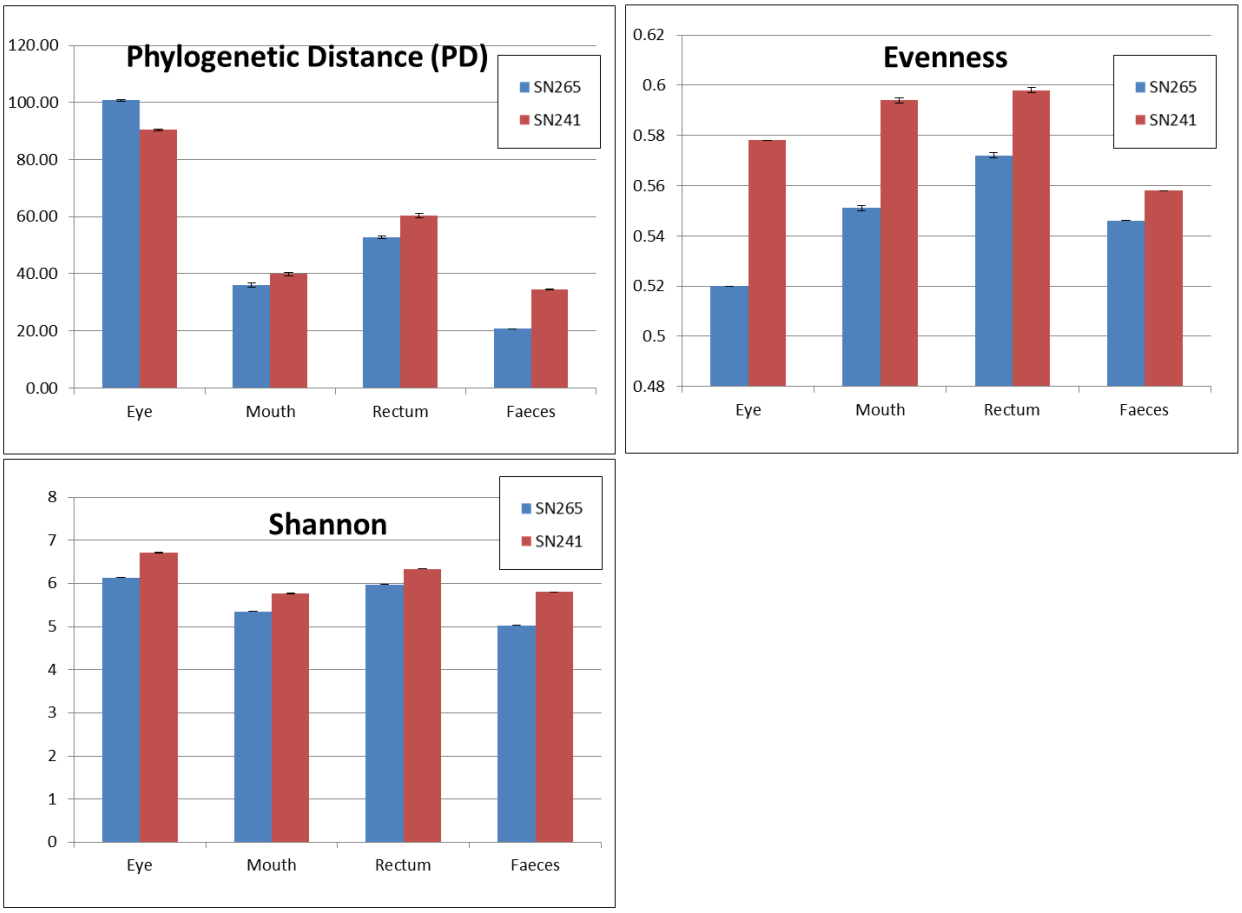
Supplemental Figure 1. Heatmap analysis of the complete list of bacterial phyla detected across all samples. The heatmap depicts the relative percentage of 16S rRNA gene sequences assigned to each bacterial phylum (y axis) across the 8 samples analysed (x axis). The heatmap colors represent the relative percentage of the microbial phylum assignments within each sample. Square colors shifted towards bright red indicate higher abundance.

Supplemental Figure 2



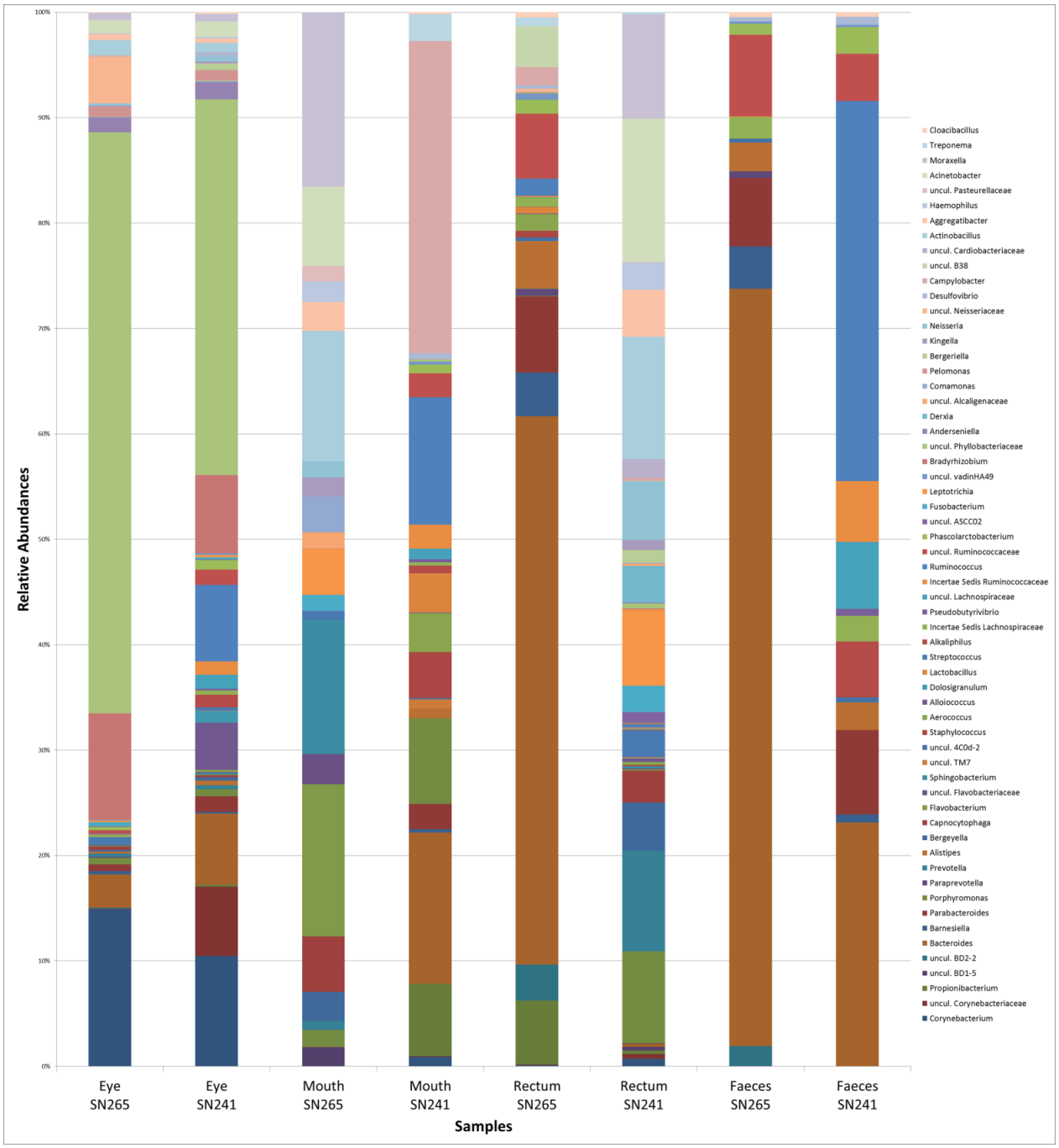
Supplemental Figure 2. Principal Coordinate Analysis (PCoA) of Unifrac distances of the eye, mouth, rectal and faecal samples of two captive koalas. Panel A is a 3 dimensional unweighted PCoA plot which is plotted by sample type. Panel B shows the 2 dimensional weighted PCoA plot by sample type.

Supplemental Figure 3



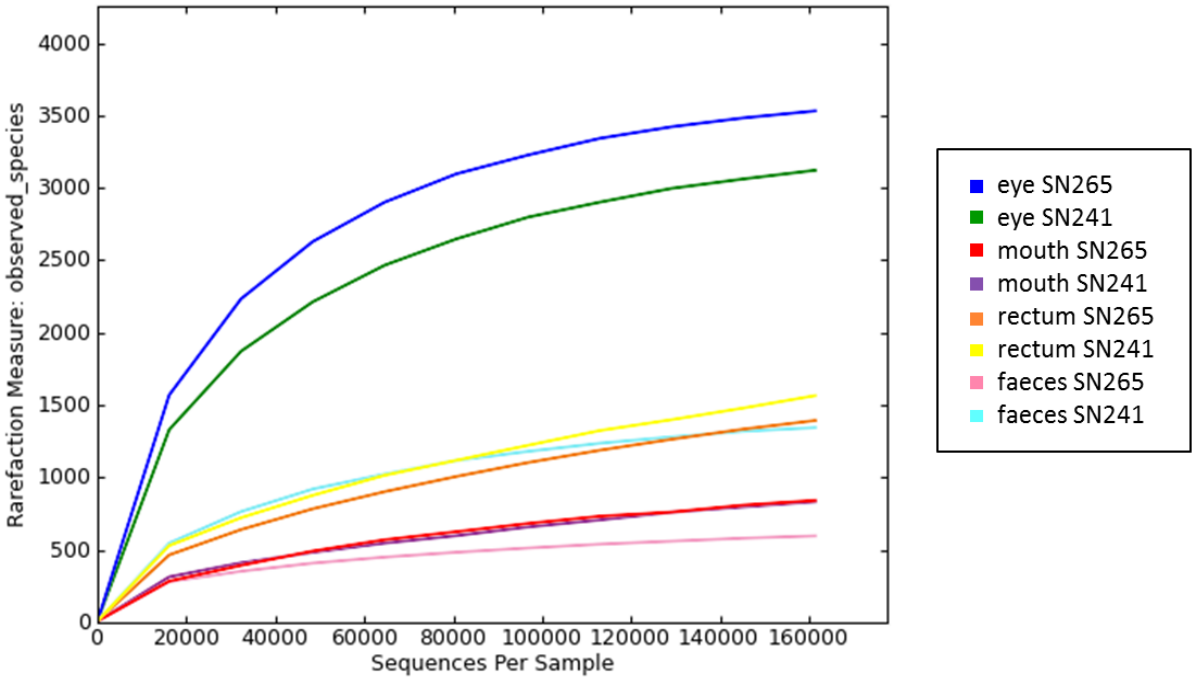
Supplemental Figure 3. Measures of alpha diversity of the eye, mouth, rectal and faecal bacterial communities of the two koalas. Phylogenetic distance (PD), Shannon diversity index and Evenness are shown in the respective panels. Each alpha diversity metric is presented as the mean value of the 10,000 iterations at the rarefaction depth of 161,378 sequences/sample. The error bars represent the standard deviations.

Supplemental Figure 4



Supplemental Figure 4. The most abundant bacterial genera found in each sample. Bar chart representations are shown of the relative abundances of the most abundant bacterial genera found in the eye, mouth, rectum and faeces of the two koalas in the present study.

Supplemental Figure 5



Supplemental Figure 5. Rarefaction analysis of the different samples sequenced. Rarefaction curves obtained from the eye, mouth, rectum and faeces of the two koalas showing the number of unique OTUs (observed species metric) as a function of sequencing depth for each sample. The curves were calculated at a maximum rarefaction depth of 161,378 sequences/sample.

Supplementary Tables

Suppl. Tab 1 Relative abundance of most abundant OTUs at phylum level in the eye, mouth, rectal and faecal microbiome from the two koalas.

	EYE		MOUTH		RECTUM		FAECES	
Phylum	SN265	SN241	SN265	SN241	SN265	SN241	SN265	SN241
Actinobacteria	14.89	17.23	0.00	7.81	6.25	1.50	-	-
BD1-5	0.02	0.01	1.80	-	-	0.35	-	-
Bacteroidetes	6.08	10.60	40.55	26.12	72.02	26.60	87.64	34.51
Candidate division TM7	-	-	-	0.86	0.02	0.02	-	-
Cyanobacteria	0.01	0.07	-	0.12	0.36	0.00	0.38	0.48
Firmicutes	1.69	20.14	0.81	31.65	13.07	4.08	10.88	63.61
Fusobacteria	0.62	0.48	5.92	0.00	0.02	10.79	-	-
Planctomycetes	0.05	0.14	0.00	0.28	0.54	0.01	0.24	0.22
Proteobacteria	76.56	51.13	50.91	30.44	6.46	56.46	0.40	0.75
Spirochaetes	0.00	0.07	0.00	2.54	0.77	0.18	-	-
Synergistetes	0.06	0.14	0.00	0.17	0.50	0.01	0.47	0.44

Data derived from OTU table where OTUs with an abundance <0.1% of the total read count were removed in order to semplify the visualization of the results.

Suppl. Tab. 2 Numbers of OTUs and measures of alpha diversity of the eye, mouth, rectal and faecal bacterial communities of the two koalas.

	EYE		MOUTH		RECTUM		FAECES	
	SN265	SN241	SN265	SN241	SN265	SN241	SN265	SN241
Number of OTUs	3592	3263	1111	1064	1531	2077	597	1381
Chao1	3192.35	2873.60	898.02	1081.17	1640.65	1866.05	577.52	1252.66
Phylogenetic diversity (PD)	83.49	74.22	26.34	29.60	39.85	45.03	16.89	27.44
Shannon (H)	5.80	6.35	5.13	5.50	5.69	6.03	4.83	5.54
Evenness (E_H)	0.569	0.628	0.618	0.659	0.637	0.662	0.604	0.615

Each alpha diversity metric - Phylogenetic distance, Shannon diversity index and Evenness - is presented as the mean value of the 10,000 iterations at the rarefaction depth of 161,378 sequences/sample.

Suppl. Tab. 3 The differences in each alpha diversity metric between the eye, mouth, rectal and faecal samples of each koala.

PD	EYE	MOUTH	RECTUM	FAECES	
EYE		64.71 63.07-66.38	47.91 46.92-48.98	80.07 79.55-80.46	SN265
MOUTH	50.45 49.02-51.89		16.8 15-18.62	15.36 13.74-16.91	
RECTUM	30.03 28.42-31.66	20.41 18.45-22.32		32.16 31.17-32.99	
FAECES	55.88 55.03-56.76	5.34 4.12-6.72	25.85 24.31-27.37		
	SN241				

Shannon	EYE	MOUTH	RECTUM	FAECES	
EYE		0.773 0.762-0.784	0.155 0.145-0.166	1.096 1.088-1.105	SN265
MOUTH	0.938 0.926-0.95		0.618 0.609-0.627	0.323 0.316-0.33	
RECTUM	0.369 0.357-0.381	0.569 0.58-0.558		0.941 0.935-0.946	
FAECES	0.903 0.893-0.914	0.035 0.026-0.044	0.534 0.525-0.544		
	SN241				

Evenness	EYE	MOUTH	RECTUM	FAECES	
EYE		0.031 0.029-0.034	0.052 0.05-0.053	0.026 0.025-0.026	SN265
MOUTH	0.017 0.014-0.019		0.02 0.018-0.023	0.006 0.004-0.008	
RECTUM	0.02 0.018-0.022	0.003 0-0.006		0.026 0.025-0.027	
FAECES	0.019 0.018-0.02	0.036 0.033-0.038	0.039 0.037-0.041		
	SN241				

The mean of the differences of the indices values (across the 10,000 iterations) among the eye, mouth, rectal and faecal bacterial communities of the two koalas for each of the three alpha diversity metrics measured - Phylogenetic distance (PD), Shannon diversity index and Evenness measure. Values above the diagonal concern individual SN265, while those below SN241. The 95% confidence intervals of the differences is given as well.

Suppl. Tab. 4 Relative abundance of most abundant OTUs at genus level in the eye, mouth, rectal and faecal microbiome from the two koalas.

PHYLUM	CLASS	ORDER	FAMILY	GENUS	EYE		MOUTH		RECTUM		FAECES	
					SN265	SN241	SN265	SN241	SN265	SN241	SN265	SN241
Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium	14.84	10.46	0.00	0.91	0.10	0.72	0.00	0.00
	Actinobacteria	Corynebacteriales	Corynebacteriaceae	uncultured	0.06	6.58	0.00	0.06	0.06	0.43	0.00	0.00
	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propionibacterium	0.00	0.19	0.00	6.85	6.08	0.35	0.00	0.00
BD1-5	uncultured bacterium	Other	Other	Other	0.02	0.01	1.80	0.00	0.00	0.35	0.00	0.00
Bacteroidetes	BD2-2	uncultured bacterium	Other	Other	0.13	0.00	0.00	0.00	3.38	0.00	1.91	0.00
	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	3.16	6.75	0.03	14.36	52.04	0.29	71.87	23.14
	Bacteroidia	Bacteroidales	Porphyromonadaceae	Barnesiella	0.31	0.16	0.00	0.32	4.13	0.01	3.99	0.74
	Bacteroidia	Bacteroidales	Porphyromonadaceae	Parabacteroides	0.65	1.47	0.01	2.39	7.24	0.05	6.50	8.01
	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas	0.63	0.66	1.60	8.13	0.06	8.71	0.00	0.00
	Bacteroidia	Bacteroidales	Prevotellaceae	Paraprevotella	0.07	0.00	0.00	0.00	0.65	0.00	0.65	0.00
	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	0.31	0.38	0.85	0.00	0.02	9.53	0.00	0.00
	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	0.19	0.43	0.00	0.92	4.48	0.01	2.72	2.62
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Bergeyella	0.18	0.32	2.73	0.00	0.01	4.57	0.00	0.00
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Capnocytophaga	0.30	0.18	5.27	0.00	0.01	3.02	0.00	0.00
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Flavobacterium	0.07	0.08	14.42	0.00	0.00	0.13	0.00	0.00
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	uncultured	0.01	0.03	2.88	0.00	0.00	0.10	0.00	0.00
	Sphingobacteria	Sphingobacteriales	Sphingobacteriaceae	Sphingobacterium	0.06	0.12	12.74	0.00	0.00	0.18	0.00	0.00
Candidate division TM7	uncultured bacterium	Other	Other	Other	0.00	0.00	0.00	0.86	0.02	0.02	0.00	0.00
Cyanobacteria	4C0d-2	uncultured bacterium	Other	Other	0.01	0.07	0.00	0.12	0.36	0.00	0.38	0.48
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	0.00	0.03	0.00	4.35	0.63	0.11	0.00	0.00
	Bacilli	Lactobacillales	Aerococcaceae	Aerococcus	0.01	0.20	0.00	3.71	1.58	0.30	0.00	0.00
	Bacilli	Lactobacillales	Carnobacteriaceae	Alloiococcus	0.04	4.49	0.00	0.05	0.04	0.32	0.00	0.00
	Bacilli	Lactobacillales	Carnobacteriaceae	Dolosigranulum	0.08	1.08	0.00	0.00	0.00	0.01	0.00	0.00
	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	0.00	0.03	0.00	3.75	0.62	0.10	0.00	0.00
	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	0.58	0.32	0.79	0.00	0.00	2.57	0.00	0.00
	Clostridia	Clostridiales	Clostridiaceae	Alkaliphilus	0.01	1.21	0.00	0.73	0.02	0.05	0.00	5.30
	Clostridia	Clostridiales	Lachnospiraceae	Incertae Sedis	0.21	0.42	0.00	0.31	0.89	0.03	2.03	2.44
	Clostridia	Clostridiales	Lachnospiraceae	Pseudobutyrivibrio	0.00	0.15	0.00	0.28	0.01	0.01	0.00	0.68
	Clostridia	Clostridiales	Lachnospiraceae	uncultured	0.03	1.30	0.00	1.02	0.02	0.06	0.00	6.34
	Clostridia	Clostridiales	Ruminococcaceae	Incertae Sedis	0.04	1.29	0.00	2.24	0.12	0.08	0.07	5.78
	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus	0.08	7.24	0.00	12.10	1.65	0.34	0.00	36.05
	Clostridia	Clostridiales	Ruminococcaceae	uncultured	0.32	1.45	0.00	2.26	6.16	0.05	7.74	4.48
	Clostridia	Clostridiales	Veillonellaceae	Phascolarctobacterium	0.30	0.93	0.00	0.84	1.33	0.05	1.02	2.54
Fusobacteria	Fusobacteria	Fusobacteriales	ASCC02	uncultured bacterium	0.09	0.06	0.05	0.00	0.00	1.05	0.00	0.00
	Fusobacteria	Fusobacteriales	Fusobacteriaceae	Fusobacterium	0.37	0.18	1.50	0.00	0.01	2.52	0.00	0.00
	Fusobacteria	Fusobacteriales	Leptotrichiaceae	Leptotrichia	0.15	0.24	4.37	0.00	0.00	7.23	0.00	0.00
Planctomycetes	vadinHA49	uncultured bacterium	Other	Other	0.05	0.14	0.00	0.28	0.54	0.01	0.24	0.22
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Bradyrhizobium	10.11	7.42	0.00	0.02	0.02	0.09	0.00	0.00
	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	uncultured	55.14	35.66	0.02	0.18	0.13	0.45	0.00	0.00
	Alphaproteobacteria	Rhizobiales	Rhodobiaceae	Anderseniella	1.38	1.62	0.01	0.00	0.00	0.10	0.00	0.00
	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Derxia	0.02	0.10	0.00	0.00	0.00	3.47	0.00	0.00
	Betaproteobacteria	Burkholderiales	Alcaligenaceae	uncultured	0.00	0.11	1.52	0.00	0.00	0.22	0.00	0.00
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Comamonas	0.03	0.04	3.43	0.00	0.00	0.06	0.00	0.00
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Pelomonas	0.99	0.89	0.00	0.01	0.01	0.02	0.00	0.00
	Betaproteobacteria	Neisseriales	Neisseriaceae	Bergeriella	0.00	0.65	0.00	0.00	0.00	1.23	0.00	0.00
	Betaproteobacteria	Neisseriales	Neisseriaceae	Kingella	0.09	0.17	1.80	0.00	0.00	0.94	0.00	0.00
	Betaproteobacteria	Neisseriales	Neisseriaceae	Neisseria	0.22	0.42	1.50	0.00	0.01	5.61	0.00	0.00
	Betaproteobacteria	Neisseriales	Neisseriaceae	uncultured	4.44	0.00	0.00	0.00	0.29	0.01	0.00	0.00
	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio	0.07	0.34	0.00	0.56	0.33	0.01	0.39	0.73
	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Campylobacter	0.01	0.02	0.01	29.64	1.73	0.24	0.00	0.00
	Gammaproteobacteria	B38	uncultured bacterium	Other	0.04	0.00	0.00	0.00	3.86	0.00	0.00	0.00
	Gammaproteobacteria	Cardiobacteriales	Cardiobacteriaceae	uncultured	0.05	0.07	0.00	0.00	0.00	1.82	0.00	0.00
	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Actinobacillus	1.39	0.92	12.36	0.01	0.02	11.61	0.01	0.01
	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Aggregatibacter	0.56	0.41	2.74	0.00	0.00	4.46	0.00	0.00
	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus	0.07	0.14	1.98	0.00	0.00	2.58	0.00	0.00
	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	uncultured	0.03	0.00	1.46	0.00	0.00	0.05	0.00	0.00
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	1.26	1.46	7.55	0.00	0.02	13.60	0.00	0.00
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Moraxella	0.65	0.67	16.52	0.00	0.02	9.89	0.00	0.00
Spirochaetes	Spirochaetes	Spirochaetales	Spirochaetaceae	Treponema	0.00	0.07	0.00	2.54	0.77	0.18	0.00	0.00
Synergistetes	Synergistia	Synergistales	Synergistaceae	Cloacibacillus	0.06	0.14	0.00	0.17	0.50	0.01	0.47	0.44

Data derived from OTU table where OTUs with an abundance <0.1% of the total read count were removed in order to simplify the visualization of the results.

Suppl. Tab. 5 Relative abundances of the main bacterial phyla detected in the faecal microbiomes of several mammalian species.

Species	Actinobacteria	Bacteroidetes	Cyanobacteria	Firmicutes	Fusobacteria	Proteobacteria	Spirochaetes	Synergistetes	FB ratio	n° of individuals	method used	type of sample
Captive koala SN265	0.00	87.64	0.38	10.88	0.00	0.40	0.00	0.47	0.12	1	Illumina	faeces
Captive koala SN241	0.00	34.51	0.48	63.61	0.00	0.75	0.00	0.44	1.84	1	Illumina	faeces
Wild healthy koala (K1)	0.3	19.79	1.59	62.91	3.61	2.22	0	6.13	3.18	1	454	faeces
Wild diseased koala (K2)	0.57	5.49	0.04	86.87	0	6.03	0.00	0.45	15.82	1	454	faeces
Tammar wallaby	10.8	29.7	0	43.9	0	15.5	0	0	1.48	42	cloning	anal swab
Kangaroo & wallaby (3 sp.)	0.64	48.02	0	47.65	0.85	0.95	0.39	0	0.99	20	454	forestomach content
Panda	0.16	0.02	0.1	83.8	0	15.8	0	0	-	15	cloning	faeces
Mouse A	0.09	70.54	0	29.21	0	0.02	0	0	0.41	12	454	faeces
Mouse B	0	53.83	0	36.35	0	7.25	0	0	0.68	121	454	caecal mucosa
Sea lion	2	NA	0	80	NA	8	NA	NA	-	1	454	faeces
Wolf	4.6	16.9	0	60	9.2	9.2	0	0	3.55	3	cloning	faeces
Dog A	1	41.22	0.52	30.52	8.64	15.26	0.53	0.76	0.74	6	454	faeces
Dog B	1.81	2.25	0	95.36	0.3	0	0	0	42.38	12	454	faeces
Cat A	7.31	0.45	0	92.1	0.04	0	0	0	204.67	12	454	faeces
Cat B	1.16	76.22	0.51	12.98	0.68	5.85	0.41	0.58	0.17	5	454	faeces
Lynx	1.78	39.43	0	43.25	10.45	4.27	0.76	0	1.10	1	454	faeces
Cheetah	15.5	5.8	0	56.2	18.1	4.2	0	0	9.69	68	Illumina	faeces
Black-backed jackal	3.8	26.1	0.2	40.5	21.8	6.9	0	0	1.55	50	Illumina	faeces
Horse	4.5	14.2	0	68	0	10.1	1.9	0	4.79	6	454	faeces
Cow	6.8	7.6	0.08	63.7	0	18.3	0.3	0	8.38	4	454	faeces
Bison	3.8	0.57	0	55.1	0.0025	30.6	0.028	0.0009	96.67	40	Illumina	faeces
Pig A	0.5	30.25	0	47.75	0	5	2.75	0	1.58	8	454	faeces
Pig B	0	52	0	33	0	13	0	0	0.63	6	454	faeces
Howler monkey	0.44	19.24	0.04	71.43	0	1.97	0.05	0.17	3.71	32	454	faeces
Pygmy loris	10.98	41.19	0.28	9.44	0.26	30.43	0.5	0	0.23	2	454	faeces
Gorilla	5.3	1.1	0	71	0	0	1.1	0	64.55	1	cloning	faeces
Chimpanzee (3 sp.)	3.97	26.44	0.00	42.31	0.01	25.70	0.78	0.00	1.60	15	454	faeces
Bonobo	6.70	18.96	0.00	71.41	0.00	1.06	0.41	0.00	3.77	5	454	faeces
Human A	2.37	8.71	0.00	63.85	0.00	13.98	0.00	0.00	7.33	2	454	faeces
Human B	8.2	27.8	0	38.8	0	2.1	0	0	1.40	39	cloning	faeces
Human C	0.2	47.7	0	50.8	0.08	0.6	0	0	1.06	3	cloning	faeces
Primates (3 sp.)	0.665	12.4	0	72	0	1.4	1.31	0	5.81	9	454	faeces
Mammals (60 sp.)	4.7	16.3	0.1	65.7	0.67	8.8	0.46	0	4.03	106	cloning	faeces

For each taxon, the number of individuals examined, the sequencing method and the sample type used are indicated. In those studies where sequences were split between Bacteroidetes/Chlorobi and Bacteroidetes groups, those data were pooled into Bacteroidetes. "NA" is used for not available abundance data.
References: wild koalas[1]; tammar wallaby[2]; kangaroo and wallaby[3]; panda[4]; mouse A[5]; mouse B[6]; sea lion[7]; wolf[8]; dog A[9]; dog B[10]; cat A[10]; cat B[11] ; lynx[12]; cheetah[13]; black-backed jackal[13]; horse[14]; cow[15]; bison[16]; pig A[17]; pig B[18]; howler monkey[19]; pigmy loris[20]; gorilla[21]; chimpanzee[22]; bonobo[22]; human A[22]; human B[23]; human C[24]; primates[25]; mammals[26].

Suppl. Tab. 6 Frequency distribution of the most abundant genera and phyla in the rectal swabs and faeces of the two koalas.

a

	SN265 genera	FAECES: FALSE	FAECES: TRUE
RECTUM	**FALSE**	4	0
	TRUE	41	15

b

	SN241 genera	FAECES: FALSE	FAECES: TRUE
RECTUM	**FALSE**	2	0
	TRUE	38	20

Fisher's Exact Test for Count Data

p-value = 0.477

c

	SN265 phyla	FAECES: FALSE	FAECES: TRUE
RECTUM	**FALSE**	1	0
	TRUE	4	6

d

	SN241 phyla	FAECES: FALSE	FAECES: TRUE
RECTUM	**FALSE**	0	0
	TRUE	5	6

Fisher's Exact Test for Count Data

p-value = 1

Contingency tables showing the frequency distribution of the binary variables defined as the presence/absence of SN265´s most abundant bacterial genera (a) and phyla (c), and of SN241´s most abundant genera (b) and phyla (d) from the rectal and faecal samples. Below the tables is indicated the p-value of the Fisher´s exact test performed to test if there was any significant difference between the contingency tables of the two koalas both at genus and phylum level. Jaccard Index computed from table a = 0.27 (N=60; C.I. 95%: 0.21-0.46); from table b = 0.34 (N=60; C.I. 95%: 0.22-0.45); from table c = 0.6 (N=11; C.I. 95%: 0-0.7); from table d = 0.54 (N=11; C.I. 95%: 0-0.64).

Suppl. Tab. 7 Correlation between the faecal samples of the two captive and two wild koalas.

correlation ρ	SN265	SN241	K1	K2
SN265	-	**0.94 (<0.0001)**	**0.80 (0.0052)**	**0.79 (0.0039)**
SN241	**0.94 (<0.0001)**	-	**0.64 (0.034)**	**0.83 (0.0017)**
K1	**0.80 (0.0052)**	**0.64 (0.034)**	-	0.56 (0.073)
K2	**0.79 (0.0039)**	**0.83 (0.0017)**	0.56 (0.073)	-

Correlation matrix of the relative abundances of the eleven phyla detected both in the faecal samples of the two captive koalas from this study (SN265 and SN241) and the two wild koalas from Barker et al. 2013 (K1 and K2). Correlation was measured using the Spearman's rank correlation coefficients. The p-values are given in brackets. The significant correlations ($p \leq 0.05$) are presented in bold characters.

Suppl. Tab. 8 Phyla distribution in the faecal samples of the two captive and two wild koalas.

	SN265	SN241	K1	K2
Actinobacteria	✓	✓	✓	✓
Bacteroidetes	✓	✓	✓	✓
Chloroflexi	✓	✓	✓	✓
Cyanobacteria	✓	✓	✓	✓
Deferribacteres	✓	✓	X	X
Firmicutes	✓	✓	✓	✓
Fusobacteria	✓	✓	✓	X
Planctomycetes	✓	✓	✓	✓
Proteobacteria	✓	✓	✓	✓
Synergistetes	✓	✓	✓	✓
Verrucomicrobia	X	X	✓	X

Contingency tables showing the presence/absence distribution of the eleven phyla detected both in the faeces of the two captive koalas analysed in this study (SN265 and SN241) and of the two wild koalas from Barker et al. 2013 (K1 and K2).

Suppl. Tab. 9 Similarity between the faecal samples of the two captive and the two wild koalas.

Pairs of faecal samples	N	Jaccard´s index	C.I. 95% +	C.I. 95% -
SN265 - K1	11	0.82*	0	0.64
SN265 - K2	10	0.80*	0	0.7
SN265 - SN241	10	1*	0	0.7
SN241 - K1	11	0.82*	0	0.64
SN241 - K2	10	0.80*	0	0.7
K1 - K2	10	0.80*	0	0.7

Jaccard´s coefficient of similarity between bacterial presence/absence profiles between faecal samples of captive (SN265 and SN241 from this study) and wild koalas (K1 and K2, from Barker et al. 2013). The table shows the lower and upper critical values of the coefficient with a probability level of $P < 0.05$ considering the total number of taxa present in either of the two samples being compared (N). * Significant values.

Suppl. Tab. 10 Statistics of the raw and quality filtered sequences from Illumina sequencing of the eye, mouth, rectal and faecal samples of the two koalas.

	EYE		MOUTH		RECTUM		FAECES		STATISTICS				
	SN265	SN241	SN265	SN241	SN265	SN241	SN265	SN241	SUM	MIN	MAX	MEAN	SD
Total raw reads	274,523	336,440	470,459	367,353	263,487	415,046	196,872	260,057	2,584,237	196,872	470,459	323,029.63	91,124.99
Merged reads	243,988	286,804	421,748	324,217	225,522	357,386	182,148	220,422	2,262,235	182,148	421,748	282,779.38	80,542.80
Quality trimmed data	225,900	262,859	384,573	297,667	203,213	326,262	163,891	200,507	2,064,872	163,891	384,573	258,109	74,134.42
after SINGLETONS removal	225,273	262,284	383,981	296,969	202,567	325,160	163,547	199,241	2,059,022	163,547	383,981	257,377.75	74,097.79
after CHIMERAS removal	223,343	260,529	383,330	274,108	199,898	324,254	161,378	194,809	2,021,649	161,378	383,330	252,706.13	73,722.19
after CHLOROPLAST removal	197,484	225,501	383,043	274,022	199,804	320,563	161,378	194,797	1,956,592	161,378	383,043	244,574	75,403.77

Supplementary References

1. Barker, C. J., Gillett, A., Polkinghorne, A. & Timms, P. Investigation of the koala (*Phascolarctos cinereus*) hindgut microbiome via 16S pyrosequencing. *Vet. Microbiol.* **167**, 554-64 (2013).
2. Chhour, K. L., Hinds, L. A., Deane, E. M. & Jacques, N.A. The microbiome of the cloacal openings of the urogenital and anal tracts of the tammar wallaby, *Macropus eugenii*. *Microbiol.* **154**, 1535-43 (2008).
3. Gulino, L. M. *et al.* Shedding light on the microbial community of the macropod foregut using 454-amplicon pyrosequencing. *PLoS One* **8**, e61463 (2013).
4. Zhu, L., Wu, Q., Dai, J., Zhang, S. & Wei, F. Evidence of cellulose metabolism by the giant panda gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 17714-9 (2011).
5. Schloss, P. D. *et al.* Stabilization of the murine gut microbiome following weaning. *Gut Microbes* **3**, 383-93 (2012).
6. Linnenbrink, M. *et al.* The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Mol. Ecol.* **22**, 1904-16 (2013).
7. Lavery, T. J., Roudnew, B., Seymour, J., Mitchell, J. G. & Jeffries, T. High nutrient transport and cycling potential revealed in the microbial metagenome of Australian sea lion (*Neophoca cinerea*) faeces. *PLoS One* **7**, e36478 (2012).
8. Zhang, H. & Chen, L. Phylogenetic analysis of 16S rRNA gene sequences reveals distal gut bacterial diversity in wild wolves (*Canis lupus*). *Mol. Biol. Rep.* **37**, 4013-22 (2010).
9. Swanson, K. S. *et al.* Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. *Isme J.* **5**, 639-49 (2011).
10. Handl, S., Dowd, S. E., Garcia-Mazcorro, J. F., Steiner, J. M. & Suchodolski, J. S. Massive parallel 16S rRNA gene pyrosequencing reveals highly diverse fecal bacterial and fungal communities in healthy dogs and cats. *FEMS Microbiol. Ecol.* **76**, 301-10 (2011).
11. Tun, H. M. *et al.* Gene-centric metagenomics analysis of feline intestinal microbiome using 454 junior pyrosequencing. *J. Microbiol. Methods* **88**, 369-76 (2012).

12. Alcaide, M. *et al.* Gene sets for utilization of primary and secondary nutrition supplies in the distal gut of endangered Iberian lynx. *PLoS One* **7**, e51521 (2012).

13. Menke, S. *et al.* Oligotyping reveals differences between gut microbiomes of free-ranging sympatric Namibian carnivores (*Acinonyx jubatus*, *Canis mesomelas*) on a bacterial species-like level. *Front. Microbiol.* **5**, 526 (2014).

14. Costa, M. C. *et al.* Comparison of the fecal microbiota of healthy horses and horses with colitis by high throughput sequencing of the V3-V5 region of the 16S rRNA gene. *PLoS One* **7**, e41484 (2012).

15. Mao, S., Zhang, R., Wang, D. & Zhu, W. The diversity of the fecal bacterial community and its relationship with the concentration of volatile fatty acids in the feces during subacute rumen acidosis in dairy cows. *BMC Vet. Res.* **8**, 237 (2012).

16. Weese, J. S., Shury, T. & Jelinski, M. D. The fecal microbiota of semi-free-ranging wood bison (*Bison bison athabascae*). *BMC Vet. Res.* **10**, 120 (2014).

17. Lamendella, R., Oerther, D. B., Martinson, J., and Santo Domingo, J. W. Comparative fecal metagenomics reveals previously unknown functionality of the distal swine gut. *BMC Microbiol.* **11**, 103 (2011).

18. Looft, T. *et al.* In-feed antibiotic effects on the swine intestinal microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1691-6 (2012).

19. Amato, K. R. *et al.* Habitat degradation impacts black howler monkey (*Alouatta pigra*) gastrointestinal microbiomes. *Isme J.* **7**, 1344-53 (2013).

20. Xu, B. *et al.* Metagenomic analysis of the pygmy loris fecal microbiome reveals unique functional capacity related to metabolism of aromatic compounds. *PLoS One* **8**, e56565 (2013).

21. Frey, J. C. *et al.* Fecal bacterial diversity in a wild gorilla. *Appl. Environ. Microbiol.* **72**, 3788-92 (2006).

22. Ochman, H. *et al.* Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol* **8**, e1000546 (2010).

23. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174-80 (2011).

24. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635-8 (2005).

25. Yildirim, S. *et al.* Characterization of the fecal microbiome from non-human wild primates reveals species specific microbial communities. *PLoS One* **5**, e13963 (2010).

26. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320**, 1647-51 (2008).

Chapter III

Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses

Published in *Journal of Virology*

http://dx.doi.org/10.1128/JVI.02745-15

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses

Niccolò Alfano,[a] Sergios-Orestis Kolokotronis,[b,c] Kyriakos Tsangaras,[a]* Alfred L. Roca,[d] Wenqin Xu,[e] Maribeth V. Eiden,[e] Alex D. Greenwood[a,f]

Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany[a]; Department of Biological Sciences, Fordham University, Bronx, New York, USA[b]; Sackler Institute for Comparative Genomics and Division of Invertebrate Zoology, American Museum of Natural History, New York, New York, USA[c]; Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA[d]; Section on Directed Gene Transfer, Laboratory of Cellular and Molecular Regulation, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland, USA[e]; Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany[f]

ABSTRACT

Gibbon ape leukemia viruses (GALVs) are part of a larger group of pathogenic gammaretroviruses present across phylogenetically diverse host species of Australasian mammals. Despite the biomedical utility of GALVs as viral vectors and in cancer gene therapy, full genome sequences have not been determined for all of the five identified GALV strains, nor has a comprehensive evolutionary analysis been performed. We therefore generated complete genomic sequences for each GALV strain using hybridization capture and high-throughput sequencing. The four strains of GALV isolated from gibbons formed a monophyletic clade that was closely related to the woolly monkey virus (WMV), which is a GALV strain that likely originated in a gibbon host. The GALV-WMV clade in turn formed a sister group to the koala retroviruses (KoRVs). Genomic signatures of episodic diversifying selection were detected among the gammaretroviruses with concentration in the *env* gene across the GALV strains that were particularly oncogenic and KoRV strains that were potentially exogenous, likely reflecting their adaptation to the host immune system. *In vitro* studies involving vectors chimeric between GALV and KoRV-B established that variable regions A and B of the surface unit of the envelope determine which receptor is used by a viral strain to enter host cells.

IMPORTANCE

The gibbon ape leukemia viruses (GALVs) are among the most medically relevant retroviruses due to their use as viral vectors for gene transfer and in cancer gene therapy. Despite their importance, full genome sequences have not been determined for the majority of primate isolates, nor has comprehensive evolutionary analysis been performed, despite evidence that the viruses are facing complex selective pressures associated with cross-species transmission. Using hybridization capture and high-throughput sequencing, we report here the full genome sequences of all the GALV strains and demonstrate that diversifying selection is acting on them, particularly in the envelope gene in functionally important domains, suggesting that host immune pressure is shaping GALV evolution.

Gibbon ape leukemia virus (GALV) is an exogenous gammaretrovirus associated with hematopoietic neoplasms in captive colonies of white-handed gibbon (*Hylobates lar*). Five strains of GALV have been isolated from gibbons. The first was isolated from an animal with lymphocytic leukemia in a colony at the San Francisco Medical Center (strain SF) (1, 2). GALV was later isolated from gibbons displaying malignant tumors, notably an individual gibbon with granulocytic leukemia, at the Southeast Asia Treaty Organization Medical Research Laboratory in Bangkok, Thailand (strain SEATO) (3, 4), and another gibbon with lymphocytic leukemia from a colony on Hall's Island, near Bermuda (strain GALV-H) (5, 6). The Brain strain was isolated from two healthy gibbons injected with brain extracts from human patients with kuru and from an uninoculated cage mate (7). The SEATO strain has been shown to cause chronic myelogenous leukemia when injected into juvenile gibbons (8).

A closely related retrovirus isolated from a 3-year-old male woolly monkey (*Lagothrix lagotricha*) with multiple fibrosarcomas was originally designated SSAV (for simian sarcoma-associated virus) and now renamed woolly monkey virus (WMV). WMV is considered a member of the GALV lineage (9). WMV isolated from the woolly monkey exists as a mixture of a replication-defective acute transforming virus and its associated replication-competent helper virus (10). Replication-competent WMV is related to GALV as supported by immunological (11) and serological tests (9), antigenic similarities in some gene products (7, 12, 13), and high RNA sequence homology (5, 7). Since the woolly monkey from which WMV was isolated was reported to have been in contact with a gibbon for the 3 months before its death, WMV is likely the product of a single horizontal transmission of GALV from a gibbon to a woolly monkey.

The GALV genomes deposited in GenBank are not representative of any one of the five GALV strains. Rather, the GALV-SEATO genome deposited by Delassus et al. (14) (M26927) rep-

Received 27 October 2015 **Accepted** 24 November 2015

Accepted manuscript posted online 4 December 2015

Citation Alfano N, Kolokotronis S-O, Tsangaras K, Roca AL, Xu W, Eiden MV, Greenwood AD. 2016. Episodic diversifying selection shaped the genomes of gibbon ape leukemia virus and related gammaretroviruses. J Virol 90:1757–1772. doi:10.1128/JVI.02745-15.

Editor: K. L. Beemon

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

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

resents a GALV-SEATO/SF chimeric genome that contains an *envelope* open reading frame (ORF) encoding a truncated form of the envelope protein lacking an R peptide (14). The R peptide in the cytoplasmic terminus of the gammaretroviral envelope protein prevents membrane fusion before budding. Transfection of this truncated form of GALV-SEATO *envelope* into human cells resulted in the expression of a hyperfusogenic GALV envelope protein with strong cytotoxic effects (15, 16). The second GALV genome sequence available in GenBank (U60065) is from a GALV discovered as a contaminant of an HIV-infected human cell line originally referred to as retrovirus X (17) and subsequently designated the GALV-X strain (18). The provenance of GALV-X remains unknown.

Only *envelope* sequences of the remaining GALV strains—GALV-Brain, Hall's Island, and SF—have been determined (19). Phylogenetic analysis of the two full-genome GenBank sequences and related retroviruses has revealed that GALV is most closely related to the koala retrovirus (KoRV) among viruses sequenced to date (20). KoRV and GALV occur in taxonomically distant mammalian hosts from different continents, suggesting that these viruses may be the products of a recent cross-species transmission, most likely originating in a common intermediate vector to both species (20, 21). In a recent study attempting to identify such an intermediate host, the *Melomys burtoni* retrovirus (MbRV) was isolated from the grassland mosaic-tailed rat, an Australian murid rodent, and showed a high nucleotide identity (93%) and close phylogenetic relatedness to GALV-SEATO (M26927) (21). Nevertheless, because of the different geographic distribution of *M. burtoni* and gibbons, MbRV cannot be considered the source of GALV, and therefore the origins of GALV are still unclear.

To better characterize GALV phylogenetic relationships and functional domains in viral control regions and structural genes besides *env*, we applied two methods to determine the complete genomic sequence of all known GALV strains. A PCR-based approach on DNA extracted from GALV-infected cell lines using primers designed on the limited GALV sequences available in GenBank was applied, but it did not recover the full genome sequences of all the strains because of the unsuccessful amplification of certain portions of the genomes. Therefore, hybridization capture and high-throughput sequencing were performed to determine the full-length GALV genomes (22, 23). We report the complete nucleotide sequence of all GALV strains, their genomic structure, the phylogenetic relationships within the GALVs, their relationship to other gammaretroviruses, and the selection pressures driving evolution within this retroviral clade.

MATERIALS AND METHODS

Cell lines and viruses. GALV wild-type viruses were obtained from the following productively infected cell lines: SEATO-88, GALV-SEATO-infected bat lung fibroblasts; GALV-4-88, GALV-Brain-infected bat lung fibroblasts; 71-AP-1, WMV-infected marmoset fibroblasts; MLA-144, GALV-SF-infected primate T cells; 6G1-PB, GALV-Hall's Island-infected lymphocytes; and HOS (ATCC CRL-1543) GALV-SF-infected human osteosarcoma cells. GALV-SF was represented by two different cultures, one from the MLA-144 cell line and another cultured in HOS cells.

DNA extraction. Genomic DNA extraction from the cell lines was performed using the Wizard Genomic DNA purification kit (Promega) according to the manufacturer's protocol. The DNA concentration was determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies).

TABLE 1 Primers that yielded PCR products and high-quality Sanger sequences

Primer	Sequence (5′-3′)	GALV strain(s)[a]
SSAVF	CAAGAACTCCCACATGACCG	WMV
SSAVR	GAACACGTCTGCTCGCTAC	WMV
U5	CCCGTGTGTCCAATAAAACCTCT	SF, SEATO, WMV
PolF1	TGGTATACAGACGGTAGCAGT	SEATO, Br, WMV
EnvR1	CACAAYYCCATTCTTTACAGTAT	SF, SEATO, H, Br, WMV
EnvR2	GGAGGTCAGCATCTATGGCGATC	SF, SEATO, H, Br, WMV
U3	AGCGAGAGGCAAGGTAAT	H, WMV
PolR2	GCAAACCCAGGGATCCAGAGTCT ACA	SF, H, Br
PolR1	CTAGCCCATACCGTCCGC	SF, SEATO
GagF1	CCCCTATCTCCCTCACTCT	SEATO, H, Br
GagF2	GACCTCGCTCAGAGTCCCCCACC ATG	SEATO
F2	GCCTTCCCCCTCAATCGACCTC	SEATO
F3	ACTAGACAAAGACCAGTGCGCAT AC	SEATO
F4	TGGCTCCAGCTTTTCCCCACTG	SEATO
EnvF	ACCTCCKGAYTCAGACTATAC	SEATO
HallsR	CACGTCTGTTCGCTACTCAC	H
HallsF	CTTCTCGCTTCTGTACCCG	H

[a] Abbreviations: SF, San Francisco; H, Hall's Island; Br, Brain.

PCR. Two primer pairs were designed, based on the alignment of the GALV sequences available in GenBank (SEATO, M26927; GALV-X, U60065), to target two regions, each ca. 4 kb in length, which together cover the GALV genome. Primers U5 (5′-CCCGTGTGTCCAA TAAAACCTCT-3′) and PolR1 (5′-CTAGCCCATACCGTCCGC-3′) were used to amplify the first 4 kb of the GALV genome (the 5′ long terminal repeat [5′ LTR], *gag*, and part of the *pol* gene) and primers PolF1 (5′-TGGTATACAGACGGTAGCAGT-3′) and U3 (5′-AGCGAGAGGC AAGGTAAT-3′) for the second 4 kb (part of the *pol* gene, *gag*, and the 3′ LTR). The PCRs were performed in a final volume of 23 μl using 100 ng of DNA extract, a 0.6 μM final concentration of each primer, 12.5 μl of 2× MyFi Mix (Bioline), and sterile-distilled water. The thermal cycling conditions were as follows: 95°C for 4 min; 40 cycles at 95°C for 30 s, 53 to 57°C (based on the best PCR product yield per strain determined empirically) for 30 s, and 72°C for 6 min; and finally 72°C for 10 min. An aliquot of each PCR product was visualized on 1.5% (wt/vol) agarose gels stained with GelRed (Biotium). In cases of positive amplification, the PCR products were purified using the MSB Spin PCRapace kit (Stratec Molecular GmbH), quantified using a NanoDrop ND-1000 spectrophotometer, and Sanger sequenced by primer walking. The primers that yielded high-quality Sanger sequences are listed in Table 1.

Illumina library preparation. The extracted DNA from each cell line was sheared using a Covaris M220 (Covaris) to an average size of 250 bp. Aliquots from each fragmented DNA extract were used to generate Illumina libraries as described by Meyer and Kircher (24) with the modifications described in Alfano et al. (25). Each library contained a unique index adapter to allow for subsequent discrimination among samples after the sequencing of pooled libraries. A negative-control extraction library was also prepared and indexed separately to monitor for experimental cross-contamination. Each library was amplified in three replicate reactions to minimize amplification bias in individual PCRs. The amplifications of the libraries were performed using Herculase II Fusion DNA polymerase (Agilent Technologies) in 50-μl volume reactions, with the cycling conditions of 95°C for 5 min, followed by five cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 40 s and then finally 72°C for 7 min. After pooling the three replicate PCR products for each sample, amplified libraries were purified using the QIAquick PCR purification kit (Qiagen) and quantified using the 2200 TapeStation (Agilent Technologies) on D1K ScreenTapes. Three additional amplification cycles were performed for SEATO and

SF-HOS libraries using Herculase II Fusion DNA polymerase with P5 and P7 Illumina library outer primers with the same cycling conditions to balance library concentrations.

Hybridization capture baits. PCR products used as baits for capturing GALV sequences from the Illumina libraries were generated from the SEATO and SF-MLA strains. A preliminary phylogenetic analysis of the *envelope* nucleotide sequences of SEATO, Hall's Island, Brain, SF, and WMV strains deposited in GenBank by Ting et al. (19) (AF055060 to AF055064) suggested that baits from these two strains would cover sufficient genetic diversity to allow for capture of unknown and divergent GALV sequences, since SEATO and SF represent each of the two main branches in which the GALV strains are clustered and thus cover much GALV diversity (data not shown). The phylogenetic analysis was carried out in Seaview v4 (26) using the neighbor-joining method (27) and the HKY model (28). Node robustness was estimated with 100 bootstrap replicates. KoRV (AF151794) was used as outgroup. Primer pairs U5-PolR1 and PolF1-U3 were used to amplify the genome of SEATO and SF-MLA, with the same reaction setup and thermal profile described in the PCR methods. PCR products were purified using the MSB Spin PCRapace kit, quantified using a NanoDrop ND-1000 and Sanger sequenced to verify that the target region had been amplified. After sequence verification, the PCR products were then pooled to equimolar concentrations to produce a mixed SEATO/SF-MLA bait and fragmented using a Covaris M220 to generate 250-bp fragments. The GALV fragments were then blunt ended using the Quick Blunting kit (New England BioLabs), ligated to a biotin adaptor using the Quick Ligation kit (New England BioLabs), and immobilized in separated individual tubes on streptavidin-coated magnetic beads as described previously (22).

Hybridization capture. Each amplified Illumina library was mixed with blocking oligonucleotides (200 μM) that help prevent cross-linking of Illumina library adapters, Agilent 2× hybridization buffer, and Agilent 10× blocking agent and heated at 95°C for 3 min to separate the DNA strands (22). Each Illumina library hybridization mixture was then combined in separate tubes with the biotinylated baits bound to the streptavidin beads. Samples were incubated in a mini-rotating incubator (Labnet) for 48 h at 65°C, during which the hybridization took place. After 48 h, the beads were washed to remove off-target DNA as described previously (22), and the hybridized libraries were eluted by incubation at 95°C for 3 min. The DNA concentration for each eluted sample was measured using the 2200 TapeStation on D1K ScreenTapes and further amplified accordingly using P5 and P7 Illumina outer primers (24). The enriched amplified libraries were then pooled in equimolar amounts to a final library concentration of 8 nM for paired-end sequencing (2 × 250) on an Illumina MiSeq platform with the v2 reagents kit at the Danish National High-Throughput DNA Sequencing Centre in Copenhagen, Denmark. As a control, a 1% PhiX genome library spike-in was used.

Genome sequence assembly and annotation. A total of 12,949,200 paired-end sequence reads 250-bp long were generated (average = 2,158,200 paired-end reads per sample, standard deviation [SD] = 451,197.4) and then sorted by index sequences. Adaptor sequences were trimmed from the reads using Cutadapt v1.2.1 (29), and low-quality reads were removed using Trimmomatic v0.27 (30), with a quality cutoff set at 20. Reads that were shorter than 20 bp were excluded from further analyses. After adaptor and quality trimming, 97.6% of the sequences were retained. Reads were then mapped to the GALV-X full genome reference sequence (U60065) using BWA v0.7.10 with default parameters (BWA-MEM algorithm) (31). Reads from the SEATO strain were also mapped to the SEATO full genome reference sequence (M26927), and the results of the two alignments were compared. Samtools v1.2 (32) was used to convert, sort, and index the aligned data files, while potential duplicates were removed using Picard (http://broadinstitute.github.io/picard). Variant call analysis was performed using GATK v1.6-11 (33), setting the minimum variant frequency to 0.2, the depth of coverage to 20, quality to 30, and the quality by depth to 5. To get better variant calling results, paired-end reads were first merged into single reads using FLASH with default parameters (34). The alignments were then visualized and manually curated using Geneious v7.1.7 (Biomatters, Inc.). Consensus sequences were generated as the majority character state at every position in an alignment of sequences. Regions that mapped poorly, likely corresponding to regions diverging from the reference sequence, were resolved by comparison with previously generated Sanger sequences. Nucleotide positions that could not be resolved by variant calling or Sanger sequencing due to the presence of multiple nucleotides at a given position were identified as polymorphisms and assigned IUPAC ambiguity codes. Exact counts for homopolymer stretches must be considered tentative due to the limitations of the Illumina platform in distinguishing their lengths. Homopolymer lengths were defined by assigning the number of nucleotides detected in the most abundant reads. In order to identify protein domains and regulatory motifs, the nucleotide sequence of each strain was compared to the annotated genome sequences available in GenBank for GALV-X (U60065), SEATO (M26927), and KoRV (AF151794) and also analyzed using the NCBI Conserved Domains Database (CDD; http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). The consensus sequence and annotations of each GALV strain genome were deposited in GenBank. Illumina reads mapping to GALV-X for each captured GALV strain were deposited in the NCBI Sequence Read Archive.

Cell lines used in the GALV/KoRV-B chimeric envelope experiment. 293T human embryonic kidney cells (ATCC CCL 11268) and murine *Mus dunni* tail fibroblast MDTF cells (35) were maintained in Dulbecco modified Eagle medium with high glucose, supplied with 10% fetal bovine serum, 100 U of penicillin/ml, and 100 μg of streptomycin/ml. MDTF cells expressing human PiT1 and human THTR1 individually were described previously (36).

Construction of GALV/KoRV-B chimeric envelope. Both chimeric envelope proteins were generated using overlap extension PCR cloning as described previously (37), and DNA sequencing analysis confirmed the sequence of each chimeric envelope. The PCR fragment of KoRV-B VRA was used to replace the corresponding VRA of GALV SEATO envelope protein (residues 46 to 100 of GALV) to construct GALV-VRA$_{KoRV-B}$. The VRA region of KoRV-B, corresponding to envelope residues 49 to 107, was PCR amplified using the following primer pairs flanking the VRA regions of the KoRV-B envelope gene: sense (5′-GTCCTGGGAACTGG AAAAGACTGATCATCCTCTTAAG-3′) and antisense (5′-CTTCTGAA AGGGTCCGGCCATCCCGGGG-3′). GALV-VRA$_{KoRV-B}$ was used as a template to replace the VRB of GALV SEATO (residues 46 to 100) with that of KoRV-B (residues 193 to 204) for the generation of GALV-VRA/VRB$_{KoRV-B}$. To generate GALV-VRA/VRB$_{KoRV-B}$, a modified overlap extension PCR cloning was used, where a primer pair containing KoRV-B VRB sequences was used instead of a PCR fragment. The sense primer of the primer pair contain GALV sequences upstream of the VRA region (underlined) sense primer, 5′-GTGTTCGCATGTCCCCGTAG GGTGGCCCAGGCCTACAGTTATGAGGTCTTTTGAGGATTTAGA TAGCCA-3′, and the antisense primer contains GALV sequences downstream of the VRA region (underlined): 5′-GTAGGCCTGGGC CACCCTACGGGGACATGCGAACACACCGGCTGGTGTAACCCCC TTAAAATAGATTTC-3′.

V5 epitope tagging of GALV and KoRV-B envelope proteins. Using the modified overlap extension PCR as mentioned above, the DNA sequence encoding the V5 epitope tag (GKPIPNPLLGLDST) was engineered into the primer pair to be used as the oversized primer for overlap extension PCR cloning to construct tagged KoRV-B and GALV SEATO envelope protein with a V5 epitope inserted downstream of signal peptide at the N-terminal of envelope sequences.

Retroviral vector production and transduction. A ProFection mammalian transfection system-calcium phosphate kit (Promega) was used for transfection of 293T cells 10-cm plates. For binding assay, 20 μg of expression plasmid encoding individual V5-epitope tagged-envelope protein was transfected into 293T cells. For assessment of envelope function of the different chimeras, pCI-neo plasmid encoding individual envelope protein was cotransfected with an MLV *gag-pol*, and a retroviral genome

encoding β-galactosidase (*lacZ* gene) as an indicator of transduction. At 48 to 72 h posttransfection, viral supernatants was collected, filtered through a 0.45-μm-pore-size syringe and stored at −80°C. For transduction, target cells were seeded at a density of 4×10^4 per well of a 24-well plate and exposed 24 h later to retroviral particles bearing one of the GALV, GALV-VRA$_{KoRV-B}$, GALV-VRA/VRB$_{KoRV-B}$, or KoRV-B envelopes in the presence of 10 μg of Polybrene/ml. At 48 h postexposure, X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) staining was performed, and β-galactosidase expression was evaluated by counting blue colonies to calculate the titers of the viral vectors. The titers of the viral vectors were averaged from at least three independent experiments and are expressed as mean numbers of β-galactosidase-expressing cells ± the SD of the mean.

Envelope binding analysis. V5 epitope-tagged envelope proteins were transfected into 293T cells and, after 48 to 72 h, the supernatant was filtered and used for binding assays. MDTFPiT1 or MDTFTHTR1 cells were trypsinized from a tissue culture flask, and 10^6 cells were resuspended with supernatant containing each of the V5-tagged envelope proteins, followed by incubation at 37°C for 45 min with shaking. To detect the presence of V5-tagged envelope on the surface of the target mouse cell, anti-V5 monoclonal antibody (Bio-Rad) was used as the first antibody, followed by a secondary antibody, a goat anti-mouse antibody conjugated to phycoerythrin (Invitrogen). The cells were then subjected to flow cytometric analysis using a FACSCalibur (BD Biosciences), and data were analyzed using CellQuest software (BD Biosciences).

Evolutionary analyses. To characterize the phylogenetic relationships among the GALV strains and other gammaretroviruses, we inferred phylogenetic trees using the translated amino acid sequences. The sequences of Env, Gag, and Pol proteins of each gammaretrovirus were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/GenBank) (Table 2). Individual gene sequences for *env*, *gag*, and *pol* were aligned by preserving the protein-coding frame in TranslatorX (38) using MAFFT (39). Sequences presenting premature stop codons were excluded from the analyses. For this reason, OOEV and MbRV were removed from the alignment of the *pol* gene. Phylogenetic analysis was carried out using maximum likelihood as an optimality criterion and the general time-reversible substitution model (40) for nucleotide sequences and the rtREV model (41) for amino acid sequences with among-site rate heterogeneity modeled by the Γ distribution and four rate categories (42), as implemented in the POSIX-threads build of RAxML v8 (43). Node robustness was assessed with rapid bootstrap pseudoreplicates (44). The bootstopping criterion (45) as implemented in RAxML showed that more than 100 (for amino acid sequences) and 500 (for nucleotide sequences) rapid bootstrap pseudoreplicates were unlikely to alter node support. Gene alignments were checked for recombination using the Φ_w test statistic (otherwise referred to as the pairwise homoplasy index) (46). The signature of natural selection was examined using the mixed effects model evolution (MEME) that allows the ratio ω of the rate of nonsynonymous substitution (*dN*) to the rate of synonymous substitution (*dS*) to vary along the tree branches and across codons (47), Fast Unconstrained Bayesian AppRoximation (FUBAR) that estimates codon-wise trends of negative or positive selection (48), and the branch-site random effects likelihood (BSREL) method that is able to detect the branches on which a proportion of codons evolve with $\omega > 1$ (49). The protein-coding sequences of *env*, *gag*, and *pol* were concatenated and analyzed in a partitioned framework, where each partition was allowed to evolve under its own substitution model.

Accession numbers. The consensus sequence and annotations of each GALV strain genome were deposited in GenBank under accession numbers KT724047 to KT724051. Illumina reads mapping to GALV-X for each captured GALVstrain were deposited in the NCBI Sequence Read Archive as BioProject PRJNA306599.

RESULTS

PCR and Sanger sequencing of GALV strains. DNA was extracted from six cell lines, each infected with a different strain of GALV. Two primer sets (U5-PolR1 and PolF1-U3) based on the full genome sequences of GALV-X (U60065) and SEATO (M26927) were designed to generate two overlapping PCR products, each 4 kb long, in order to cover the whole GALV genome from each cell line. However, full sequences of the GALV strains were not recovered by PCR, since one of the two primer pairs generally failed to yield an amplification product or readable Sanger sequence, presumably due to the coamplification of different products. Furthermore, the PCR approach has the disadvantage of omitting sequences at the genome ends covered by the primers. The primers that yielded products and high-quality Sanger sequences are listed in Table 1. The Sanger sequences, however, were subsequently used to confirm the proper assembly of high-throughput sequences obtained by hybridization capture.

Hybridization capture and high-throughput sequencing of GALV strains. Illumina libraries were prepared from each cell line DNA extract and indexed to allow all samples to be processed in a single Illumina sequencing experiment. Two amplicons 4 kb in length, together covering the entire GALV genome from SF-MLA and SEATO strains, were generated as hybridization capture baits (23). Equimolar amounts of indexed libraries were hybridized to the GALV baits and the enriched GALV libraries sequenced on an Illumina MiSeq platform. The enrichment (proportion of on-target reads mapping to GALV), which ranged from 0.6% (Brain) to 15% (Hall's Island), was comparable to previous reports (22), although the rates for the Brain, SEATO, and SF-HOS strains were relatively low (0.6 to 0.9%). This might be in part due to low sequence identity between baits used and some of the strains targeted. Nonetheless, full coverage of the GALV genome was obtained from each of the cell lines included in the study. The capture enrichment yielded very high per-base coverage, with average values ranging from 2,362× for SF-MLA to 116× for Brain (Fig. 1A and B). Although the per-base coverage differed among strains, the coverage profiles were similar among the GALV strains (Fig. 1A and B). The negative control generated few sequence reads, which only sporadically mapped to GALV (33 of 560 total reads) (Fig. 1A and B). This low frequency of target-mapping reads was well within the known misindex error reading rate on the Illumina platform (0.3%) (50) and is consistent with the rate reported by previous studies (23).

GALV consensus sequence determination. A nucleotide consensus sequence was generated for each GALV strain, with the exception of SF-MLA, in which the presence of multiple distinct viral sequences prevented assembly. Therefore, the genome of GALV-SF was derived from an infected HOS cell line (SF-HOS), which lacks the defective GALV-SF variants (M. V. Eiden, unpublished data).

The consensus sequences were confirmed by the previously generated Sanger sequences covering parts of the GALVs genomes (Fig. 1C). There was concordance between the hybridization capture and PCR-derived sequences. Polymorphisms detected among sequences in the hybridization capture data were confirmed as double peak signals in the Sanger electropherograms. By comparison of the GALV consensus sequences with the primer sequences, we found that the failures in the PCRs or Sanger sequencing were due to indels and polymorphisms that presumably prevented the primers from binding to the templates.

GALV strain genome structures and regulatory motifs. All GALV strains had comparable genome sizes ranging from 8,370 bp (Brain) to 8,534 bp (SEATO) (Table 3 and Fig. 2A). In an

TABLE 2 Gammaretrovirus sequences used for phylogenetic analyses in this study

Strain (accession no.)	Full name	Host	*gag*	*pol*	*env*	Reference or GenBank accession no.
GALV SF	Gibbon ape leukemia virus strain San Francisco	Gibbon	✓	✓	✓	This study
GALV Brain	Gibbon ape leukemia virus strain Brain	Gibbon	✓	✓	✓	This study
GALV Hall's Island	Gibbon ape leukemia virus strain Hall's Island	Gibbon	✓	✓	✓	This study
GALV SEATO	Gibbon ape leukemia virus strain SEATO	Gibbon	✓	✓	✓	This study
WMV	Woolly monkey virus	Gibbon	✓	✓	✓	This study
GALV SEATO (M26927)	Gibbon ape leukemia virus strain SEATO	Gibbon	✓	✓	✓	M26927
GALV-X	Gibbon ape leukemia virus strain X	Gibbon	✓	✓	✓	U60065
GALV SF (AF055063)	Gibbon ape leukemia virus strain San Francisco	Gibbon			✓	AF055063
GALV SEATO (AF055060)	Gibbon ape leukemia virus strain SEATO	Gibbon			✓	AF055060
WMV (AF055064)	Woolly monkey virus	Gibbon			✓	AF055064
GALV Brain (AF055062)	Gibbon ape leukemia virus strain Brain	Gibbon			✓	AF055062
GALV Hall's Island (AF055061)	Gibbon ape leukemia virus strain Hall's Island	Gibbon			✓	AF055061
KoRV-A (KF786280)	Koala retrovirus, variant A	Koala	✓	✓	✓	KF786280
KoRV-A (KF786284)	Koala retrovirus, variant A	Koala	✓	✓	✓	KF786284
KoRV-A (AF151794)	Koala retrovirus, variant A (strain "Cindy")	Koala	✓	✓	✓	AF151794
KoRV-A (AB721500)	Koala retrovirus, variant A (strain "Aki")	Koala	✓	✓	✓	AB721500
KoRV-B	Koala retrovirus, variant B (strain Br2-1CETTG)	Koala	✓	✓	✓	KC779547
KoRV-A (AB823238)	Koala retrovirus, variant A (strain OJ-4)	Koala			✓	AB823238
KoRV-C	Koala retrovirus, variant C (strain OJ-4)	Koala			✓	AB828005
KoRV-D	Koala retrovirus, variant D (strain OJ-4)	Koala			✓	AB828004
KoRV-J	Koala retrovirus, variant J (strain OJ-4)	Koala			✓	AB822553
MDEV	*Mus dunni* endogenous virus	Mouse	✓	✓	✓	AF053745
McERV	*Mus caroli* endogenous virus	Mouse	✓	✓	✓	KC460271
MmERV	*Mus musculus* retrovirus	Mouse	✓	✓	✓	AC005743
MbRV	*Melomys burtoni* retrovirus	Mouse		✓	✓	KF572483 to KF572486
PERV-A 1	Porcine endogenous retrovirus A	Pig	✓	✓	✓	AJ293656
PERV-A 2	Porcine endogenous retrovirus A	Pig	✓	✓	✓	HQ540592
PERV-B 1	Porcine endogenous retrovirus B	Pig	✓	✓	✓	HQ540593
PERV-B 2	Porcine endogenous retrovirus B	Pig	✓	✓	✓	AY099324
PERV-C 1	Porcine endogenous retrovirus C	Pig	✓	✓	✓	HQ536013
PERV-C 2	Porcine endogenous retrovirus C	Pig	✓	✓	✓	AM229311
PERV-C MSL	Porcine endogenous retrovirus MSL	Pig	✓	✓	✓	AF038600
RlRV	*Rousettus leschenaultii* retrovirus	Bat	✓	✓		JQ951957 to JQ951958
MlRV	*Megaderma lyra* retrovirus	Bat	✓	✓		JQ951955 to JQ951956
RfRV	*Rhinolophus ferrumequinum* retrovirus	Bat	✓	✓	✓	JQ303225
CrERV	*Odocoileus hemionus* endogenous virus	Mule deer	✓	✓	✓	JN592050
OOEV	*Orcinus orca* endogenous retrovirus	Killer whale	✓	✓	✓	GQ222416
BaEV	Baboon endogenous virus	Baboon	✓	✓	✓	D10032
RD114	Feline RD114 retrovirus	Cat	✓	✓	✓	EU030001
REV	Reticuloendotheliosis virus	Bird	✓	✓	✓	AY842951
PreXMRV-1	Prexenotropic MuLV-related virus 1	Mouse	✓	✓	✓	FR871849
M-CRV	Murine type C retrovirus	Mouse	✓	✓	✓	X94150
M-MuLV	Moloney murine leukemia virus	Mouse	✓	✓	✓	AF033811
F-MuLV	Friend murine leukemia virus	Mouse	✓	✓	✓	Z11128
R-MuLV	Rauscher murine leukemia virus	Mouse	✓	✓	✓	U94692
FeLV	Feline leukemia virus	Cat	✓	✓	✓	AF052723

attempt to precisely localize the coding regions and the regulatory motifs within the genome of each strain, the nucleotide sequence of each strain was compared to the annotated genomes available in GenBank of GALV-X (U60065) and SEATO (M26927) and of the closely related KoRV (AF151794). Each strain was characterized by the common genetic structure of simple type C mammalian retroviruses with a 5′ LTR-*gag*-*pol*-*env*-3′ LTR organization. Furthermore, the following regulatory motifs were readily identified in each strain: a tRNAPro primer binding site, a CAAT box, a TATA box, a Cys-His box, a polypurine tract, and a polyadenylation [poly(A)] signal. No differences in these motifs were detected among GALV strains with the exception of four polymorphisms in the Cys-His box, three of which were mutations unique to WMV (positions 2518, 2536, and 2539), along with a G-to-A (position 2530) transition and a C-to-G (position 2536) transversion, both found in GALV-X and SF (data not shown).

The 5′ and 3′ LTRs of the GALV strains were 463 to 559 bp long (Table 3) with a retrovirus-typical U3-R-U5 region structure (Fig. 2B and C). The 5′ and 3′ LTRs were compared for each strain and were found to be identical, further validating the sequencing and assembly methods used. The overall average nucleotide identity of LTRs across the GALV strains was 82.2%, lower than that calculated for the open reading frames (ORFs). However, between GALV-X and SF-HOS, the LTRs were 100% identical, and the

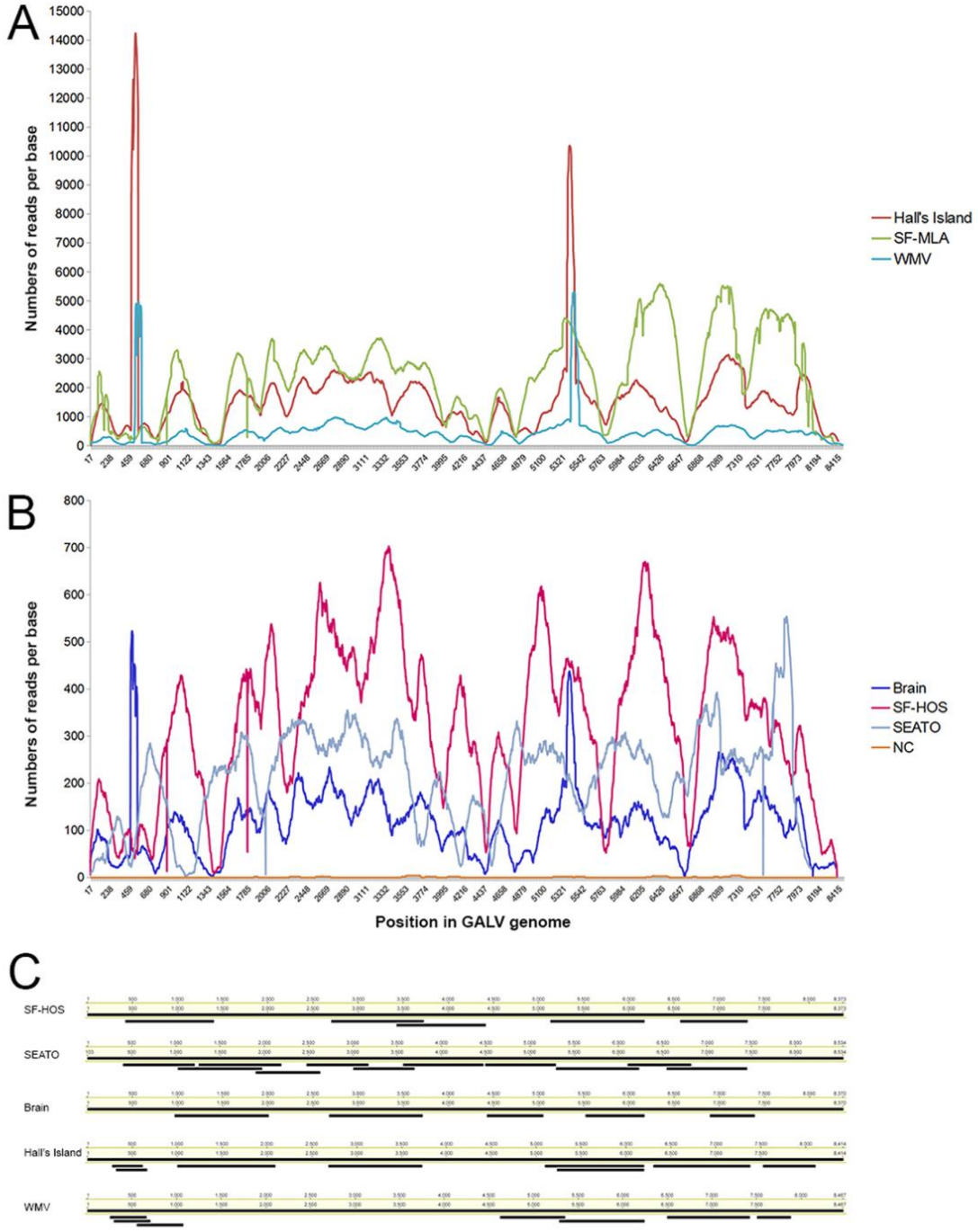


FIG 1 Hybridization capture sequence and Sanger sequence coverage across the proviral genome for the GALV strains. The sequence coverage is shown for each nucleotide position, numbered as in the corresponding strain consensus sequence. Mapping results for a negative control (NC) are also shown. Each sample is color coded. Panel A shows a coverage profile of the strains that reached very high values (up to 14,000 reads per base), while panel B shows the coverage profile of the strains with lower coverage (up to 700 reads per base). Panel C shows the position of each Sanger sequence generated by PCR in comparison to the full genome consensus sequences of the GALV strain from which it was generated. The Sanger sequences presented here were all of high quality and were used to confirm the bioinformatics assembly of sequences obtained by hybridization capture.

TABLE 3 Length and coordinates of the genomic regions of the GALV strains[a]

Strain	Total length (nt)	5′ LTR		*gag*			*pol*			*env*			3′ LTR	
		Length (nt)	Coordinates (nt)	Length (nt)	Coordinates		Length (nt)	Coordinates		Length (nt)	Coordinates		Length (nt)	Coordinates (nt)
					nt	aa		nt	aa		(nt)	aa		
SF-HOS	8,373	463	1–463	1,566	910–2475	1–521	3,384	2590–5973	1–1127	2,013	5855–7867	1–670	463	7911–8373
SEATO	8,534	463	1–463	1,563	954–2516	1–520	3,384	2631–6014	1–1127	2,058	5875–7932	1–685	559	7976–8534
Brain	8,370	453	1–453	1,572	899–2470	1–523	3,375	2585–5959	1–1124	2,046	5829–7874	1–681	453	7918–8370
Hall's Island	8,414	469	1–469	1,572	915–2486	1–523	3,384	2601–5984	1–1127	2,058	5845–7902	1–685	469	7946–8414
WMV	8,467	507	1–507	1,566	963–2528	1–521	3,384	2643–6026	1–1127	2,010	5908–7917	1–669	507	7961–8467

[a] aa, amino acids; nt, nucleotides.

LTRs of Brain and Hall's Island were similar (93.2% sequence identity) (Table 4). The differences among GALV strain LTRs were concentrated in the U3 region, which was the most variable (average identity, 75.8%). In addition to small insertions, deletions, and point mutations, there were three notable differences among the strains: (i) a 16-bp deletion at the 5′ end of the U3 region of the Brain strain (compared to GALV-X, positions 9 to 25); (ii) two fragments, 21 and 22 bp in length, present only in WMV (positions 52 to 72 and positions 101 to 122, respectively); and (iii) a 48-bp perfect tandem direct repeat present only in

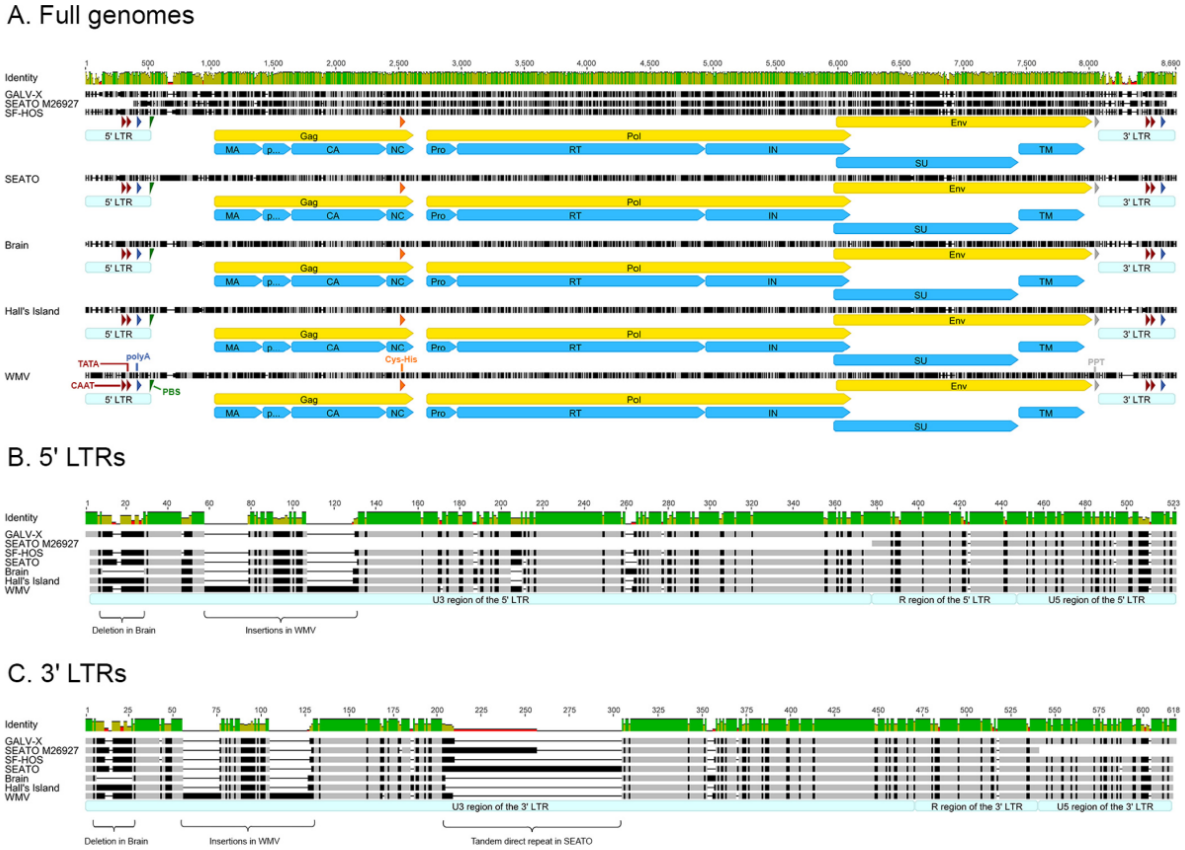


FIG 2 Genomic structure of the GALV strains. Alignment of the newly generated nucleotide sequences of the GALV strains with GALV GenBank reference sequences (SEATO, M26927; GALV-X, U60065). Panel A shows the full genomes of each GALV strain, with the positions of proviral genes, proteins, and regulatory motifs indicated. Panels B and C show the differences among the GALV strains in the 5′ and 3′ LTRs, respectively. Nucleotide positions identical among the strains are indicated in light gray, while mismatches are shown in black. Gaps are shown as dashes. The green bar above the alignment indicates the percent identity among the sequences (green, highest identity; red, lowest identity). The following structural regions are shown: the 5′ and 3′ LTRs with the typical U3-R-U5 structure (in light blue), the CAAT box and TATA box (in red), the polyadenylation [poly(A)] signal (in dark blue), the primer binding site (PBS) (in green), the Cys-His box (in orange), and the polypurine tract (PPT) (in gray). The ORFs of *gag*, *pol*, and *env* genes are shown in yellow, while protein domains are in sky blue. Protein domain abbreviations: MA, matrix; CA, capsid; NC, nucleocapsid; Pro, protease; RT, reverse transcriptase; IN, integrase; SU, surface unit; TM, transmembrane subunit.

TABLE 4 Similarities among the GALV strains in the LTRs and in the full genomes sequences[a]

Strain	LTRs (% identity)							Full genome (% identity)						
	GALV-X	SEATO*	SF-HOS	SEATO	Brain	Hall's Island	WMV	GALV-X	SEATO*	SF-HOS	SEATO	Brain	Hall's Island	WMV
GALV-X		84	100	83.5	80	82.9	80.8		87.6	99	87.1	88.2	88.3	90
SEATO*	75.2		84	100	87.7	87	90.3	87.6		87.9	98.7	91.4	91.5	89.6
SF-HOS	100	75.2		83.5	80	82.9	80.8	99	87.9		87.4	88.4	88.5	90.5
SEATO	69.6	89.5	69.6		84	88	81.9	87.1	98.7	87.4		90.9	91.1	89
Brain	79.5	74.5	79.5	69.9		93.2	79.4	88.2	91.4	88.4	90.9		97.7	89.9
Hall's Island	82.5	78.9	82.5	73.1	93.2		81.6	88.3	91.5	88.5	91.1	97.7		89.9
WMV	80.8	72.4	80.8	68.9	79.4	81.5		90	89.6	90.5	89	89.9	89.9	

[a] The similarities are reported as percent nucleotide identities between nucleotide sequences. For the LTRs, the values above the diagonal represent the percent nucleotide identities among the 5′ LTR sequences of GALV strains, whereas the values below the diagonal represent the percent identities among the 3′ LTR sequences. GALV reference sequences from GenBank (SEATO, M26927, indicated by SEATO*; GALV-X, U60065) are included in the comparison.

SEATO (positions 136 to 183), as previously reported (51) (Fig. 2B and C). The 48-bp motif is found in two copies in the 3′ LTR in the SEATO sequence from Delassus et al. (14) and Trainor et al. (51). However, in the current study different variants with two to four copies were observed among the Illumina sequences (three copies are reported in the 3′ LTR of the consensus sequence). The GenBank entry for SEATO (M26927) (14) does not include the first 320 bp of the 5′ LTR, and the data presented here fill in the genome sequence.

An imperfect 7-bp inverted repeat (e.g., TGAAAGA/TCT CTCA in SF-HOS), which is known to mark the boundaries of the LTR ends (18, 51), was identified in each strain with minor differences. An AAAAATAC motif, which was found to correlate with leukemogenicity in several MuLVs (52), was identified in SEATO, Brain, and Hall's Island GALVs. The insertions and deletions previously reported by Trainor et al. (51) in the U5 region of GALV strains, including a deletion affecting the poly(A) signal in SEATO, were not detected in the current study. In fact, among the GALV strains the U5 region was overall more conserved (85.2% sequence identity) than the U3 region (75.8%).

When the full nucleotide sequences were compared, all of the GALV strains demonstrated a high degree of similarity overall, with an average nucleotide identity of 90.6% (Table 4). Specifically, as expected, the SEATO sequence generated here was almost identical to the GenBank SEATO (98.7% identity), while SF-HOS shared 99% identity with GALV-X. The Brain and Hall's Island strains were very closely related (97.7% nucleotide identity) and together more similar to GenBank SEATO (average nucleotide identity, 91.4%) than to GALV-X (88.2%). WMV did not show strong affinity with any specific GALV strain, although identity with the other GALVs was high (89 to 90.5%, Table 4).

Three ORFs corresponding to the *gag*, *pol*, and *env* genes were identified in the genome of each GALV strain. The ORF average length was 1,568 bp (1,563 to 1,572 bp) for *gag*, 3,382 bp (3,375 to 3,384 bp) for *pol*, and 2,037 bp (2,010 to 2,058 bp) for *env*, indicating low ORF size variability among the GALV strains. All ORFs were undisrupted. The *gag* and *pol* ORFs were in the same reading frame, while *env* was in a different frame, with the end of *pol* and the beginning of *env* ORFs overlapping, as found in many retroviruses. The GALV strains displayed a 93.3% average amino acid similarity for *gag*, 96.2% for *pol*, and 87.6% for *env* (Table 5).

For each GALV strain, we identified the matrix p15 (MA), p12, capsid p30 (CA), and nucleocapsid p10 (NC) proteins within Gag; the protease (PR), reverse transcriptase (RT), and integrase (IN) proteins within Pol, and the surface unit gp70 (SU) and transmembrane subunit p15E (TM) within Env (Fig. 2A). Furin sites with the motif R-X-K-R for the cleavage of the Env precursor into SU and TM subunits were identified in each GALV strain at the C terminus of the SU. Also, the CWLC motif, which is thought to play a role in the assembly and function of the Env complex (53), was conserved across all GALV strains (positions 355 to 358 of the Env protein). Among Gag protein domains, the capsid was by far the most conserved among GALV strains with 98.5% amino acid identity, while the nucleocapsid was the most variable (85.6% among strain similarity). All Pol protein domains were highly conserved, while within Env the surface unit was much more variable than the transmembrane subunit (84.8 and 94.8% identity, respectively) (Table 5 and Fig. 2A). On average, 34% of the poly-

TABLE 5 Amino acid similarity among the GALV strains from this study for the Gag, Pol, and Env proteins

Strain	Similarity (% identity)[a]														
	Gag (avg, 93.3%)					Pol (avg, 96.2%)					Env (avg, 87.6%)				
	SF-HOS	SEATO	Brain	Hall's Island	WMV	SF-HOS	SEATO	Brain	Hall's Island	WMV	SF-HOS	SEATO	Brain	Hall's Island	WMV
SF-HOS		91.4	90.8	90.6	92.1		95.7	95.5	94.9	96.5		85.3	85	86.1	85.6
SEATO	91.4		96.4	96.6	92.5	95.7		96	95.5	96.5	85.3		92.8	94	83.1
Brain	90.8	96.4		97.7	92.7	95.5	96		99.2	96.4	85	92.8		97.8	82.8
Hall's Island	90.6	96.6	97.7		92.4	94.9	95.5	99.2		96	86.1	94	97.8		83.7
WMV	92.1	92.5	92.7	92.4		96.5	96.5	96.4	96		85.6	83.1	82.8	83.7	

[a] The similarities are reported as percent identities between amino acid sequences. The average amino acid similarity among strains for each of the protein is indicated in parentheses in the column heading. The average amino acid similarities among strains for each of the protein domains were as follows: (i) within Gag, p15 MA (89.14%), p12 (88.81), p30 CA (98.53), and p10 NC (85.6%); (ii) within Pol, Pro (97.31%), RT (96.44%), and IN (95.66%); and (iii) within Env, gp70 SU (84.83%) and p15e TM (94.8%). Abbreviations: MA, matrix; CA, capsid; NC, nucleocapsid; Pro, protease; RT, reverse transcriptase; IN, integrase; SU, surface unit; TM, transmembrane subunit.

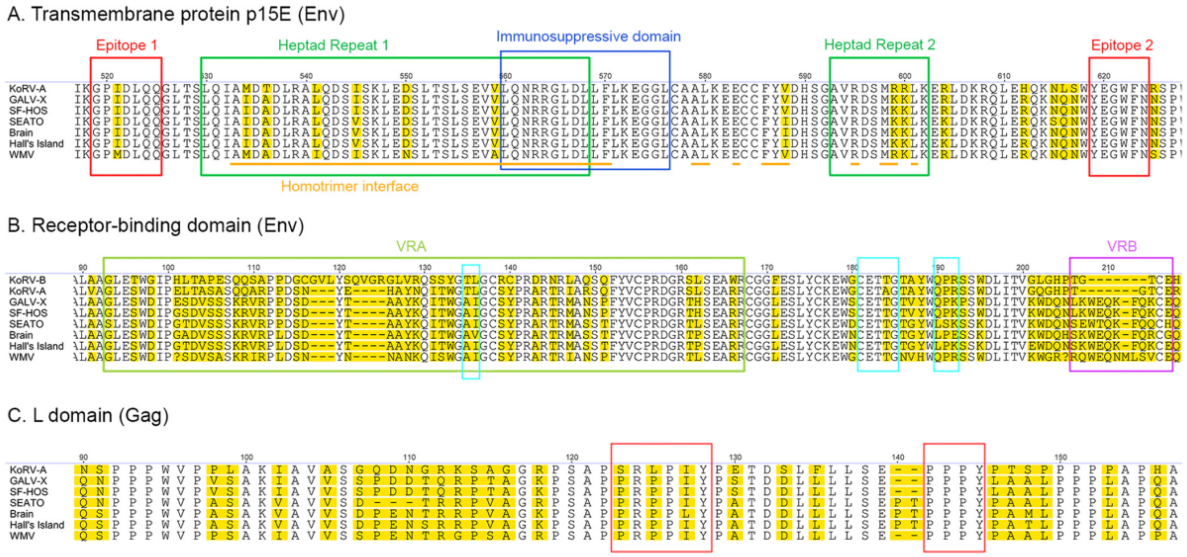


FIG 3 Differences among GALV strains and KoRV in the Env and Gag domains regulating viral fusion, infectivity, and host range. Alignment of Env and Gag amino acid sequences of GALV strains with relevant GenBank reference sequences (GALV-X, U60065; KoRV, AF151794) for the domains affecting viral fusion (epitopes 1 and 2, heptad repeats 1 and 2, homotrimer interface, and immunosuppressive domain of the transmembrane protein p15E of Env) (A), receptor specificity (variable regions A and B of Env) (B), and viral infectivity (receptor-binding domain of Env and L domain of Gag) (B and C, respectively). The three motifs influencing infectivity within the receptor-binding domain are marked by turquoise squares (B), while the PRPPIY and PPPY motifs are marked by brown squares within the L domain (C). Positions where amino acids vary are highlighted in yellow. Since KoRV-B was used to investigate the functional differences between GALV and KoRV in the VRA and VRB regions, KoRV-B has been included in panel B.

morphisms identified in Gag, Pol, and Env were mutations unique to SF-HOS (17.5, 22.2, and 16.3%, respectively) and WMV (12.5, 16.6, and 18.2%, respectively). These unique polymorphisms were concentrated in the p12 domain in Gag, in the integrase domain in Pol, and in the surface unit in Env.

The transmembrane protein p15E of the envelope is known to contain several motifs that are highly conserved among gammaretroviruses (54). The epitopes E1 (residues 519 to 525) and E2 (residues 619 to 624), the immunosuppressive domain (residues 560 to 576), the homotrimer interface (interspersed residues 533 to 601), and the heptad repeats 1 and 2 (residues 530 to 568 and residues 593 to 602, respectively) were conserved across all GALV strains (Fig. 3A). These domains are mainly involved in viral fusion and are highly conserved among GALVs, KoRVs, and PERVs (54). Nevertheless, one polymorphism each within the E1 and heptad repeat 2 and five polymorphisms in the overlapping region between heptad repeat 1 and the homotrimer interface were observed among GALVs. Six of the seven detected polymorphisms were identified in WMV. Of these six polymorphisms identified in WMV, two were shared with KoRV (Fig. 3A).

Differences in the variable regions A and B (VRA and VRB) of the receptor-binding domain (RBD) of the envelope protein are responsible for variation in receptor specificity for WMV and the other GALV strains (19). Sixteen polymorphisms in the VRA, and eight polymorphisms in the VRB were observed, as well as an insertion of one amino acid in the VRB of WMV compared to other GALVs (Fig. 3B). WMV, which is the only GALV strain to show a difference from other strains in the host range (it cannot infect E36 hamster cells), exhibited a high degree of diversification in these two regions, with an average of 13 amino acid residue differences in the VRA and of 8.5 amino acid residue differences in VRB sequences relative to other GALVs (Fig. 3B). Similarly to WMV, KoRV-A also fails to infect E-36 cells (Eiden, unpublished). Thus, the ability to infect hamster E36 cells is a distinguishing feature of the GALVs, with the exception of WMV. It has been previously shown that glycosylation does not account for the inability of WMV to use the E36 GALV receptors, and it has been postulated that cellular factors, such as the expression of inhibiting factors or the lack of accessory proteins, may be involved (19).

We also confirmed the high variability detected by Oliveira et al. (55) among GALV strains in the motifs of the RBD of the envelope protein, which are known to influence the differential infectivity of GALV and KoRV (55). All GALV strains presented the AI residues at positions 135 to 136 of the envelope surface unit, with the exception of Brain, which had AV at these positions. WMV was the only strain to show at residues 190 to 192 the same QPR residues displayed by KoRV (55) (Fig. 3B). Oliveira et al. (55) showed that when these five residues of the GALV envelope are replaced by the corresponding residues of KoRV, the resulting mutant vectors exhibit substantially reduced titers similar to those observed with KoRV vectors. In contrast, no polymorphisms among GALV strain envelopes were observed in the CETTG motif (residues 181 to 185 of the surface unit) (Fig. 3B), which is highly conserved among infectious gammaretroviruses, including KoRV-B, although is mutated in KoRV-A (55). It has been hypothesized that these mutations played a key role in the endogenization process of KoRV-A into the koala genome (55).

Few differences were observed among GALV strains in the PRPPIY and PPPY motifs of the L domain of the Gag protein (residues 123 to 128 and residues 142 to 145, respectively, of the

matrix protein) (Fig. 3C), which are known to play a key role in the release of viral particles from the plasma membrane after viral budding. Replacement of GALV PRPPIY with KoRV SRLPIY motif causes a substantial reduction in viral titer (55), while the disruption of the PPPY motif has been reported to be involved in the reduction of KoRV viral budding (56, 57). The only difference observed in the PRPPIY motif was an I-to-L residue replacement in GALV-Brain at the fifth position of the motif, while the PPPY motif was identical across all GALV strains (Fig. 3C). A high level of conservation was observed in the major homology region, which is the most conserved region among retroviruses of the Gag CA protein and whose residues are necessary for the proper assembly of mature capsids (58). Only one polymorphism (an A-to-T change in Brain) was found at the sixth position of the motif (VLQGPAEPPSVFLERLMEAY, positions 348 to 367 of the Gag protein).

Functional differences between GALV and KoRV VRA and VRB regions. The GALV polymorphisms identified within the VRA and VRB regions may have functional consequences for receptor binding. Within the KoRV/GALV group, KoRV-A and all GALVs use the sodium-dependent phosphate transporter 1 (PiT1) as a receptor (59), whereas KoRV-B and -J infect cells via the thiamine transporter 1 (THTR1) (36). In order to understand which part of the envelope of KoRV and GALV influences receptor specificity, we constructed vectors endowed with GALV-SEATO chimeric envelopes in which regions of the RBD were replaced by the corresponding region of KoRV-B (Fig. 4). These vectors were used to infect *Mus dunni* tail fibroblast (MDTF) cells. Murine MDTF cells are resistant to all KoRVs and GALVs, but the expression of PiT1 renders them susceptible to KoRV-A and GALVs but not KoRV-B, whereas the expression of THTR1 renders them susceptible to KoRV-B but not GALVs or KoRV-A (36). Chimeric vectors with a GALV envelope in which the GALV VRA was replaced by the VRA from KoRV-B failed to infect MDTF cells expressing PiT1 or THTR1 (Fig. 4A). However, when the GALV vector had both VRA and VRB replaced by the corresponding regions from KoRV-B, MDTF cells expressing THTR1 were successfully infected, and the vector titer was similar to that of vectors bearing the full-length KoRV-B envelope (Fig. 4A). Therefore, although KoRV-B VRA was by itself insufficient to confer infectivity, the combination of VRA and VRB was sufficient to confer the infectivity properties of KoRV-B to GALV. Binding studies involving MDTF cells expressing either PiT1 or THTR1 were conducted (Fig. 4B). These studies demonstrated that the reason why the vector bearing both KoRV-B VRA and VRB does not infect MDTF cells expressing PiT1 (Fig. 4A) is that this vector does not bind PiT1 (Fig. 4B). Therefore, the block to infection is not mediated at a postbinding stage of entry. Similarly, the inability of vector bearing only KoRV-B VRA to infect MDTF cells expressing THTR1 is due to the failure to bind THTR1 (Fig. 4B).

Phylogenetic and selection analysis of GALV strains. Nucleotide mismatches were observed between the sequences from GenBank and those generated in this study for the same GALV strain, many of the differences representing nonsynonymous substitutions. This was pronounced in *env* for which sequences of each GALV strain are available in GenBank. For example, we detected 24 nucleotide differences in the GALV Hall's Island *env*, 8 of which were nonsynonymous substitutions. All GALV GenBank sequences were generated more than 15 years ago (14, 18, 19) by Sanger sequencing, while the sequences reported here were confirmed both by hybridization capture and bidirectional Sanger sequencing with an updated BigDye chemistry kit (v3.1). In order to account for the potential of errors in the GenBank sequences, the selection analysis was run with and without the GenBank sequences. While the results of the selection analysis for *gag* and *pol* did not change, in *env* three GenBank-derived GALV sequences (Hall's Island AF055061, SEATO M26927, and SF AF055063) were found to have undergone episodic diversifying selection, whereas all other GALV tree terminal branches were not. Even though the GALV *env* GenBank sequences grouped with their strain counterparts from our sequences (data not shown), the evidence of episodic diversifying selection on the GenBank sequences is likely an artifact of either mistakes in the GenBank sequences or mutations that have occurred over time in cell culture. Therefore, the results of the evolutionary analyses on the *env* are presented without GALV GenBank sequences (Fig. 5).

All GALV strains formed a monophyletic clade sister to WMV, with the clade of the GALVs and WMV forming a sister group to

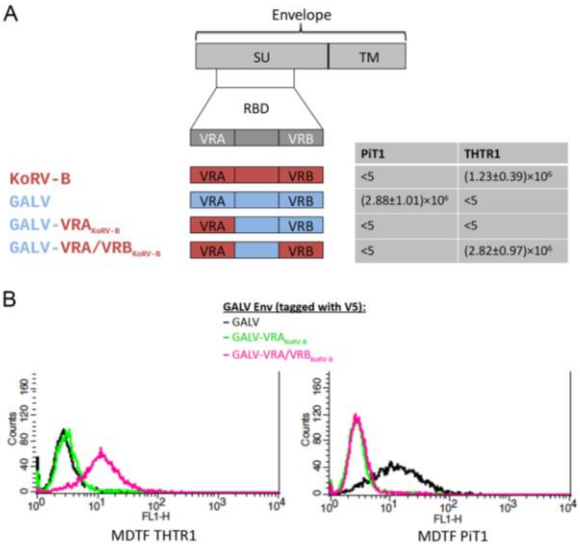


FIG 4 Exposition of murine MDTF cells expressing the receptor for GALV (PiT1) or the receptor for KoRV-B (THTR1) to vectors bearing different GALV/KoRV-B envelopes. The structure of a gammaretroviral envelope protein with the surface unit (SU) and the transmembrane subunit (TM) is schematically depicted at the top of the figure underneath which is a depiction of the receptor-binding domain (RBD) located within the surface unit gene. In the schematic representation of the chimeric envelopes, sequences from KoRV-B envelope are in red, and those from GALV-SEATO are in blue. The GALV chimeric envelope within which the VRA of GALV-SEATO was replaced by the corresponding region of KoRV-B is designated GALV-VRA$_{KoRV-B}$, whereas the GALV-SEATO chimeric envelope containing both KoRV-B VRA and VRB is designated GALV-VRA/VRB$_{KoRV-B}$. Murine MDTF cells expressing PiT1 or THTR1 were exposed to vectors bearing GALV, KoRV-B, GALV-VRA$_{KoRV-B}$, or GALV-VRA/VRB$_{KoRV-B}$ envelopes and assessed for susceptibility to these vectors using a conventional β-galactosidase assay. The titers of the viral vectors were averaged from at least three independent experiments and are expressed as mean numbers of β-galactosidase-expressing cells ± the SD of the mean. Panel B demonstrates the ability of GALV (black line), GALV-VRA$_{KoRV-B}$ (green line) and GALV-VRA/VRB$_{KoRV-B}$ (pink line) envelopes, each with a V5 epitope tag, to bind to MDTF cells expressing either PiT1 or THTR1. The binding ability of the vectors was assessed using flow cytometry.

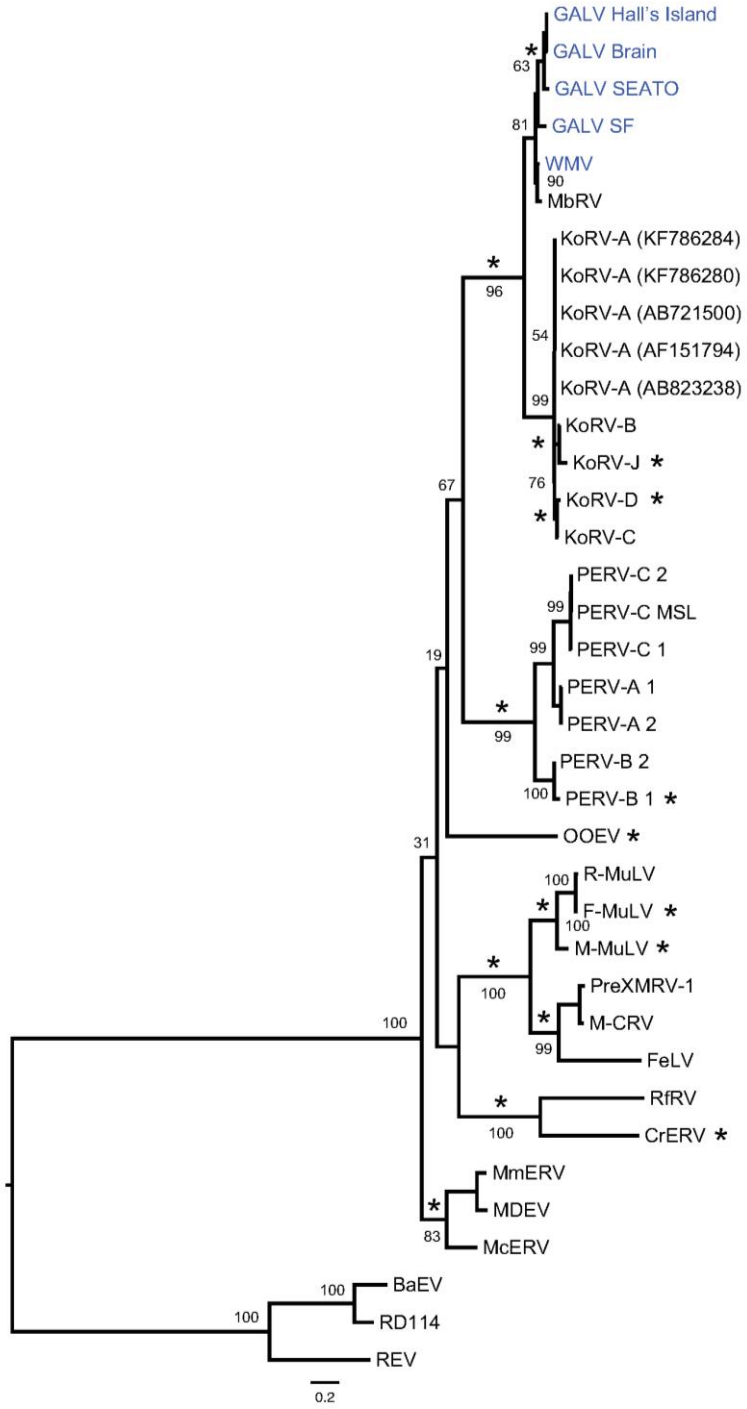


FIG 5 Maximum-likelihood phylogenetic tree of gammaretroviruses inferred using complete *env* nucleotide sequences, excluding GALV GenBank sequences. GALV GenBank sequences were excluded to avoid any influence of possible errors in these sequences on the analysis of selection. Node robustness was assessed with 500 rapid bootstrap pseudoreplicates. Numbers above or below the internode branches indicate bootstrap support. The GALV strain sequences generated in this study are highlighted in blue. Branches with significant ($P < 0.05$) evidence of episodic diversifying selection as indicated by the BSREL method are marked with an asterisk. GenBank accession codes are shown in brackets. The scale bar indicates 0.2 nucleotide substitutions per site. The tree is midpoint-rooted for purposes of clarity. All abbreviations can be found in Table 2.

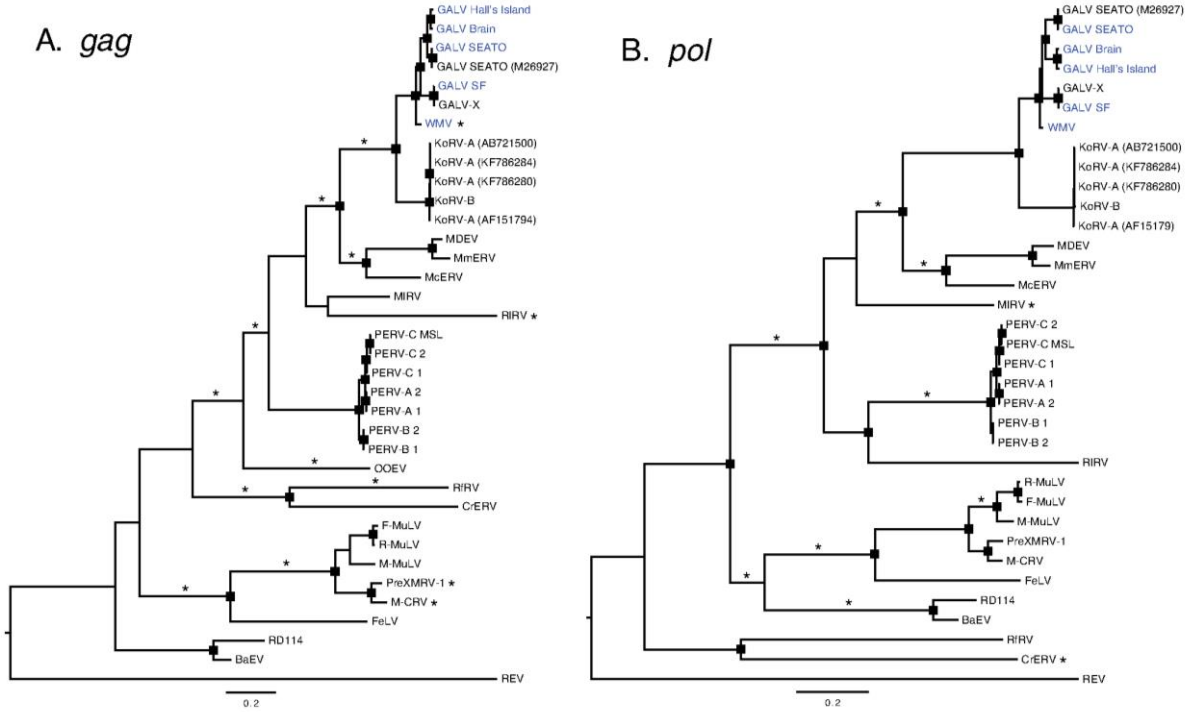


FIG 6 Maximum-likelihood phylogenetic trees of gammaretroviruses inferred using complete *pol* (A) and *gag* (B) nucleotide sequences. Node robustness was assessed with 500 rapid bootstrap pseudoreplicates. The rectangles on the nodes indicate a bootstrap support of >80%. The GALV strain sequences generated in this study are highlighted in blue. Branches with significant ($P < 0.05$) evidence of episodic diversifying selection, as indicated by the BSREL method, are marked with an asterisk. GenBank accession codes are shown in brackets. The scale bar indicates 0.2 nucleotide substitutions per site. The trees are midpoint rooted for purposes of clarity. All abbreviations can be found in Table 2.

the KoRVs, both at the nucleotide (Fig. 5 and Fig. 6) and amino acid level (data not shown). The highest level of internode branch support was observed when the protein sequences of the three protein-coding genes were concatenated and analyzed in a partitioned maximum likelihood framework (data not shown). The evolutionary relationships among GALV strains were robust regardless of the data type analyzed. Both concatenated, partitioned protein sequences (data not shown) and concatenated, partitioned nucleotide sequences (data not shown) (that included noncoding LTRs and spacers) grouped the two SEATO isolates, sister to the Brain and Hall's Island strains with the SF and X strains (98 to 100% bootstrap support).

Recombination was not detected in any of the protein-coding loci. Signs of positive diversifying selection were detected using the consensus results of MEME and FUBAR methods: only codons found to be under positive selection by both methods were considered. FUBAR detected only codons 98 and 360 under positive selection in the *env* gene, with a posterior probability (PP) > 0.97 and an empirical Bayes factor (EBF) of >180. By relaxing the PP threshold to 0.7 (EBF > 12), 11 more codons were found under episodic diversifying selection. MEME analysis identified many more codons (data not shown). The consensus consisted only in codon 98 of the *env* gene or with the relaxed threshold in codons 89, 96, 98, 211, 212, 282, 345, and 396. These codons correspond to residues 14, 21, 23, 118, 119, 154, 202, and 227, respectively, of the surface unit gp70 (SU) of the Env protein. Residues 118, 119, 154, and 202 represent four of the polymorphisms that we detected among GALV strains in the variable regions A and B (VRA/VRB) of the N-terminal region of the envelope and which are thought to influence the receptor specificity of these viruses. Although identified by both FUBAR and MEME, the codons identified by FUBAR only at a lower threshold should be treated with caution. We uncovered signs of episodic diversifying selection along the branches of the *gag*, *pol*, and *env* gene trees using the BSREL method (Fig. 5 and 6). A fraction of the codons of *gag* were found to deviate from purifying selection and neutrality on the branch unifying the GALV and KoRV clades and on the WMV terminal branch (Fig. 6). In the *env* gene, the branches connecting GALV-Hall's Island, Brain, and SEATO, and KoRV-B/KoRV-J strains were found to be under episodic diversifying selection (Fig. 5).

DISCUSSION

Because of its broad host range, GALV-based retroviral vectors have been developed for use in gene transfer (60). GALV has also been used in cancer gene therapy. GALV envelope fusogenic membrane glycoprotein (a C-terminal truncated form of GALV envelope glycoprotein, GALV.fus), which has strong cytotoxic effects, can be transduced into a range of human tumor cells to efficiently kill the cells through a process of syncytial formation (61). The use of this system in the treatment of lung cancer has already given encouraging results (62). In addition to its utility as

a clinical tool, GALV is an epizootic agent. Therefore, it is surprising that, with the exception of two strains, SEATO and GALV-X, most GALV laboratory strains have not been fully sequenced.

Hybridization capture advantages for viral genomics. Part of the difficulty in characterizing the GALV strains by PCR was the high failure rate of primer combinations given that the underlying diversity was unknown. Hybridization capture outperformed PCR amplification and Sanger sequencing in determining the uncharacterized genomic regions of the five GALV strains. PCR is subject to primer target mismatches and is sensitive to GC content. Hybridization capture, in contrast, can tolerate bait and target mismatches well over 15% (63). Multiplexing can be performed and yields high per-base coverage across the genome while allowing for discrimination of polymorphism or viral variant cooccurrence. The result was full coverage of all GALV strain genomes (Fig. 1A and B) with an average per-base fold coverage of 848.6. Where the Sanger sequencing and hybridization capture results overlapped, the sequences were identical. The consistency of results between Sanger sequencing and hybridization capture suggests that the capture results can be relied upon to yield the correct sequences. The GALV-SEATO and SF derived baits were suitable for examining viruses with up to 12.9% divergence and will likely be applicable to viruses with greater divergence, as observed by whole-genome cross hybridization experiments (64). Therefore, hybridization capture will likely be a valuable tool for viral discovery among closely and distantly related gammaretroviruses, which could be generally applied to retroviral discovery. However, when multiple similar viral strains are present in a sample, genome assembly can be hindered due to their sequence similarity. In our case it was not possible to recover the genome sequence of SF-MLA because of the presence of multiple distinct viral sequences. MLA-144 is a T-lymphoid cell line established from tumor cells of a gibbon with lymphoid leukemia (1). In contrast to other GALV cell lines, MLA-144 is thought to harbor several different defective recombinant GALV-SF proviruses, which may contain cell-derived, nonviral sequences (65). Furthermore, it was found that the MLA-144 cell line contains two GALV insertions in the *IL-2* gene, which allow the cell line to produce interleukin 2 constitutively (66). Together, these anomalies of the MLA-144 cell line hindered the capture experiment and complicated the assembly of the sequencing reads of SF-MLA.

Significance of genomic structural differences of GALV. As with other gammaretroviruses, malignancies induced by GALV or KoRV involve both viral and cellular determinants. The viral determinants include the transcription elements contained within the long terminal repeats (LTRs) and the envelope protein that affects cell tropism, *in vivo* spread and cytopathicity. Cellular determinants of infectivity and pathogenesis include viral receptors and cellular oncogenes activated by the adjacent integration of a transcriptionally active LTR. The only GALV sequences previously available in GenBank were the SEATO and GALV-X genome sequences (14, 18) and the envelope sequences of each GALV strain (19). Therefore, the sequences of the LTRs and the *gag* and *pol* genes were missing for most of the strains. Furthermore, the GenBank entry for SEATO (14) is chimeric, with part of the *pol* gene of SF strain incorporated into the SEATO genome, and also excludes the first 320 bp of the 5′ LTR. We have determined that the *env* gene of the GenBank SEATO is wrongly annotated since it does not include the sequence corresponding to the R peptide. Thus, the data presented in this study fill in these gaps in the SEATO genome completing its sequence. GALV-X was found to be almost identical to GALV-SF, suggesting that they could represent the same virus.

The five GALV strains showed high degree of similarity at the genome level with an average nucleotide identity above 90% (Table 4). However, we found high variability among the GALV strains in the LTRs (Table 4), especially in the U3 region. Notably, the insertions in the LTRs of WMV compared to the other strains and the 48-bp perfect tandem direct repeat present only in SEATO (Fig. 2B and C) are located in an area likely to contain transcriptional enhancers and could be relevant to the leukemogenic potential of these two strains, as already suggested (51). Of note, an AAAAATAC motif, reported by Villemur et al. (52) to be present specifically in the U3 of leukemogenic strains of MuLV, was identified in the LTRs of the SEATO, Brain, and Hall's Island strains.

At the amino acid level, the GALV strains demonstrated high degree of conservation in the *pol* and *gag* genes, with an average amino acid identity above 93% (Table 5). However, multiple distinct mutations could be identified in SF-HOS and WMV in both proteins, particularly in the p12 domain of Gag and in the integrase domain of Pol. The *env* gene was more variable, particularly in the surface unit (average amino acid identity 84.8%), which is known to contain motifs influencing viral infectivity (e.g., RBD) and receptor specificity (e.g., VRA/VRB). A high percentage of the polymorphisms were attributable to mutations found in SF-HOS and WMV in this domain. Functional analysis of differences, particularly between these two strains and the other GALVs may reveal further insights into the different biological properties of these viruses.

Until now only the *env* gene sequences were available for all the GALV strains, thus most functional analyses have been confined to domains within this gene. The only two determinants of infectivity identified in *gag*—the PRPPIY and PPPY motifs of the L domain, which are known to influence the release of viral particles from the plasma membrane after viral budding (56, 57)—were highly conserved across the GALV strains (Fig. 3C). The only exception was one amino acid difference found in Brain.

Our study confirmed the high degree of conservation in *env*, already highlighted among gammaretroviruses and specifically between KoRV and GALV (54), in the amino acid sequences of the domains and epitopes of the transmembrane envelope protein p15E that are important for viral fusion (Fig. 3A). The exception was WMV, which was variable in most motifs in comparison with other GALVs and shared some polymorphisms with KoRV. Similarly, WMV demonstrated unique amino acid changes relative to the other GALVs in the variable regions A and B (VRA and VRB) within the RBD of the envelope (Fig. 3B). These two regions are involved in receptor utilization and variation has been demonstrated to be responsible for the difference in host range between WMV and the other GALVs (19). Although both WMV and GALVs use PiT1 (SLC20A1) to infect human cells, WMV cannot infect hamster E36 cells that are susceptible to all other GALVs (19). The difference in host range is due to residues in the RBD of WMV (19). When GALV-SEATO RBD residues were substituted for the corresponding residues in WMV, the block to E36 infection was circumvented (19). A similar host range restriction extends to KoRV-A with respect to its inability to infect hamster E36 cells. The high degree of residue variation detected in the RBD region between WMV and KoRV-A and the other GALVs (Fig. 3B) sup-

ports the role of VRA and VRB in modulating receptor specificity (19).

Despite their genetic similarity, KoRV-B, unlike KoRV-A and the GALVs, does not use PiT1 as a receptor. THTR1 serves instead as KoRV-B receptor (36). Using chimeric envelopes derived from KoRV-B and GALV, we determined that both VRA and VRB comprising the RBD are required for GALV to switch to KoRV-B receptor usage (Fig. 4). Thus, we provided a second example among the KoRVs, WMV, and GALVs of the importance of RBD in receptor utilization.

Evolutionary analyses. Episodic diversifying selection is associated with selection pressure at the host-pathogen interface. Less pathogenic or endogenous retroviruses may be expected to elicit a less severe immune or antiretroviral response and exhibit reduced evidence of selection. Episodic diversifying selection was found to be acting on most of the gammaretroviral clades examined (Fig. 5 and 6). However, each gene exhibited a different pattern of selection. Selection on *gag* was observed on most clades except for the BaEV/RD114 clade (Fig. 6). There was also no evidence for specific selection on the GALV/KoRV lineages, even though the general clade to which GALV and KoRV belong is under selection. This was also true for the *pol* gene. In contrast, for the *env* gene there was evidence for selection on the GALV/KoRV clade, and specifically on the GALV Hall's Island/Brain/SEATO, KoRV-B/KoRV-J, and KoRV-C/D subclades (Fig. 5). In the case of the GALV strains under episodic diversifying selection, they represent some of the strains associated with leukemias in captive gibbons, GALV-SEATO and Hall's Island strains. The codon-oriented FUBAR and MEME analyses indicated that positive selection in these gammaretroviruses was concentrated on eight amino acids within the SU of the envelope, the most accessible portion of the virus to the immune system, supporting the potential involvement of host-pathogen interactions.

The only KoRVs exhibiting episodic diversifying selection are those associated with greater pathogenicity and which have switched receptor usage from Pit-1 to THTR1 (36, 67). In both cases it has been posited that these variants of KoRV are recently evolved strains that are exogenous (23, 36, 67). The concentration of selection in the *env* gene is consistent with analysis of historical koala KoRV-A derived sequences that suggest that the *env* gene is one of the few genes under longer-term selection, although weak (23, 68). The results are also consistent with our functional analysis of the importance of the VRA and VRB domains to receptor specificity in KoRV and GALV. The concentration of polymorphisms in the VRA and VRB regions among GALVs and the selective forces acting on the SU region of the *env* gene suggest that selection is strongly influencing GALV and KoRV interactions with host cells. The lack of observable positive selection on the KoRV-A clade is consistent with the endogenization of KoRV-A viruses in the koala genome (54).

Conclusions. Although most GALV strains are highly similar at the nucleotide and amino acid sequence level, WMV is the most divergent GALV, and it shares some traits with KoRV, i.e., host range and infectivity motifs in the *env* gene, which could explain the biological differences observed between WMV and other GALV strains. Episodic diversifying selection is concentrated on the Env protein likely as a consequence of adaptation to host immune responses. Among the GALVs and KoRVs, episodic diversifying selection acts most prominently on GALVs associated with leukemia in captive gibbons and KoRVs thought to be exogenous. Because viruses with affinity to GALVs are regularly being discovered in wildlife species such as rodents and bats (21, 69), our findings and the methods applied provide a comparative framework for analyzing GALV-like retroviruses as they are discovered. The full GALV strain genomes reported here provide a resource to functionally explore and augment or improve existing retroviral vector biology.

FUNDING INFORMATION

National Institute of Mental Health Intramural Research Program provided funding to Maribeth Eiden under grant number 1ZIAMH002592. International Max Planck Research School for Infectious Diseases and Immunology provided funding to Niccolo Alfano. HHS | NIH | National Institute of General Medical Sciences (NIGMS) provided funding to Sergios Kolokotronis, Alfred Roca, and Alex David Greenwood under grant number R01 GM092706. Morris Animal Foundation (MAF) provided funding to Alex David Greenwood under grant number D14ZO-94.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the National Institutes of Health. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. **Kawakami TG, Huff SD, Buckley PM, Dungworth DL, Synder SP, Gilden RV.** 1972. C-type virus associated with gibbon lymphosarcoma. Nat New Biol **235**:170–171. http://dx.doi.org/10.1038/newbio235170a0.
2. **Snyder SP, Dungworth DL, Kawakami TG, Callaway E, Lau DT.** 1973. Lymphosarcomas in two gibbons (*Hylobates lar*) with associated C-type virus. J Natl Cancer Inst **51**:89–94.
3. **DePaoli A, Johnsen DO, Noll MD.** 1973. Granulocytic leukemia in white handed gibbons. J Am Vet Med Assoc **163**:624–628.
4. **Kawakami TG, Buckley PM.** 1974. Antigenic studies on gibbon type-C viruses. Transplant Proc **6**:193–196.
5. **Gallo RC, Gallagher RE, Wong-Staal F, Aoki T, Markham PD, Schetters H, Ruscetti F, Valerio M, Walling MJ, O'Keeffe RT, Saxinger WC, Smith RG, Gillespie DH, Reitz MS, Jr.** 1978. Isolation and tissue distribution of type-C virus and viral components from a gibbon ape (*Hylobates lar*) with lymphocytic leukemia. Virology **84**:359–373. http://dx.doi.org/10.1016/0042-6822(78)90255-6.
6. **Reitz MS, Jr, Wong-Staal J-F, Haseltine WA, Kleid DG, Trainor CD, Gallagher RE, Gallo RC.** 1979. Gibbon ape leukemia virus-Hall's Island: new strain of gibbon ape leukemia virus. J Virol **29**:395–400.
7. **Todaro GJ, Lieber MM, Benveniste RE, Sherr CJ.** 1975. Infectious primate type C viruses: three isolates belonging to a new subgroup from the brains of normal gibbons. Virology **67**:335–343. http://dx.doi.org/10.1016/0042-6822(75)90435-3.
8. **Kawakami TG, Kollias GV, Jr, Holmberg C.** 1980. Oncogenicity of gibbon type-C myelogenous leukemia virus. Int J Cancer **25**:641–646. http://dx.doi.org/10.1002/ijc.2910250514.
9. **Theilen GH, Gould D, Fowler M, Dungworth DL.** 1971. C-type virus in tumor tissue of a woolly monkey (*Lagothrix* spp.) with fibrosarcoma. J Natl Cancer Inst **47**:881–889.
10. **Wolfe LG, Smith RK, Deinhardt F.** 1972. Simian sarcoma virus, type 1 (*Lagothrix*): focus assay and demonstration of nontransforming associated virus. J Natl Cancer Inst **48**:1905–1908.
11. **Hino S, Stephenson JR, Aaronson SA.** 1975. Antigenic determinants of the 70,000 molecular weight glycoprotein of woolly monkey type C RNA virus. J Immunol **115**:922–927.
12. **Rangan SR.** 1974. Antigenic relatedness of simian C-type viruses. Int J Cancer **13**:64–70. http://dx.doi.org/10.1002/ijc.2910130108.
13. **Reitz MS, Jr, Luczak JC, Gallo RC.** 1979. Mapping of related and nonrelated sequences of RNA from woolly monkey virus and gibbon ape leukemia virus. Virology **93**:48–56. http://dx.doi.org/10.1016/0042-6822(79)90274-5.
14. **Delassus S, Sonigo P, Wain-Hobson S.** 1989. Genetic organization of gibbon ape leukemia virus. Virology **173**:205–213. http://dx.doi.org/10.1016/0042-6822(89)90236-5.
15. **Fielding AK, Chapel-Fernandes S, Chadwick MP, Bullough FJ, Cosset FL, Russell SJ.** 2000. A hyperfusogenic gibbon ape leukemia envelope

glycoprotein: targeting of a cytotoxic gene by ligand display. Hum Gene Ther **11**:817–826. http://dx.doi.org/10.1089/10430340050015437.
16. **Bateman A, Bullough F, Murphy S, Emiliusen L, Lavillette D, Cosset FL, Cattaneo R, Russell SJ, Vile RG.** 2000. Fusogenic membrane glycoproteins as a novel class of genes for the local and immune-mediated control of tumor growth. Cancer Res **60**:1492–1497.
17. **Burtonboy G, Delferriere N, Mousset B, Heusterspreute M.** 1993. Isolation of a C-type retrovirus from an HIV infected cell line. Arch Virol **130**:289–300. http://dx.doi.org/10.1007/BF01309661.
18. **Parent I, Qin Y, Vandenbroucke AT, Walon C, Delferriere N, Godfroid E, Burtonboy G.** 1998. Characterization of a C-type retrovirus isolated from an HIV infected cell line: complete nucleotide sequence. Arch Virol **143**:1077–1092. http://dx.doi.org/10.1007/s007050050357.
19. **Ting YT, Wilson CA, Farrell KB, Chaudry GJ, Eiden MV.** 1998. Simian sarcoma-associated virus fails to infect Chinese hamster cells despite the presence of functional gibbon ape leukemia virus receptors. J Virol **72**:9453–9458.
20. **Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF.** 2000. The nucleotide sequence of koala (*Phascolarctos cinereus*) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus. J Virol **74**:4264–4272. http://dx.doi.org/10.1128/JVI.74.9.4264-4272.2000.
21. **Simmons G, Clarke D, McKee J, Young P, Meers J.** 2014. Discovery of a novel retrovirus sequence in an Australian native rodent (*Melomys burtoni*): a putative link between gibbon ape leukemia virus and koala retrovirus. PLoS One **9**:e106954. http://dx.doi.org/10.1371/journal.pone.0106954.
22. **Maricic T, Whitten M, Paabo S.** 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One **5**:e14004. http://dx.doi.org/10.1371/journal.pone.0014004.
23. **Tsangaras K, Siracusa MC, Nikolaidis N, Ishida Y, Cui P, Vielgrader H, Helgen KM, Roca AL, Greenwood AD.** 2014. Hybridization capture reveals evolution and conservation across the entire koala retrovirus genome. PLoS One **9**:e95633. http://dx.doi.org/10.1371/journal.pone.0095633.
24. **Meyer M, Kircher M.** 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protoc pdb.prot5448. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
25. **Alfano N, Courtiol A, Vielgrader H, Timms P, Roca AL, Greenwood AD.** 2015. Variation in koala microbiomes within and between individuals: effect of body region and captivity status. Sci Rep **5**:10189. http://dx.doi.org/10.1038/srep10189.
26. **Gouy M, Guindon S, Gascuel O.** 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol **27**:221–224. http://dx.doi.org/10.1093/molbev/msp259.
27. **Saitou N, Nei M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**:406–425.
28. **Hasegawa M, Kishino H, Yano T.** 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol **22**:160–174. http://dx.doi.org/10.1007/BF02101694.
29. **Martin M.** 2012. Cutadapt removes adapter sequences from high-throughput sequencing reads. Bioinformatics Action **17**:10–12.
30. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**:2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.
31. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv **1303**:3997.
32. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25**:2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.
33. **DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ.** 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43**:491–498. http://dx.doi.org/10.1038/ng.806.
34. **Magoc T, Salzberg SL.** 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics **27**:2957–2963. http://dx.doi.org/10.1093/bioinformatics/btr507.
35. **Lander MR, Chattopadhyay SK.** 1984. A Mus dunni cell line that lacks sequences closely related to endogenous murine leukemia viruses and can be infected by ectropic, amphotropic, xenotropic, and mink cell focus-forming viruses. J Virol **52**:695–698.
36. **Xu W, Stadler CK, Gorman K, Jensen N, Kim D, Zheng H, Tang S, Switzer WM, Pye GW, Eiden MV.** 2013. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proc Natl Acad Sci U S A **110**:11547–11552. http://dx.doi.org/10.1073/pnas.1304704110.
37. **Bryksin AV, Matsumura I.** 2010. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. Biotechniques **48**:463–465. http://dx.doi.org/10.2144/000113418.
38. **Abascal F, Zardoya R, Telford MJ.** 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res **38**:W7–W13. http://dx.doi.org/10.1093/nar/gkq291.
39. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30**:772–780. http://dx.doi.org/10.1093/molbev/mst010.
40. **Lanave C, Preparata G, Saccone C, Serio G.** 1984. A new method for calculating evolutionary substitution rates. J Mol Evol **20**:86–93. http://dx.doi.org/10.1007/BF02101990.
41. **Dimmic MW, Rest JS, Mindell DP, Goldstein RA.** 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol **55**:65–73. http://dx.doi.org/10.1007/s00239-001-2304-y.
42. **Yang Z.** 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39**:306–314. http://dx.doi.org/10.1007/BF00160154.
43. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033.
44. **Stamatakis A, Hoover P, Rougemont J.** 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst Biol **57**:758–771. http://dx.doi.org/10.1080/10635150802429642.
45. **Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A.** 2010. How many bootstrap replicates are necessary? J Comput Biol **17**:337–354. http://dx.doi.org/10.1089/cmb.2009.0179.
46. **Bruen TC, Philippe H, Bryant D.** 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics **172**:2665–2681.
47. **Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL.** 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet **8**:e1002764. http://dx.doi.org/10.1371/journal.pgen.1002764.
48. **Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K.** 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol **30**:1196–1205. http://dx.doi.org/10.1093/molbev/mst030.
49. **Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K.** 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol **28**:3033–3043. http://dx.doi.org/10.1093/molbev/msr125.
50. **Kircher M, Sawyer S, Meyer M.** 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res **40**:e3. http://dx.doi.org/10.1093/nar/gkr771.
51. **Trainor CD, Scott ML, Josephs SF, Fry KE, Reitz MS, Jr.** 1984. Nucleotide sequence of the large terminal repeat of two different strains of gibbon ape leukemia virus. Virology **137**:201–205. http://dx.doi.org/10.1016/0042-6822(84)90025-4.
52. **Villemur R, Rassart E, DesGroseillers L, Jolicoeur P.** 1983. Molecular cloning of viral DNA from leukemogenic Gross passage A murine leukemia virus and nucleotide sequence of its long terminal repeat. J Virol **45**:539–546.
53. **Pinter A, Kopelman R, Li Z, Kayman SC, Sanders DA.** 1997. Localization of the labile disulfide bond between SU and TM of the murine leukemia virus envelope protein complex to a highly conserved CWLC motif in SU that resembles the active-site sequence of thiol-disulfide exchange enzymes. J Virol **71**:8073–8077.
54. **Ishida Y, McCallister C, Nikolaidis N, Tsangaras K, Helgen KM, Greenwood AD, Roca AL.** 2015. Sequence variation of koala retrovirus transmembrane protein p15E among koalas from different geographic regions. Virology **475**:28–36. http://dx.doi.org/10.1016/j.virol.2014.10.036.
55. **Oliveira NM, Satija H, Kouwenhoven IA, Eiden MV.** 2007. Changes in viral protein function that accompany retroviral endogenization. Proc Natl Acad Sci U S A **104**:17506–17511. http://dx.doi.org/10.1073/pnas.0704313104.
56. **Demirov DG, Freed EO.** 2004. Retrovirus budding. Virus Res **106**:87–102. http://dx.doi.org/10.1016/j.virusres.2004.08.007.

57. **Shojima T, Hoshino S, Abe M, Yasuda J, Shogen H, Kobayashi T, Miyazawa T.** 2013. Construction and characterization of an infectious molecular clone of koala retrovirus. J Virol **87:**5081–5088. http://dx.doi.org/10.1128/JVI.01584-12.
58. **Purdy JG, Flanagan JM, Ropson IJ, Rennoll-Bankert KE, Craven RC.** 2008. Critical role of conserved hydrophobic residues within the major homology region in mature retroviral capsid assembly. J Virol **82:**5951–5961. http://dx.doi.org/10.1128/JVI.00214-08.
59. **Oliveira NM, Farrell KB, Eiden MV.** 2006. In vitro characterization of a koala retrovirus. J Virol **80:**3104–3107. http://dx.doi.org/10.1128/JVI.80.6.3104-3107.2006.
60. **Miller AD, Garcia JV, von Suhr N, Lynch CM, Wilson C, Eiden MV.** 1991. Construction and properties of retrovirus packaging cells based on gibbon ape leukemia virus. J Virol **65:**2220–2224.
61. **Higuchi H, Bronk SF, Bateman A, Harrington K, Vile RG, Gores GJ.** 2000. Viral fusogenic membrane glycoprotein expression causes syncytium formation with bioenergetic cell death: implications for gene therapy. Cancer Res **60:**6396–6402.
62. **Zhu B, Yang JR, Jiang YQ, Chen SF, Fu XP.** 2014. Gene therapy of lung adenocarcinoma using herpes virus expressing a fusogenic membrane glycoprotein. Cell Biochem Biophys **69:**583–587. http://dx.doi.org/10.1007/s12013-014-9836-4.
63. **Mason VC, Li G, Helgen KM, Murphy WJ.** 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Res **21:**1695–1704. http://dx.doi.org/10.1101/gr.120196.111.
64. **Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard JM, Poinar HN.** 2014. Ancient whole genome enrichment using baits built from modern DNA. Mol Biol Evol **31:**1292–1294. http://dx.doi.org/10.1093/molbev/msu074.
65. **Eiden M, Trainor CD, Reitz MS.** 1986. Gibbon ape leukaemia virus RNA in leukaemic T-lymphoid cell lines: expression of a novel RNA transcript. J Gen Virol **67**(Pt 7):1455–1460.
66. **Chen SJ, Holbrook NJ, Mitchell KF, Vallone CA, Greengard JS, Crabtree GR, Lin Y.** 1985. A viral long terminal repeat in the interleukin 2 gene of a cell line that constitutively produces interleukin 2. Proc Natl Acad Sci U S A **82:**7284–7288. http://dx.doi.org/10.1073/pnas.82.21.7284.
67. **Shojima T, Yoshikawa R, Hoshino S, Shimode S, Nakagawa S, Ohata T, Nakaoka R, Miyazawa T.** 2013. Identification of a novel subgroup of koala retrovirus from koalas in Japanese zoos. J Virol **87:**9943–9948. http://dx.doi.org/10.1128/JVI.01385-13.
68. **Avila-Arcos MC, Ho SY, Ishida Y, Nikolaidis N, Tsangaras K, Honig K, Medina R, Rasmussen M, Fordyce SL, Calvignac-Spencer S, Willerslev E, Gilbert MT, Helgen KM, Roca AL, Greenwood AD.** 2013. One hundred twenty years of koala retrovirus evolution determined from museum skins. Mol Biol Evol **30:**299–304. http://dx.doi.org/10.1093/molbev/mss223.
69. **Cui J, Tachedjian G, Tachedjian M, Holmes EC, Zhang S, Wang LF.** 2012. Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats. J Gen Virol **93:**2037–2045. http://dx.doi.org/10.1099/vir.0.043760-0.

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

ERRATUM

Erratum for Alfano et al., Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses

Niccolò Alfano,[a] Sergios-Orestis Kolokotronis,[b,c] Kyriakos Tsangaras,[a] Alfred L. Roca,[d] Wenqin Xu,[e] Maribeth V. Eiden,[e] Alex D. Greenwood[a,f]

Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany[a]; Department of Biological Sciences, Fordham University, Bronx, New York, USA[b]; Sackler Institute for Comparative Genomics and Division of Invertebrate Zoology, American Museum of Natural History, New York, New York, USA[c]; Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA[d]; Section on Directed Gene Transfer, Laboratory of Cellular and Molecular Regulation, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland, USA[e]; Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany[f]

Volume 90, no. 4, p. 1757–1772, 2016. Page 1768, Fig. 6 legend, line 1: "*pol* (A) and *gag* (B)" should read "*gag* (A) and *pol* (B)."

Citation Alfano N, Kolokotronis S-O, Tsangaras K, Roca AL, Xu W, Eiden MV, Greenwood AD. 2016. Erratum for Alfano et al., Episodic diversifying selection shaped the genomes of gibbon ape leukemia virus and related gammaretroviruses. J Virol 90:4254. doi:10.1128/JVI.00210-16.

Chapter IV

An endogenous gibbon ape leukemia virus (GALV) identified in a rodent (*Melomys burtoni* subsp.) from Wallacea

In review *Journal of Virology*

http://dx.doi.org/10.1128/JVI.00723-16

An endogenous gibbon ape leukemia virus (GALV) identified in a rodent (*Melomys burtoni* subsp.) from Wallacea

Niccolo Alfano[1], Johan Michaux[2,3], Serge Morand[4], Ken Aplin[5], Kyriakos Tsangaras[1*], Ulrike Löber[1], Pierre-Henri Fabre[5,6], Yuli Fitriana[7], Gono Semiadi[7], Yasuko Ishida[8], Kristofer M. Helgen[5], Alfred L. Roca[8], Maribeth V. Eiden[9], Alex D. Greenwood[1,10#]

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

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

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

[4] CIRAD-CNRS, Centre d'Infectiologie Christophe Mérieux du Laos, Vientiane, Lao PDR

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

[6] Institut des Sciences de l'Evolution de Montpellier (ISEM - UMR 5554 UM2-CNRS-IRD), Montpellier University, Montpellier, France

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

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

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

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

Running Title: Discovery of a GALV in a Wallacean rodent species

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

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

ABSTRACT

Gibbon ape leukemia virus (GALV) and koala retrovirus (KoRV) most likely originated from a cross-species transmission of an ancestral retrovirus into koalas and gibbons via one or more intermediate as yet unknown hosts. A highly similar virus to GALV has been identified in an Australian rodent (*Melomys burtoni*) after extensive screening of Australian wildlife. GALV-like viruses have also been discovered in several Southeast Asian species although screening has not been extensive and viruses discovered to date are only distantly related to GALV. We therefore screened 26 Southeast Asian rodent species for KoRV- and GALV-like sequences, using hybridization capture and high-throughput sequencing, in the attempt to identify potential GALV and KoRV hosts. Only the individuals belonging to a population of *Melomys burtoni* from Wallacea were positive yielding an endogenous provirus very closely related to a strain of GALV. The sequence of the critical receptor domain for GALV infection in the Wallacean *M. burtoni* subsp. was consistent with the susceptibility of the species to GALV infection. The second record of a GALV in *M. burtoni* provides further evidence that *M. burtoni*, and potentially other lineages within the widespread subfamily Murinae, may play a role in the spread of GALV-like viruses. The discovery of such a close GALV relative in the most western part of the Australo-Papuan distribution of *M. burtoni*, specifically in a transitional zone between Asia and Australia, may be relevant to the cross-species transmission to gibbons in Southeast Asia and broadens the known distribution of GALVs in wild rodents.

IMPORTANCE

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

INTRODUCTION

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

Mobile species such as bats, birds or commensal rats have been proposed as potential intermediate hosts of GALV and KoRV (9, 14). Bats can fly and disperse rapidly; they have been linked to the spread of several zoonotic diseases (17) and some Southeast Asian species harbor retroviruses related to GALV and KoRV (18). Rodents, however, are plausible intermediate hosts as they have migrated from Southeast Asia to Australia multiple times with several Southeast Asian species having established themselves in Australia (19). Furthermore, endogenous retroviruses related to GALV have been reported to be present in the genome of several Southeast Asian rodents such as *Mus caroli*, *Mus cervicolor* and *Vandeleuria oleracea* (20-22). However, these reports were based on DNA hybridization techniques and sequences were not reported. In 2008, the full genome sequence of an endogenous retrovirus found in the genome of *Mus caroli* (McERV) was reported (23). Despite the relatively high similarity to the genomic sequences of GALV and KoRV, McERV has a different host range and uses a different receptor, and therefore it is unlikely a progenitor of GALV and KoRV (23). McERV is most closely related to *Mus dunni* endogenous

virus (MDEV) (24) and the *Mus musculus* endogenous retrovirus (MmERV) (25), which together form a sister clade to the KoRV/GALV clade (2). Recently Simmons et al. (16) discovered fragments belonging to a retrovirus closely related to GALV and KoRV in the Australian native rodent *Melomys burtoni* (MbRV). MbRV sequence share 93 and 83% nucleotide identity with GALV and KoRV respectively, and *Melomys burtoni* overlaps with the geographic distribution of koalas. However, it is hard to explain how this Australian murid species could have come in contact with gibbons in Southeast Asia. Consequently it is unlikely that MbRV represents the direct or immediate ancestor virus of KoRV and GALV (16).

The aim of this work was to screen a wide range of rodent species from Southeast Asia for the presence of KoRV and GALV-like sequences and characterize polymorphisms in their viral receptor proteins in the attempt to identify the intermediate host(s) of KoRV and GALV using a non-PCR based approach called hybridization capture (26, 27). We focused on Southeast Asian rodent species since 42 Australian vertebrate species were previously screened, with MbRV the only virus identified (16), and most of the rodent species with GALV-like sequences identified are from Southeast Asia suggesting that GALVs and KoRVs may be circulating naturally in rodent populations residing there. Twenty-six rodent species were screened of which only a newly identified Australasian subspecies of *Melomys burtoni*, in the process of being taxonomically described and geographically reported (Fabre et al. unpublished data), was positive for a GALV sequence distinct from MbRV and none were positive for KoRV-like sequences. Specifically, this new subspecies has been discovered in the biogeographical region comprising a group of mainly Indonesian islands between the Asian and Australian continental shelves and called Wallacea (Fabre et al. unpublished data). We report the complete nucleotide sequence of the identified GALV-like virus, which we term *Melomys* Woolly Monkey Virus (MelWMV), its genomic structure, and its phylogenetic relationships with other related gammaretroviruses. We also examine GALV receptor variation among permissive and restrictive hosts including species belonging to the genus *Melomys*.

MATERIALS AND METHODS

Sample collection

The rodents used for the screening for GALV and KoRV were captured using folding rat traps during fieldwork expeditions in Southeast Asia and Asia in the periods January-February 2010, June-July 2010 and September 2013. Muscle samples were collected and

conserved in ethanol. All 49 samples belonging to the 26 species analyzed in the current study are listed in table 1. For the sequencing of the receptor of GALV, a blood sample was collected from a male white-handed gibbon (*Hylobates lar*) from Nuremberg zoo, Germany, during a routine health check on 24th July 1996.

Ethics statement

All animal experiments were performed according to the directive 2010/63/EEC on the Protection of Animals Used for Experimental and Other Scientific Purposes. The animal work also complied with the French law (nu 2012–10 dated 05/01/2012 and 2013-118 dated 01/02/2013). The rodents were captured using Sherman traps and the study of the species used in this project did not require the approval of an ethics committee (European directives 86-609 CEE and 2010/63/EEC). The species used are not protected, and no experiment was performed on living animals. No permit approval was needed as the species were trapped outside any preserved areas (national parks or natural reserves). The rodents were euthanized by vertebrate dislocation immediately after capture in agreement with the legislation and the ethical recommendations (2010/63/EEC annexe IV) (see also protocol available on http://www.ceropath.org/references/rodent_protocols_book). All experimental protocols involving animals were carried out by qualified personnel (accreditation number of the Center of Biology and Management of the Populations (CBGP) for wild and inbred animal manipulations: A34-1691). For the samples from Laos and Thailand, approval notices for trapping and investigation of rodents were provided by the Ministry of Health Council of Medical Sciences, National Ethics Committee for Health Research (NHCHR) Lao PDR, number 51/NECHR, and by the Ethical Committee of Mahidol University, Bangkok, Thailand, number 0517.1116/661. Oral agreements for trappings from obtained for local community leaders and land owners. Aplin's rodent sampling in Southeast Asia was carried out under CSIRO Sustainable Ecosystems Animal Ethics Committee Approval Numbers 00/01-27, 00/01-28 and 02/03-18. For the samples from Indonesia, rodent capture and handling in the field followed animal care and use guidelines recommended by the American Society of Mammalogists (28). Permits to collect scientific specimens were requested and provided by the State Ministry of Research and Technology (RISTEK) and the Ministry of Forestry, Republic of Indonesia. Specimens were prepared in the field by Museum Zoologicum Bogoriense personnel.

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

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

	Melomys burtoni subsp.	Indonesia	WD283
	Melomys burtoni subsp.	Indonesia	WD310
	Melomys burtoni subsp.	Indonesia	WD144
	Melomys burtoni subsp.	Indonesia	WD279
26	*Melomys paveli*	Indonesia	YS284

Cell lines, viruses and DNA extraction

GALV DNA for hybridization capture bait generation (26, 27) was obtained from the following productively infected cell lines: SEATO-88, GALV-SEATO infected Tb 1 Lu bat lung fibroblasts (ATCC CCL-88); GALV-4-88, GALV-Brain infected Tb 1 Lu bat lung fibroblasts (ATCC CCL-88); 71-AP-1, WMV infected marmoset fibroblasts; 6G1-PB, GALV-Hall´s Island infected lymphocytes; HOS (ATCC CRL-1543) GALV-SF infected human osteosarcoma cells. Genomic DNA extraction from the cell lines was performed using the Wizard Genomic DNA Purification Kit (Promega), following the manufacturer's protocol. Rodent tissue samples were first homogenized using a Precellys 24 (Bertin Technologies), with genomic DNA then extracted using the QIAamp DNA mini kit (QIAGEN) according to manufacturer's instructions. The genomic DNA of the white-handed gibbon was extracted following the method described in Sambrook and Russell (29). For all DNA extracts, DNA concentration was determined using the dsDNA High Sensitivity Assay Kit on a Qubit 2.0 fluorometer (Invitrogen).

Illumina library preparation

All rodent sample DNA extracts were sheared using a Covaris S220 (Covaris) to an average size of 300-bp prior to building Illumina sequencing libraries. Libraries were generated as described in Meyer and Kircher (30) with the modifications described in Alfano et al. (31), except for using a variable starting amount of DNA extract according to each sample availability and using 1 µl Illumina adapter mix (20 µM) in the adapter ligation step. Each library contained a unique combination of index adapters, one at each end of the library molecule (double-indexing) (32), to allow for subsequent discrimination among samples after the sequencing of pooled libraries. Negative control extraction libraries were also prepared and indexed separately to monitor for experimental cross contamination. Each library was amplified in three replicate reactions to minimize amplification bias in individual PCRs. The amplifications of the libraries were performed using Herculase II Fusion DNA polymerase (Agilent Technologies) in 50 µl volume reactions, with the cycling conditions of 95°C for 5 min, followed by 7 cycles of 95°C for 30 s, 60°C for 30 s, 72°C for 40 s and finally 72°C for 7 min. After pooling the three replicate PCR products for each sample, amplified libraries were

purified using the QIAquick PCR Purification Kit (QIAGEN) and quantified using a 2200 TapeStation (Agilent Technologies) on D1K ScreenTapes. Additional amplification cycles were performed for some of the libraries, when needed to balance library concentrations, using Herculase II Fusion DNA polymerase with P5 and P7 Illumina library outer primers with the same cycling conditions.

Hybridization capture baits

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

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

Hybridization capture

The 50 rodent indexed libraries were pooled in groups of 5 in order to reach a library input of 2 µg for each capture reaction. The negative controls for library preparation were also included in the capture reactions. Each indexed library pool was mixed with blocking oligos (200 µM) to prevent crosslinking of Illumina library adapters, Agilent 2× hybridization buffer, Agilent 10× blocking agent, and heated at 95°C for 3 min to separate the DNA strands (26). Each hybridization mixture was then combined with the biotinylated bait bound streptavidin beads. Samples were incubated in a mini rotating incubator (Labnet) for 48 hours at 65°C. After 48 hours the beads were washed to remove off-target DNA as described previously (26) and the hybridized libraries eluted by incubating at 95°C for 3 min. The DNA concentration for each captured sample was measured using the 2200 TapeStation on D1K ScreenTapes and further amplified accordingly using P5 and P7 Illumina outer primers (30). The enriched amplified libraries were then pooled in equimolar amounts to a final library concentration of 4.5 nM for paired-end sequencing (2×250) on an Illumina MiSeq platform with the v2 reagents kit at the Berlin Centre for Genomics in Biodiversity Research (BeGenDiv).

Genome sequence assembly

A total of 12,502,407 paired-end sequence reads 250-bp long were generated (average = 250,046.8 paired-end reads per sample, SD = 113,859.9) and sorted by their double indexes sequences. Cutadapt v1.2.1 (33) and Trimmomatic v0.27 (34) were used to remove adaptor sequences and low-quality reads using a quality cutoff of 20 and a minimal read length of 30 nt. After trimming, 97.6% of the sequences were retained. Thereafter reads were aligned to the NCBI nucleotide database using BLASTn (35) and the taxonomic profile of BLAST results were visualized using Krona (36) in order to assess the taxonomic content of the captured libraries. Reads were then mapped to the genome sequences of GALV strains (KT724047-51), KoRV (AF151794) and closely related gammaretroviruses (McERV - KC460271; MDEV - AF053745; MmERV - AC005743) using BWA v0.7.10 with default parameters (BWA-MEM algorithm)(37). The alignments were further processed using Samtools v1.2 (38) and Picard (http://broadinstitute.github.io/picard) for sorting and removal

of potential duplicates, respectively. Mapping was used as a preliminary screen to identify samples potentially positive or negative for viral sequences. Only samples that produced reads mapping across the genome of a viral reference were considered positive and subjected to further analyses. Samples that exhibited reads mapping only to limited portions of the reference, likely due to random homology of part of the bait to host genomic regions, were not further considered. Reads from positive samples were mapped to the reference of interest and the resulting alignments visualized and manually curated using Geneious v7.1.7 (http://www.geneious.com; Biomatters, Inc.).

PCR amplifications

Two primer pairs based on the GALV consensus sequences generated from the hybridization capture data were designed to fill in gaps found in the bioinformatics assembly. Primers GagF1 (5'-TGAGTAGCGAGCAGACGTGTT-3') and GagR1 (5'-GGCAAAATCACAGTGGAGTCA-3') were used to amplify a region encompassing part of the *gag* gene and the interspace fragment between 5' LTR and *gag*, while primers EnvF1 (5'-CAGTTGACCATTCGCTTGGA-3') and EnvR1 (5'-CCGAGGGTGAGCAACAGAA-3') were used to amplify part of the *env* gene. The PCR reaction mix comprised 12.5 µl of 2× MyFi Mix (Bioline), 0.6 µM final concentration of forward primer, 0.6 µM final concentration of reverse primer, approximately 100 ng of DNA template and sterile distilled water to a final volume of 22 µl. Thermal cycling conditions were: 95°C for 3 min; 40 cycles at 95°C for 15 s, 59°C for 20 s, 72°C for 30 s; and 72°C for 30 s. For EnvF1-EnvR1, the annealing temperature was set to 61°C instead of 59°C, and the extension time to 40 s instead of 30 s.

Five primer sets were designed based on the alignment of the phosphate transporter 1 (*PiT1* or *SLC20A1*) and the phosphate transporter 2 (*PiT2* or *SLC20A2*) sequences available in GenBank of *Mus musculus*, *Rattus norvegicus*, *Cricetulus griseus*, *Homo sapiens*, *Macaca mulatta*, *Nomascus leucogenys* to sequence the region A of *PiT1* and *PiT2* from *Hylobates lar*, *Melomys* sp., *Melomys paveli* and *Mus caroli*. Primers PiT1-F1long (5'-AGATCCTTACAGCCTGCTTTGG-3') and PiT1-R1 (5'-TCCTTCCCCATRGTCTGGAT-3') were designed to amplify a region approximately 600-bp long and encompassing the exons 7 and 8 of *PiT1* – which contains region A – compared to *M. musculus* sequence (800-bp long and targeting exons 8 and 9 compared to *H. sapiens* sequence). Primers PiT1-F1short (5'-CCTCTGGTTGCTTTGTATCTTGTT-3') for the rodent templates and PiT1-F1short_apes for the gibbon template (5'-GGCCTCTGGTTGCTTTATATTTG-3'), both in combination with the above mentioned PiT1-R1, were designed to amplify a 150-bp long fragment including region

A. Two primer pairs – PiT2-F1 (5'-TGCTATTGGTCCCCTTGTGG-3') and PiT2-R1 (5'-CCCCAAACCCAGAGACCTGT-3') for the rodents, and PiT2-F1_apes (5'-CCTGGTAGCCTTGTGGCTGA-3') and PiT2-R1_apes (5'-TGATGGGAGTGAGGTCCTTC-3') for the gibbon – were designed to amplify a fragment approximately 150-bp long including PiT2 region A. The PCRs were performed using approximately 100 ng of DNA extract, 0.6 µM of final concentration of each primer, 12.5 µl of 2× MyFi Mix (Bioline) and sterile distilled water to a final volume of 22 µl. Cycling conditions were: 95°C for 3 min; 35 cycles at 95°C for 15 s, 57°C for 20 s, 72°C for 10 s; and 72°C for 10 s. For PiT1-F1long and PiT1-R1, the extension at 72°C was prolonged to 30 s.

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

Evolutionary analyses

To characterize the phylogenetic relationships among the identified viral consensus sequences, the known GALV strains, MbRV and other related gammaretroviruses, phylogenetic trees were inferred based on the viral nucleotide sequences. The following reference sequences were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/GenBank): GALV-SEATO (KT724048), GALV-SF (KT724047), GALV-Brain (KT724049), GALV-Hall's Island (KT724050), woolly monkey virus (WMV; KT724051), *Melomys burtoni* retrovirus (MbRV; KF572483-6). KoRV (AF151794) was used as an outgroup. Genomic sequences and individual gene (*env*, *gag*, and *pol*) sequences were aligned using MAFFT (39). Phylogenetic analysis was performed using the maximum-likelihood (ML) method available in RAxML v8 (40), including 500 bootstrap replicates to determine the node support. The general time-reversible substitution model (41) with among-site rate heterogeneity modeled by the Γ distribution and four rate categories (42) were used. Nucleotide sequences of *env*, *gag*, and *pol* were concatenated and analyzed in a partitioned framework, where each partition was allowed to evolve under its own substitution model. In order to infer the phylogenetic trees, the nucleotide sequences of *env*, *gag*, and *pol* were both analyzed separately and concatenated including noncoding LTRs and spacers and analyzed in a partitioned framework.

Data accession

The complete sequence and annotations of MelWMV genome was deposited in GenBank under accession number KX059700. Illumina reads mapping to WMV for the six *Melomys burtoni* subsp. samples were deposited in the NCBI Sequence Read Archive as BioProject PRJNA318360.

RESULTS

Screening for GALV and KoRV in rodents using hybridization capture

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

Melomys woolly monkey virus (MelWMV)

Seven *Melomys* spp. samples were screened, of which six were from a new subspecies of *Melomys burtoni* from Wallacea which is in the process of being described (Fabre et al. unpublished data) (here referred to as *Melomys burtoni* subsp.). In addition, a sample of *Melomys paveli* from Seram Island (Moluccas, Indonesia) was included. Only *Melomys burtoni* subsp. yielded GALV-like sequences, with reads mapping to the woolly monkey virus (WMV) detected in all six *Melomys burtoni* subsp. samples. For most of the samples only few reads were found: from a minimum of 24 to a maximum of 1,008 mapping reads, but in each case distributed evenly across the WMV genome. However, in sample WD279 almost full coverage of the viral genome was obtained with an average per-base coverage of 18×. The enrichment (proportion of on-target reads mapping to WMV) was low

(below 1%) in all samples, similarly to our previous experiments (2). The negative control generated few sequence reads, none mapping to GALV.

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

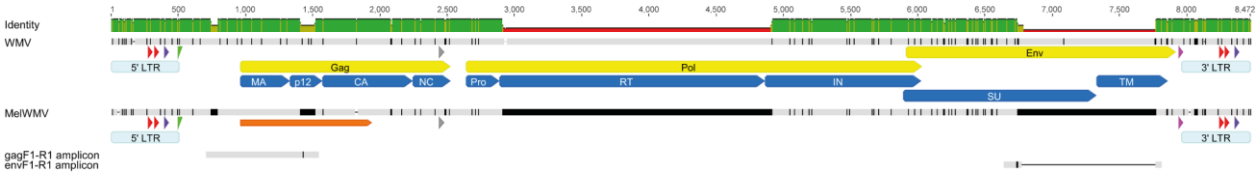


Figure 1. MelWMV genomic assembly and structure. Alignment of WMV and MelWMV consensus sequence generated from hybridization capture data combined with the PCR products that were produced to fill in the gaps in the bioinformatics assembly, shown as continuous black bars. Nucleotide positions identical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

The primers were applied to the *Melomys paveli* sample as well and confirmed the absence of GALV-like sequences suggested by the hybridization capture experiment. Identical amplification products from each primer set were produced for all 6 *Melomys burtoni*

subsp. samples. Based on the Sanger sequences and the hybridization capture Illumina reads, we determined that the viral sequences were identical in the 6 *Melomys burtoni* subsp. samples. The identified virus was characterized by the common genetic structure of simple type C mammalian retroviruses with a 5' LTR-*gag*-*pol*-*env*-3' LTR organization (Fig. 1). The 5' and 3' LTRs were identical. Nevertheless, the virus lacked approximately 60% of *pol*, with the whole reverse transcriptase domain missing, and almost half of the surface unit gp70 (SU) and most of the transmembrane subunit p15E (TM) of *env* (Fig. 1). The remaining protein domains of Pol – the protease (PR) and integrase (IN) – and all Gag protein domains – the matrix p15 (MA), p12, capsid p30 (CA), and nucleocapsid p10 (NC) – were intact. However, the ORF of *gag* was truncated by a premature stop codon. Therefore, the Gag protein was 324 amino acids long, instead of the 521 residues expected for WMV. The same regulatory motifs found in WMV and in the other GALVs (2) were identified: a tRNAPro primer binding site, a CAAT box, a TATA box, a Cys-His box, a polypurine tract, and a polyadenylation signal (Fig. 1). Furthermore, no differences between MelWMV and WMV were observed in the domains known to affect GALV and KoRV differential infectivity: the CETTG motif (43) of the Env protein (residues 167 to 171) and the PRPPIY and PPPY motifs (43, 44) of the Gag protein (residues 123-128 and 140-143). In addition, MelWMV showed high levels of conservation compared to WMV in the variable regions A and B (VRA and VRB) of the Env protein (residues 86-153 and 192-203, respectively), which are known to influence receptor specificity (45): only 6 out of 80 residues were polymorphic between the two viruses.

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

The detection in all the *Melomys burtoni* subsp. individuals tested and the identification of identical integration sites in each sample suggest that the virus is endogenous. To estimate a maximum age of endogenization, we used a molecular clock relying on the divergence between the 5' and 3' LTR sequences within the same provirus, as described in Ishida et al. (46). No differences were observed in the 1012 bp of 5' and 3' LTRs (each 506-bp long). Using mouse mutation rate of approximately 4.5×10^{-9} mutations per site per year (47-49) to estimate nuclear mutation rate of *Melomys burtoni*, we calculated that the first mutation anywhere within the LTRs would be expected to occur within 219,600 years of integration. Since no mutations were detected in LTRs, this would represent a maximum age estimate for the integration of the virus.

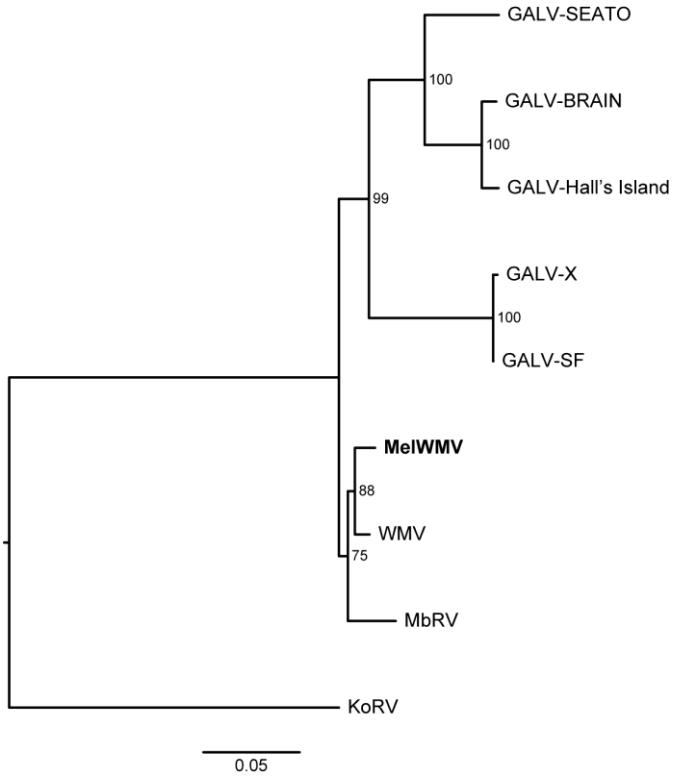


Figure 2. GALVs maximum likelihood phylogenetic tree inferred using concatenated partitioned full genome nucleotide sequences. Coding sequences, non-coding LTRs and inter-gene spacers were included in the analysis. The sequences obtained from GenBank with corresponding accession codes are: GALV-SEATO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

each node. The scale bar indicates 0.05 nucleotide substitutions per site. The tree is midpoint-rooted for purposes of clarity.

The newly identified virus shared 98% nucleotide identity with WMV and 96.7% with the *Melomys burtoni* retrovirus (MbRV). A phylogenetic analysis was performed including sequences from the genomes of the GALV strains and MbRV, using KoRV as an outgroup. The evolutionary relationships among these viruses were robust regardless of the data type analyzed, full genome (Fig. 2) or individual gene (*gag*, *pol* and *env*) nucleotide sequences (data not shown). The new virus formed a sister taxon to WMV, which together formed a monophyletic group with MbRV (Fig. 2). These three viruses in turn constituted a sister clade to the other GALV strains. The evolutionary relationship between the new virus and WMV was well supported (bootstrap 88 – 91%) using both concatenated partitioned nucleotide sequences (Fig. 2) and *gag* and *env* nucleotide sequences (data not shown). Therefore the new virus can be considered a strain of GALV and is here designated *Melomys* woolly monkey virus (MelWMV). Lower support was found using *pol* nucleotide sequences (bootstrap 51%), likely due to the large deletion of the gene in MelWMV, which reduced the number of phylogenetically informative sites (data not shown). The support for the relationship among the WMV-MelWMV clade and MbRV was not very robust (bootstrap 61 – 75%) since only partial sequences of *pol* and *env* were recovered for MbRV (Fig. 2; data not shown for *pol* and *env* trees).

Sequencing of region A of PiT1 and PiT2

Residues present in the C-terminal region of the fourth extracellular domain of PiT1, the receptor used by GALV to infect host cells (50), have been identified as critical for receptor function and therefore GALV infection (51-54). This nine-residue region, designated region A, has been extensively analyzed by mutational analysis and by comparative alignment of PiT1 orthologs that function as GALV receptor to PiT1 orthologs that fail to support GALV entry. Substitution of region A residues of PiT1 for the corresponding residues of two proteins that do not support GALV entry, Pit2 (a PiT1 paralog) (54) and the distantly related phosphate transporter Pho-4 from the filamentous fungus *Neurospora crassa* (53), renders these proteins functional as GALV receptors. Five primer sets were designed to sequence region A of *PiT1* and *PiT2* from *Hylobates lar*, *Melomys burtoni* subsp., *Melomys paveli* and *Mus caroli*. PiT2 was also sequenced since it is used by GALV to infect Chinese hamster and Japanese feral mouse cells (52, 55). An amplification product was obtained from

each of the five primer sets. Sanger sequencing of the amplicons and the subsequent BLAST search confirmed the amplification of the region A of *PiT1* and *PiT2*. The sequences were then aligned with the reference sequences of *Mus musculus*, *Rattus norvegicus*, *Cricetulus griseus*, *Homo sapiens*, *Macaca mulatta* and *Nomascus leucogenys* available in GenBank and translated into amino acid sequences. The amino acid sequences were then aligned and compared with the amino acid sequences of region A of PiT1 and PiT2 of all the species known to be permissive (*Homo sapiens*, *Rattus norvegicus*, *Mus musculus molossinus*, *Cricetulus griseus*) or resistant (*Mus musculus musculus* and *Mus dunni*) to GALV infection according to the literature (table 2) (50, 52, 55-57).

Residues at positions 550-558 and 522-530 comprise region A of PiT1 and PiT2 respectively. Positions 550 and 553 of PiT1, and 522 and 529 of PiT2 are crucial for receptor function (52-54). Functional GALV receptors have an acidic residue, either Asp(D) or Glu(E), at one or both of these positions. However, a Lys(K) at position 550 (522 in PiT2) is known to abrogate receptor function (52, 58). The PiT1 sequence of *M. caroli* had an Asp(D) at position 553 but also a Lys(K) at position 550, and overall it was identical to that of *M. dunni*, the cells of which are resistant to GALV infection (57). The sequence of PiT2 was identical to that of *Mus musculus molossinus* which serves as a functional GALV receptor (57): they both have a Gln(Q) at position 522, but a Glu(E) at position 529. The sequence of *H. lar* PiT1 region A had an Asp(D) at both positions 550 and 553, and was identical to the human sequence (50), whereas PiT2 displayed one amino acid difference – Thr(T) to Met(M) at position 527 – when compared to human (56). Both human cells and gibbons are permissive to GALV infection, but human PiT2, which has a Lys(K) at positions 522, like gibbon PiT2, does not function as a GALV receptor. The sequence of PiT1 region A of *Melomys burtoni* subsp. was very similar to the sequence carried by susceptible species such as rats, humans, gibbons and *Mus musculus molossinus*. *Melomys burtoni* subsp. had a Glu(E) at position 550 and an Asp(D) at position 553, identical to rat. The Thr(T), Val(V) and Lys(K) at positions 551, 554 and 557 respectively were invariant among *Melomys burtoni* subsp. and the other permissive species, with the Lys(K)-557 shared with both resistant and permissive species. The residues at positions 555, 556 and 558 of PiT1 varied randomly among resistant and susceptible species, while residue 552 was missing in the resistant ones. The PiT2 sequence of *Melomys burtoni* subsp. had a Glu(E) at position 522 and differed in only one residue – Met(M) to Thr(T) at position 527 – compared to *C. griseus* (59), which is also susceptible to GALV infection. The sequence was identical to *Mus musculus molossinus* PiT2, which is also considered a functional GALV receptor (57). The PiT1 and PiT2 region A sequences of *Melomys paveli*

were almost identical to *Melomys burtoni* subsp., but the PiT1 sequence of *Melomys paveli* lacked the residue – a Gly(G) in *Melomys burtoni* subsp. – at position 552, like in the resistant species.

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

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

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

DISCUSSION

KoRV and GALV are closely related retroviruses (2). However, their respective hosts, koalas and gibbons, share neither a recent common ancestor nor overlapping geographic

distributions. Thus, KoRV and GALV may have arisen from a cross-species transmission that involved an intermediate host (9, 14-16). In order to identify such a vector, Simmons et al. (16) screened 42 Australian vertebrate species (birds and mammals including rodents and bats) for KoRV and GALV-like sequences. An ERV closely related to GALV (MbRV) was found in the Australian subspecies of the murid species *Melomys burtoni*, but, even if related to GALVs, particularly WMV, it does not represent an ancestor of GALV or KoRV because the distribution of *Melomys burtoni* and gibbons do not overlap (16). Because GALV-like viruses have been identified in Southeast Asian rodents (20, 21, 60), we screened rodent species from this geographic area in the attempt to identify potential intermediate hosts and retrieve ancestral viral strains of KoRV and GALV. Twenty-six rodent species were screened (table 1). Some of the species tested (*Bandicota savilei*, *Bandicota indica*, *Bandicota bengalensis*, *Berylmys berdmorei*, *Mus musculus*) had been reported as negative for GALV and KoRV by Simmons et al. (16), consistent with the absence of GALV and KoRV from the Southeast Asian samples from the same species in this study. None of the species tested in the current study or in Simmons et al. (16) was positive for KoRV-like sequences, while only two different subspecies of *Melomys burtoni*, one from Australia from Simmons et al. (16) and one from Wallacea from the current study, were found positive for GALV-like sequences. Based on the homology to WMV (98%) and phylogenetic affinity, the *Melomys* woolly monkey virus (MelWMV) that we discovered in the Wallacean *Melomys burtoni* subsp. is a subtype of WMV, whereas MbRV from the Australian subspecies is a sister taxon (Fig. 2).

Only one integration site was found for MelWMV. Therefore there may be only a single copy of MelWMV in the genome of *Melomys burtoni* subsp., and this would explain the low hybridization capture coverage. Furthermore, MelWMV was detected in all 6 individuals of *Melomys burtoni* subsp. tested and the integration site was identical in all 4 individuals for which they were identified by hybridization capture. This result, the premature stop codon in *gag* and the deletions in *pol* and *env* (Fig. 1) strongly indicate that MelWMV is an endogenous retrovirus. Furthermore, we estimate that MelWMV has recently (within the last 200,000 years) integrated into the genome of the Wallacean *Melomys burtoni* subsp., based on the identical 5' and 3' LTR sequences and the mutation rate of a murid host (46). MelWMV is not present in *M. paveli*, tested in this study, and in the Australian *M. burtoni* subspecies and the endemic Australian *M. cervinipes*, tested in Simmons et al. (16). The different species of *Melomys* diverged from a common ancestor between one and two million years ago (61), consistent with the date of integration of MelWMV into the *Melomys burtoni* genome based on the LTR sequences.

MelWMV along with WMV and MbRV represent the most basal clade of the GALV phylogeny described to date, so it can be argued that the WMV-like viruses are the most ancestral GALV strains currently known to be circulating and most likely the closest viruses to the progenitor of GALV and KoRV. Such close GALV relatives were only found in two different populations of the murine species *Melomys burtoni* out of 68 total species tested in Australia (16) and SE Asia. Furthermore, more distantly related GALV-like ERVs are found in rodents belonging to the genus *Mus* (20, 60). Taken together, this suggests an overall rodent origin of the clade, more specifically an Australo-Papuan murine origin. However, since MelWMV is an ERV in *Melomys burtoni* subsp. but *M. paveli* did not yield any GALV-like sequences, it is not clear whether *Melomys* is a reservoir or a susceptible host for GALVs. Thus, it is formally possible that GALV did not originate in *Melomys* and the two *Melomys burtoni* subspecies were independently infected with GALV in Wallacea and Australia from an unknown reservoir species. As the vast majority of samples in the current study were from Southeast Asia and those of Simmons et al. (16) exclusively from Australia, Wallacea and Papua New Guinea remain largely unexplored. In addition, only three species of *Melomys* have been tested out of a total of 22 *Melomys* species, 19 of which are found in the Moluccas, Melanesia and Papua New Guinea (IUCN 2015. *The IUCN Red List of Threatened Species. Version 2015-4*. http://www.iucnredlist.org), suggesting that many more GALVs, including potentially exogenous GALVs, and possibly KoRV-like sequences may be present. Of particular relevance to the current host range of GALV, *Melomys* species are found in both Australia and Wallacea. Since Wallacea is a transitional zone between Asia and Australia, the discovery of MelWMV in a Wallacean subspecies of *M. burtoni* represents the most proximate record of GALV to the Asian continent and to the distribution of gibbons. However, even if the genus *Melomys* is one of the most widespread murine genera in the Australo-Papuan region, and specifically one of those which has dispersed furthest to the West (to the Moluccas), it has never been reported, not even from the fossil record, in Sulawesi or the Sunda Shelf (mainland Southeast Asia) (62), and thus it has probably never been in direct contact with gibbons. Therefore, it is still not clear how the virus moved from Australia and Wallacea to mainland Southeast Asia crossing the Wallace Line, a line running between the islands of Bali and Lombok and dividing the Australian and the Asian biogeographic zones. An intermediate and mobile host which is distributed across the Wallace Line must have played a critical role in the viral transmission. However, our study suggests that any intermediate host which eventually infected gibbons in Southeast Asia came in contact with *M. burtoni* in Wallacea. *Rattus* species would be good candidates as GALV and KoRV hosts given their

widespread distribution in this region (Australia, Papua New Guinea, both insular and mainland Southeast Asia). However, nine *Rattus* species, including both species endemic to Australia and with an Australo-Papuan distribution, were tested and reported as negative for GALV and KoRV by Simmons et al. (16). Similarly, in a preliminary screening of *Rattus exulans* and *R. rattus* from Southeast Asia using a single GALV and KoRV as hybridization capture bait, we did not identify any GALV sequence (data not shown). Other candidate hosts are lineages belonging to the same molecular tribe of *Melomys*, such as *Hydromys* and *Uromys* genera, which display a similarly wide Australo-Papuan distribution, but have been not included in this study and not extensively sampled in Simmons et al. (16). Gibbons in particular are surprising hosts as GALVs have only been isolated from captive and not wild gibbons suggesting they have had infrequent but regular contact with a GALV reservoir or host species but only in captive facilities. This is particularly relevant for the gibbon colony housed at the SEATO Laboratory in Bangkok, Thailand (12), from which the other non-Asian gibbon colonies originated.

GALV infects cells using a ubiquitous transmembrane protein that functions as a sodium-dependent phosphate transporter called PiT1 or SLC20A1 (50). GALV can alternatively infect cells using a related phosphate transporter, PiT2 or SLC20A2, originally recognized as the amphotropic murine leukemia virus (A-MuLV) and 10A1 MuLV receptor, to infect Chinese hamster and Japanese feral mouse cells (52, 55, 56). This similarity of receptor usage is consistent with the phylogenetic relationship of GALVs and MuLVs, which belong to the same overall retroviral group (2).

Mutagenesis studies have shown that region A of PiT1, a stretch of nine residues corresponding to residues 550-558 of human PiT1, which is highly polymorphic among species, is crucial for GALV entry into cells (51, 52). Because of its highly polymorphic nature, it is not clear which of the residues of region A are essential for GALV infection. Schneiderman et al. (52) had suggested that the functional GALV receptors have an acidic residue at either position 550 or 553 of PiT1 (522 or 529 of PiT2) or both, but lysine at position 550 (522 in PiT2) abrogates GALV receptor function, even when an acidic residue is present at position 553 (529 in PiT2). A subsequent study (58) demonstrated that PiT1 and PiT2 serve as receptors for GALV when lysine is absent from the first position, regardless of the presence of acidic residues at the above mentioned positions. We have sequenced PiT1 and PiT2 region A from species resulted both positive (*Melomys burtoni* subsp.) and negative (*Melomys paveli* and *Mus caroli*) to our GALV screening, and also from *Hylobates lar*, another natural host of GALV. When comparing with the previously reported sequences of species

both permissive (human *Homo sapiens*, rat *Rattus norvegicus*, Japanese feral mouse *Mus musculus molossinus*, Chinese hamster *Cricetulus griseus*) and resistant (*Mus musculus*, *Mus dunni*) to GALV infection (table 2), the sequences generated here were consistent with the findings of previous functional studies (51, 52, 58). Positions 551-2 and 554-8 of PiT1 are not critical determinants of receptor function. All permissive species have a Thr(T) and a Val(V) at positions 551 and 554, whereas resistant species have a Gln(Q) and Ala(A) respectively. However, these positions in PiT1 may not be crucial as PiT2 of both resistant and permissive species have a Gln(Q) and a Val(V) at positions 523 and 526 respectively, which correspond to residues 551 and 554 of PiT1. Positions 555, 556 and 558 of PiT1, which varied randomly among resistant and susceptible species, and the Lys(K) at position 557, which was present in all species, are unlikely to be determinants of GALV susceptibility.

In contrast, positions 550 and 553 of PiT1 may play a key role, as previously proposed by Schneiderman et al. (52). All permissive species have an acidic residue – Asp(D) or Glu(E) – at either position 550 or 553 of PiT1. In PiT2 an acidic residue is found at either position 522 or 529 among permissive species. A Lys(K) is present at the first position, 550 of PiT1 or 522 of PiT2, in all resistant species and therefore it is likely to be the residue which determines the resistance to GALV infection. Therefore, the *Mus caroli* PiT1 sequenced in this study, which has a Lys(K) at position 550 and is identical to *Mus dunni* in region A, is unlikely to serve as a GALV receptor. This is consistent with the absence of any GALV-like sequence in this species. McERV sequences were detected but this virus uses a different receptor than GALV (23). However, GALV could potentially infect *Mus caroli* using PiT2, since *Mus caroli* PiT2 sequence is identical to that of *Mus musculus molossinus* PiT2 that is a functional GALV receptor. Regions A of human and gibbon PiT1 are identical, and both humans and gibbons have a Lys(K) at the first position of PiT2 region A. Human PiT1 functions as GALV receptor, while PiT2 does not. Given the similarity between human and gibbon PiT receptors captive gibbons were likely infected via PiT1.

Both PiT1 and PiT2 of *Melomys burtoni* subsp. are potentially functional GALV receptors, consistent with our discovery of MelWMV in this species. However, MelWMV and WMV are highly similar in the VRA and VRB domains of the envelope, and WMV is known to be unable to use the PiT2 receptor to infect hamster cells due to a block mediated by WMV envelope, specifically VRA and VRB (45). Therefore, it is likely that *Melomys burtoni* subsp. was infected by WMV via the PiT1 receptor. *Melomys paveli* is also potentially susceptible to GALV infection, since its PiT1 and PiT2 region A are identical to *Melomys burtoni* subsp., with the exception that residue 552 is missing in PiT1, as observed in resistant species (*Mus*

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

In conclusion, our screen of Southeast Asian rodents identified MelWMV in a *Melomys burtoni* subspecies from Wallacea. MelWMV represents the most closely related retrovirus to GALV identified from rodents to date and the second GALV relative identified from two different subspecies of *Melomys burtoni*, suggesting that either *Melomys burtoni* is a host of GALVs or more species within the genus *Melomys* are sympatric with the reservoir. With the current data, we cannot distinguish between the two possibilities that MelWMV derives from MbRV and represents a single infection of *M. burtoni* with subsequent evolution or whether the two viruses represent independent infections. However, WMV itself must represent a distinct infection event because *Melomys* do not overlap with gibbons geographically. The PiT1 and PiT2 region A sequences of the *Melomys* species tested in the current study are consistent with the general susceptibility of these species to GALV infection. Further screening of GALV and KoRV in *Melomys* across the range of this genus, in older *Melomys* related lineages of the Murinae subfamily, and in potential host species that have crossed the Wallace Line would be promising for identifying additional GALV sequences.

ACKNOWLEDGEMENTS

The authors wish to thank Karin Hönig (Leibniz Institute for Zoo and Wildlife Research) and Susan Mbedi of the Berlin Centre for Genomics in Biodiversity Research (BeGenDiv) for sequencing support. We thank the State Ministry of Research and Technology (RISTEK, permit number: 028/SIP/FRP/SMII/2012) and the Ministry of Forestry of the Republic of Indonesia for providing permits to carry out fieldwork in the Moluccas. Likewise, we thank the Research Center for Biology, Indonesian Institute of Sciences (RCB-LIPI) and the Museum Zoologicum Bogoriense for providing staff and support to carry out fieldwork in the Moluccas. P.-H.F. want to acknowledges Carsten Rahbek (Center for Macroecology Evolution and Climate, University of Copenhagen, Denmark) who generously funds part of his recent field research in the Moluccas.

FUNDING INFORMATION

Y.I., A.L.R., K.M.H. and A.D.G. were supported by Grant Number R01GM092706 from the National Institute of General Medical Sciences (NIGMS). M.V.E. contribution was supported by National Institute of Mental Health Intramural Research Program Project ZIAMH002592. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the National Institutes of Health. N.A. was supported by the International Max Planck Research School for Infectious Diseases and Immunology (IMPRS-IDI) at the Interdisciplinary Center of Infection Biology and Immunity (ZIBI) of the Humboldt University Berlin (HU). P.-H.F. was funded by a Marie-Curie fellowship (PIOF-GA-2012-330582-CANARIP-RAT). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. **Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE.** 2010. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science **329:**676-679.
2. **Alfano N, Kolokotronis SO, Tsangaras K, Roca AL, Xu W, Eiden MV, Greenwood AD.** 2015. Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses. J Virol **90:**1757-1772.
3. **Kawakami TG, Huff SD, Buckley PM, Dungworth DL, Synder SP, Gilden RV.** 1972. C-type virus associated with gibbon lymphosarcoma. Nat New Biol **235:**170-171.
4. **DePaoli A, Johnsen DO, Noll MD.** 1973. Granulocytic leukemia in white handed gibbons. J Am Vet Med Assoc **163:**624-628.
5. **Reitz MS, Jr., wong-Staal F, Haseltine WA, Kleid DG, Trainor CD, Gallagher RE, Gallo RC.** 1979. Gibbon ape leukemia virus-Hall's Island: new strain of gibbon ape leukemia virus. J Virol **29:**395-400.
6. **Todaro GJ, Lieber MM, Benveniste RE, Sherr CJ.** 1975. Infectious primate type C viruses: Three isolates belonging to a new subgroup from the brains of normal gibbons. Virology **67:**335-343.
7. **Theilen GH, Gould D, Fowler M, Dungworth DL.** 1971. C-type virus in tumor tissue of a woolly monkey (*Lagothrix* spp.) with fibrosarcoma. J Natl Cancer Inst **47:**881-889.
8. **Wolfe LG, Smith RK, Deinhardt F.** 1972. Simian sarcoma virus, type 1 (Lagothrix): focus assay and demonstration of nontransforming associated virus. J Natl Cancer Inst **48:**1905-1908.
9. **Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF.** 2000. The nucleotide sequence of koala (*Phascolarctos cinereus*) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus. J Virol **74:**4264-4272.
10. **Shojima T, Yoshikawa R, Hoshino S, Shimode S, Nakagawa S, Ohata T, Nakaoka R, Miyazawa T.** 2013. Identification of a novel subgroup of Koala retrovirus from Koalas in Japanese zoos. J Virol **87:**9943-9948.

11. **Xu W, Stadler CK, Gorman K, Jensen N, Kim D, Zheng H, Tang S, Switzer WM, Pye GW, Eiden MV.** 2013. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proc Natl Acad Sci U S A **110:**11547-11552.
12. **Kawakami TG, Kollias GV, Jr., Holmberg C.** 1980. Oncogenicity of gibbon type-C myelogenous leukemia virus. Int J Cancer **25:**641-646.
13. **Tarlinton R, Meers J, Hanger J, Young P.** 2005. Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. J Gen Virol **86:**783-787.
14. **Tarlinton R, Meers J, Young P.** 2008. Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci **65:**3413-3421.
15. **Fiebig U, Hartmann MG, Bannert N, Kurth R, Denner J.** 2006. Transspecies transmission of the endogenous koala retrovirus. J Virol **80:**5651-5654.
16. **Simmons G, Clarke D, McKee J, Young P, Meers J.** 2014. Discovery of a novel retrovirus sequence in an Australian native rodent (*Melomys burtoni*): a putative link between gibbon ape leukemia virus and koala retrovirus. PLoS One **9:**e106954.
17. **Wong S, Lau S, Woo P, Yuen KY.** 2007. Bats as a continuing source of emerging infections in humans. Rev Med Virol **17:**67-91.
18. **Cui J, Tachedjian G, Tachedjian M, Holmes EC, Zhang S, Wang LF.** 2012. Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats. J Gen Virol **93:**2037-2045.
19. **Martin J, Herniou E, Cook J, O'Neill RW, Tristem M.** 1999. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. J Virol **73:**2442-2449.
20. **Lieber MM, Sherr CJ, Todaro GJ, Benveniste RE, Callahan R, Coon HG.** 1975. Isolation from the asian mouse Mus caroli of an endogenous type C virus related to infectious primate type C viruses. Proc Natl Acad Sci U S A **72:**2315-2319.
21. **Callahan R, Meade C, Todaro GJ.** 1979. Isolation of an endogenous type C virus related to the infectious primate type C viruses from the Asian rodent Vandeleuria oleracea. J Virol **30:**124-131.
22. **Benveniste RE, Callahan R, Sherr CJ, Chapman V, Todaro GJ.** 1977. Two distinct endogenous type C viruses isolated from the asian rodent Mus cervicolor: conservation of virogene sequences in related rodent species. J Virol **21:**849-862.
23. **Miller AD, Bergholz U, Ziegler M, Stocking C.** 2008. Identification of the myelin protein plasmolipin as the cell entry receptor for Mus caroli endogenous retrovirus. J Virol **82:**6862-6868.
24. **Wolgamot G, Bonham L, Miller AD.** 1998. Sequence analysis of Mus dunni endogenous virus reveals a hybrid VL30/gibbon ape leukemia virus-like structure and a distinct envelope. J Virol **72:**7459-7466.
25. **Bromham L, Clark F, McKee JJ.** 2001. Discovery of a novel murine type C retrovirus by data mining. J Virol **75:**3053-3057.
26. **Maricic T, Whitten M, Paabo S.** 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One **5:**e14004.
27. **Tsangaras K, Siracusa MC, Nikolaidis N, Ishida Y, Cui P, Vielgrader H, Helgen KM, Roca AL, Greenwood AD.** 2014. Hybridization capture reveals evolution and conservation across the entire Koala retrovirus genome. PLoS One **9:**e95633.
28. **Sikes RS, Gannon WL, Animal Care and Use Committee of the American Society of Mammalogists.** 2011. Guidelines of the American Society of Mammalogists for the use of wild mammals in research. Journal of Mammalogy **92:**235–253.
29. **Sambrook J, Russell DW.** 2006. Purification of nucleic acids by extraction with phenol:chloroform. CSH Protoc **2006**.

30. **Meyer M, Kircher M.** 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc **2010:**pdb.prot5448.
31. **Alfano N, Courtiol A, Vielgrader H, Timms P, Roca AL, Greenwood AD.** 2015. Variation in koala microbiomes within and between individuals: effect of body region and captivity status. Sci Rep **5:**10189.
32. **Kircher M, Sawyer S, Meyer M.** 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res **40:**e3.
33. **Martin M.** 2012. Cutadapt removes adapter sequences from high-throughput sequencing reads. . Bioinformatics in Action **17:**10-12.
34. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:**2114-2120.
35. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-3402.
36. **Ondov BD, Bergman NH, Phillippy AM.** 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics **12:**385.
37. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv **1303**.
38. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-2079.
39. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30:**772-780.
40. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30:**1312-1313.
41. **Lanave C, Preparata G, Saccone C, Serio G.** 1984. A new method for calculating evolutionary substitution rates. J Mol Evol **20:**86-93.
42. **Yang Z.** 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39:**306-314.
43. **Oliveira NM, Satija H, Kouwenhoven IA, Eiden MV.** 2007. Changes in viral protein function that accompany retroviral endogenization. Proc Natl Acad Sci U S A **104:**17506-17511.
44. **Shojima T, Hoshino S, Abe M, Yasuda J, Shogen H, Kobayashi T, Miyazawa T.** 2013. Construction and characterization of an infectious molecular clone of Koala retrovirus. J Virol **87:**5081-5088.
45. **Ting YT, Wilson CA, Farrell KB, Chaudry GJ, Eiden MV.** 1998. Simian sarcoma-associated virus fails to infect Chinese hamster cells despite the presence of functional gibbon ape leukemia virus receptors. J Virol **72:**9453-9458.
46. **Ishida Y, Zhao K, Greenwood AD, Roca AL.** 2015. Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. Mol Biol Evol **32:**109-120.
47. **Kumar S, Subramanian S.** 2002. Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A **99:**803-808.
48. **Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al.** 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420:**520-562.
49. **Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y, et al.** 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol **19:**1453-1457.

50. **O'Hara B, Johann SV, Klinger HP, Blair DG, Rubinson H, Dunn KJ, Sass P, Vitek SM, Robins T.** 1990. Characterization of a human gene conferring sensitivity to infection by gibbon ape leukemia virus. Cell Growth Differ **1:**119-127.
51. **Johann SV, van Zeijl M, Cekleniak J, O'Hara B.** 1993. Definition of a domain of GLVR1 which is necessary for infection by gibbon ape leukemia virus and which is highly polymorphic between species. J Virol **67:**6733-6736.
52. **Schneiderman RD, Farrell KB, Wilson CA, Eiden MV.** 1996. The Japanese feral mouse Pit1 and Pit2 homologs lack an acidic residue at position 550 but still function as gibbon ape leukemia virus receptors: implications for virus binding motif. J Virol **70:**6982-6986.
53. **Pedersen L, Johann SV, van Zeijl M, Pedersen FS, O'Hara B.** 1995. Chimeras of receptors for gibbon ape leukemia virus/feline leukemia virus B and amphotropic murine leukemia virus reveal different modes of receptor recognition by retrovirus. J Virol **69:**2401-2405.
54. **Pedersen L, van Zeijl M, Johann SV, O'Hara B.** 1997. Fungal phosphate transporter serves as a receptor backbone for gibbon ape leukemia virus. J Virol **71:**7619-7622.
55. **Wilson CA, Farrell KB, Eiden MV.** 1994. Properties of a unique form of the murine amphotropic leukemia virus receptor expressed on hamster cells. J Virol **68:**7697-7703.
56. **van Zeijl M, Johann SV, Closs E, Cunningham J, Eddy R, Shows TB, O'Hara B.** 1994. A human amphotropic retrovirus receptor is a second member of the gibbon ape leukemia virus receptor family. Proc Natl Acad Sci U S A **91:**1168-1172.
57. **Wilson CA, Farrell KB, Eiden MV.** 1994. Comparison of cDNAs encoding the gibbon ape leukaemia virus receptor from susceptible and non-susceptible murine cells. J Gen Virol **75 (Pt 8):**1901-1908.
58. **Chaudry GJ, Eiden MV.** 1997. Mutational analysis of the proposed gibbon ape leukemia virus binding site in Pit1 suggests that other regions are important for infection. J Virol **71:**8078-8081.
59. **Eiden MV, Farrell KB, Wilson CA.** 1996. Substitution of a single amino acid residue is sufficient to allow the human amphotropic murine leukemia virus receptor to also function as a gibbon ape leukemia virus receptor. J Virol **70:**1080-1085.
60. **Callahan R, Benveniste RE, Sherr CJ, Schidlovsky G, Todaro GJ.** 1976. A new class of genetically transmitted retravirus isolated from Mus cervicolor. Proc Natl Acad Sci U S A **73:**3579-3583.
61. **Bryant LM, Donnellan SC, Hurwood DA, Fuller SJ.** 2011. Phylogenetic relationships and divergence date estimates among Australo-Papuan mosaic-tailed rats from the Uromys division (Rodentia: Muridae). Zoologica Scripta **40:**433-447.
62. **Breed WG, Aplin KP.** 2008. The 'Uromys Group': Papuan Old Endemics. *In* Van Dyck SM, Strahan R (ed.), The Mammals of Australia, 3rd edn (pp. 666). Reed New Holland, Sydney, Australia.

Chapter V

Concluding Remarks

Concluding Remarks

KoRV has attracted much scientific attention since its discovery in the late 1990s for several reasons (1). First, it is the only infectious retrovirus which is currently in the process of invading the germ line of its host species, the koala, and therefore provides the unique opportunity to study the process of retroviral endogenization as it happens (2). Second, KoRV is believed to induce leukemia, lymphomas and immunosuppression in koalas, which may eventually cause higher susceptibility to secondary infections, such as the highly prevalent *Chlamydia* infection in koalas (3, 4). The combined effects of KoRV and *Chlamydia* infection may lead to local extinctions of koalas (5). Third, KoRV is most closely related to GALV, a retrovirus which infects gibbons in Southeast Asia (1). KoRV and GALV are likely the results of cross-species transmissions which likely occurred via intermediate as yet unknown host(s) (1, 6, 7). This thesis had two primary aims: to evaluate the effect of KoRV on koala health through the study of the koala microbiome (chapter II) and to investigate the evolutionary history of KoRV and GALV trying to identify the intermediate host(s) involved in the cross-species transmission (chapters III and IV).

In **chapter II**, I estabilished the healthy baseline for koala ocular, oral, rectal and fecal microbiomes. Future comparisons with the microbiomes of KoRV negative and *Chlamydia* infected koalas will help to elucidate if KoRV has an effect on koala microbial communities and, more specifically, which changes occur in koala bacterial communities following infection with *Chlamydia*. Since *Chlamydia* frequently causes ocular infections and keratoconjunctivitis in koalas, which can progress to blindness (8-10), the future comparison with *Chlamydia* infected eye microbiomes is of special interest because it will show how the pathogen interacts with the resident bacteria of the koala eye. Furthermore, since a high proportion of the koala ocular community was found to be represented by a group of bacteria never described before in the eye (family Phyllobacteriaceae), further studies are warranted to clarify the role of these bacteria. The characterization of the koala microbiome in digestion-associated body regions, and the comparison with other mammalian species microbiomes, demonstrated that koalas, despite their highly specialized diet based almost exclusively on *Eucalyptus* leaves, do not show unique features in their bacterial community composition. However, since the comparison with other mammals microbiomes was based on the most abundant taxa (e.g. predominant phyla and genera), it is important to consider that koala adaptation to the *Eucalyptus* diet may be reflected by the presence of low-abundance bacterial species performing specialized functions, such as degradation of lignin and tannins, which are abundant in *Eucalyptus* leaves (11, 12). For example, the tannin degrading bacterium *Lonepinella*

koalarum has been detected at very low abundance (<0.2%) in koala feces in a previous study on the koala gut microbiome (13). Since abundant molecular functions are not necessarily provided by abundant taxa (14), a metagenomic functional analysis of koala microbiome could help identify abundant functions shared by several low-abundance taxa and reveal peculiar features of koala digestion-associated microbiomes. Furthermore, given the growing evidence that gut microbiome composition relates to health status of the host (15), the findings of this study can be relevant for the management and assessment of the health of koalas in zoos. The evidence that the fecal microbiomes of the captive koalas analysed in this study were similar to those reported for wild koalas suggests that captivity does not shift microbiome communities in koalas and may not compromise koala microbial health. This should reassure managers of koalas in captive facilities, because this differs from many other species, especially carnivores, for which captivity can pose serious dietary and behavior based health issues. Moreover, since feces were found to subsample the microbial diversity detected by rectal swabs, this study questions the common use of fecal samples to investigate gut microbiome in mammals. Future studies should compare fecal and rectal samples in describing gut microbiome composition of other species in order to understand if the recommendation for future gut microbiome investigations should be to use non-fecal samples in general.

It has to be acknowledged that this study was limited by only comparing two animals and caution should be used in establishing definitive conclusions based on these results. The small sample size is due, on one hand, to the fact that the population of captive koalas in Europe consists only of few individuals (around 30), and, on the other hand, to the difficulty of convincing zoos to collect invasive samples, such as rectal and conjunctival swabs. Similarly, obtaining wild koala samples is difficult as indicated by a similar study on wild koala microbiomes (13) that was also based on two individuals. The main goal of this study consisted in comparing multiple body regions in each koala, some of which have been rarely described in the literature (eye, rectum) and getting a wider range of sample types per koala rather than a single sample type from multiple individuals was the priority. However, further studies on more koala individuals are needed to further confirm the findings of this study.

Chapters III and IV describe the evolutionary history of KoRV and GALV. The study presented in **chapter III** was preliminary to the one in chapter IV, and consisted of generating the complete nucleotide sequence of all GALV strains, describing their genomic structure and analyzing the phylogenetic relationships within the GALVs and with the other gammaretroviruses, all information beneficial for the following study on KoRV and GALV evolutionary history. Hybridization capture, which was used in combination with high-throughput sequencing to recover the GALVs genome sequences, proved to be a

valuable tool for viral discovery. This technique was able to capture sequences with up to 12.9% nucleotide divergence from the baits used, and most likely can be applied to capture viruses with higher divergence, as suggested by cross-species hybridization studies (16, 17). With this study, for the first time the whole genomic diversity of the five strains of GALV isolated to date has been described. All strains were characterized by the typical genetic structure of simple type C mammalian retroviruses with a 5' LTR-*gag-pol-env*-3' LTR organization and showed an average nucleotide identity above 90%, well within the threshold of 80–90% nucleotide identity for retroviral isolates to be considered as the same "species" (18). The comparison among the GALV strains in the motifs regulating viral pathogenicity and receptor usage helped identify the differences between the GALVs in these important biological features. Generally, high levels of sequence conservation were observed in the domains of Gag and Env proteins influencing viral infectivity (19-22), even though few polymorphisms were detected, mainly concentrated in WMV and shared with KoRV. High variability was observed among the GALVs only in two of the motifs of the surface unit of Env regulating viral infectivity (19) and in the LTRs, which are known to contain transcriptional enhancers and could influence the leukemogenicity of the strains. Future functional analysis of these polymorphisms may reveal further insight into the differential pathogenicity of the GALV strains. This study also confirmed the importance of VRA and VRB envelope domains in influencing receptor specificity in gammaretroviruses (23, 24). High variability was detected in these motifs among the GALVs, with WMV being the most divergent strain. This is consistent with the fact that WMV is the only GALV which is unable to infect hamster E36 cells, similarly to KoRV-A, and that VRA/VRB of the WMV envelope have been shown to be responsible for the infection block (25). Further functional analyses targeted at those residue positions which differ in WMV from the other GALVs but are identical to KoRV-A will help elucidate which specific residues are involved in the infection block. This study also demonstrated that the substitution of both VRA and VRB of GALV with the corresponding residues of KoRV-B is required to switch GALV receptor usage from PiT1 to THTR1, further confirming the role of these domains as determinants of receptor usage. The phylogenetic analyses showed that the GALVs strains formed a monophyletic clade, which was sister group to the KoRVs clade. Within the GALV clade, WMV occupied the most basal position, suggesting that WMV may be the most ancestral GALV, as supported by the fact that WMV seems to share some ancestral traits with KoRV, such as host range (inability to infect hamster cells) and infectivity motifs in the *env* gene. Signatures of episodic diversifying selection, which are bursts of positive selection from an otherwise negative selection pressure pattern (26, 27), were detected on the GALV/KoRV clade in the *env* gene and, specifically, on eight amino acids within the surface unit of the envelope.

Positive selection is a hallmark of evolutionary struggle between a virus and its host, and is usually concentrated at the interacting surfaces between host and viral proteins (28). More pathogenic and infectious viruses are expected to face harsher immune and antiviral responses from the hosts. Consistent with this prediction, the clades found under selection included the Hall's Island and SEATO strains, which are among the GALVs with a stronger association with leukemias in captive gibbons and likely more pathogenic, and KoRV-B and KoRV-J, which are thought to be infectious exogenous variants of KoRV and which have switched receptor usage from PiT1 to THTR1 (29, 30), possibly to evade host infection blocks. Conversely, no evidence of selection was found on the endogenous KoRV variant (KoRV-A), which is expected to be more adapted to its host and to be confronted by less severe immune responses. Furthermore, the eight residues where signs of episodic diversifying selection were detected are located in the surface unit of the envelope, which is the portion of the virus binding to host receptors, and four of them in the VRA and VRB domains, which are major determinants for receptor specificity (25). Such evidence suggests that the conflicting interaction of these viruses with host immune systems ("arms race") has shaped the evolution of their genomes at the contact surfaces with the hosts.

Part of the information gathered in this study were used in the follow-up study presented in **chapter IV**, which was aimed at identifying possible intermediate hosts involved in the cross-species transmission between koalas and gibbons from which KoRV and GALV originated. Twenty-six species of Southeast Asian rodents were screened for the presence of KoRV- and GALV-like sequences using hybridization capture and high-throughput sequencing. The GALV strains genomic sequences from the previous study were used, on the one hand, to design primers to produce the baits needed in the hybridization capture experiment to enrich for GALV sequences, and, on the other hand, to provide a comparative framework for analyzing GALV-like retroviruses as they are discovered. Only the individuals belonging to a new subspecies of *Melomys burtoni* from Indonesia were positive yielding an endogenized provirus which was phylogenetically very closely related to WMV and was named MelWMV. The new Indonesian subspecies of *M. burtoni* is in the process of being described and its geographical distribution defined, but is known to be distributed in Wallacea, a group of Indonesian islands which separates the Asian and Australian continental shelves. The virus was also related to MbRV, another GALV-like virus discovered in the Australian population of *Melomys burtoni* in a wide screening of Australian wildlife for KoRV and GALV (31). Even though GALV-like retroviruses have been isolated in the genome of several Southeast Asian rodents (32-34), MelWMV and MbRV are the most closely related viruses, among those sequenced so far, to GALV and KoRV. Overall, this evidence support the hypothesis that GALV

originated in rodents and spread secondarily to gibbons and koalas. Furthermore, the discovery of two close GALV relatives (MelWMV and MbRV) in two populations of the Australo-Papuan species *M. burtoni* suggests that this species may have played an important role not only in the spread of GALV-like viruses in this region, but also in the cross-species transmission between koalas and gibbons. Indeed, MbRV was isolated in the Australian subspecies of *M. burtoni* which therefore overlaps with koala distribution, while MelWMV represents the most proximate record of GALV to the Asian continent and to the distribution of gibbons, since it was identified in a transitional zone between Asia and Australia (Wallacea). Even though this study expanded both the geographic and taxonomic distribution of GALV, *M. burtoni* is not present in mainland Southeast Asia, where gibbons are distributed, and therefore gibbons were infected by another yet unknown intermediate host. In particular, it is still not clear how the progenitor virus of KoRV and GALV crossed the Wallace Line, a deep sea trench separating the Australian and the Asian biogeographic zones. Future efforts in the quest of the intermediate host of KoRV and GALV should focus on species which are distributed on both sides of Wallace line. Even though GALV most likely has a rodent origin, it still possible that other vertebrates may have played a key role in the viral transfer. Bats fly and are vectors in several zoonotic diseases, and are therefore candidates. Furthermore, the Southeast Asian bat species *Megaderma lyra* was found to harbor a retrovirus related to GALV (35). Some bat species have been screened for KoRV and GALV by Simmons et al. (31) yielding negative results, but only 7 species and 2 genera of Australian bats were tested. Concerning rodents, further screening of lineages of the subfamily Murinae distributed in the Australo-Papuan region may lead to the discovery of more GALV-like retroviruses. The vast majority of samples from this study and Simmons et al. (31) were either from Australia or mainland Southeast Asia. Therefore Wallacea, Indonesia and Papua New Guinea are still almost completely unexplored. Since these areas are located in the transitional zone between Australia and Southeast Asia, several species from this region may have been involved in the cross-species transmission between koalas and gibbons. In particular, more *Melomys* species need to be tested. Even though *Melomys paveli* and *M. cervinipes* were found negative for KoRV and GALV in the present study and in Simmons et al. (31) respectively, the most closely related retroviruses to GALV identified to date were discovered in *Melomys burtoni*, and *Melomys* is a widespread murine genus in the Australo-Papuan region which accounts for a total of 23 species, 20 of which are found in Indonesia and Papua New Guinea. Furthermore, within the subfamily Murinae, the genera *Hydromys* and *Uromys*, which belong to the same molecular tribe of *Melomys* (Hydromyini), display similar wide Australo-Papuan distribution and should be targeted in future screening for KoRV and GALV. Also the genus *Rattus*, which has been already

partially screened in this study and Simmons et al. (31) but not extensively, requires further investigation, since several *Rattus* species (e.g. *R*. *exulans*, *R*. *nitidus* and *R*. *tiomanicus*) are distributed both in the Australo-Papuan region and in mainland Southeast Asia, making these species particularly interesting potential hosts. Moreover, in order to clarify the relationship between MelWMV and MbRV, it would be necessary to recover the full genome of MbRV, since only four proviral partial sequences were recovered by Simmons et al. (31). At this stage, indeed, since MelWMV and MbRV share very high nucleotide identity, it is not possible to understand if they represent two independent infections or if one of the two derived from the other. Sequencing the integration sites of MbRV could solve this issue and hybridization capture would be useful for this purpose.

In conclusion, this thesis shows the great potential of high-throughput sequencing in combination with target enrichment techniques such as hybridization capture or amplicon sequencing in both microbiological and virological research. In particular, 16S amplicon high-throughput sequencing helped to characterize the complex microbial communities inhabiting different body regions of koalas. This data will be useful to investigate the health status of koalas based on their bacterial communities and to assess the effect of diseases seriously affecting koala health, such as KoRV and *Chlamydia* infection. Hybridization capture of KoRV- and GALV-like sequences from a large set of possible intermediate vectors has revealed a new GALV host which may represent one of the intermediate hosts used by the ancestors of GALV and KoRV to move across Southeast Asia and Australia and finally infect gibbons and koalas.

References

1. **Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF.** 2000. The nucleotide sequence of koala (*Phascolarctos cinereus*) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus. J Virol **74:**4264-4272.
2. **Stoye JP.** 2006. Koala retrovirus: a genome invasion in real time. Genome Biol **7:**241.
3. **Tarlinton R, Meers J, Hanger J, Young P.** 2005. Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. J Gen Virol **86:**783-787.
4. **Tarlinton R, Meers J, Young P.** 2008. Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci **65:**3413-3421.
5. **Macphee RD, Greenwood AD.** 2013. Infectious disease, endangerment, and extinction. Int J Evol Biol **2013:**571939.

6. **Fiebig U, Hartmann MG, Bannert N, Kurth R, Denner J.** 2006. Transspecies transmission of the endogenous koala retrovirus. J Virol **80:**5651-5654.
7. **Martin J, Herniou E, Cook J, O'Neill RW, Tristem M.** 1999. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. J Virol **73:**2442-2449.
8. **Cockram FA, Jackson AR.** 1981. Keratoconjunctivitis of the koala, Phascolarctos cinereus, caused by *Chlamydia psittaci*. J Wildl Dis **17:**497-504.
9. **Polkinghorne A, Hanger J, Timms P.** 2013. Recent advances in understanding the biology, epidemiology and control of chlamydial infections in koalas. Vet Microbiol **165:**214-223.
10. **Jackson M, White N, Giffard P, Timms P.** 1999. Epizootiology of Chlamydia infections in two free-range koala populations. Vet Microbiol **65:**255-264.
11. **Cork SJ, Hume, I. D. & Dawson, T. J. .** 1983. Digestion and metabolism of a natural foliar diet (*Eucalyptus punctata*) by an arboreal marsupial, the koala (*Phascolarctos cinereus*). Journal of Comparative Physiology B **153:**181-190.
12. **Eberhard IH, McNamara, J., Pearse, R. J. & Southwell, I.A. .** 1975. Ingestion and excretion of *Eucalyptus punctata* DC and its essential oil by the Koala, *Phascolarctos cinereus* (Goldfuss). Aust. J. Zool. **23:**169-179.
13. **Barker CJ, Gillett A, Polkinghorne A, Timms P.** 2013. Investigation of the koala (*Phascolarctos cinereus*) hindgut microbiome via 16S pyrosequencing. Vet Microbiol **167:**554-564.
14. **Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, et al.** 2011. Enterotypes of the human gut microbiome. Nature **473:**174-180.
15. **Shreiner AB, Kao JY, Young VB.** 2015. The gut microbiome in health and in disease. Curr Opin Gastroenterol **31:**69-75.
16. **Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard JM, Poinar HN.** 2014. Ancient whole genome enrichment using baits built from modern DNA. Mol Biol Evol **31:**1292-1294.
17. **Mason VC, Li G, Helgen KM, Murphy WJ.** 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Res **21:**1695-1704.
18. **Coffin JM, Hughes SH, Varmus HE.** 1997. The Interactions of Retroviruses and their Hosts. *In* Coffin JM, Hughes SH, Varmus HE (ed.), Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

19. **Oliveira NM, Satija H, Kouwenhoven IA, Eiden MV.** 2007. Changes in viral protein function that accompany retroviral endogenization. Proc Natl Acad Sci U S A **104:**17506-17511.
20. **Ishida Y, McCallister C, Nikolaidis N, Tsangaras K, Helgen KM, Greenwood AD, Roca AL.** 2015. Sequence variation of koala retrovirus transmembrane protein p15E among koalas from different geographic regions. Virology **475:**28-36.
21. **Demirov DG, Freed EO.** 2004. Retrovirus budding. Virus Res **106:**87-102.
22. **Shojima T, Hoshino S, Abe M, Yasuda J, Shogen H, Kobayashi T, Miyazawa T.** 2013. Construction and characterization of an infectious molecular clone of Koala retrovirus. J Virol **87:**5081-5088.
23. **Battini JL, Danos O, Heard JM.** 1995. Receptor-binding domain of murine leukemia virus envelope glycoproteins. J Virol **69:**713-719.
24. **Overbaugh J, Miller AD, Eiden MV.** 2001. 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 **65:**371-389, table of contents.
25. **Ting YT, Wilson CA, Farrell KB, Chaudry GJ, Eiden MV.** 1998. Simian sarcoma-associated virus fails to infect Chinese hamster cells despite the presence of functional gibbon ape leukemia virus receptors. J Virol **72:**9453-9458.
26. **Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL.** 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet **8:**e1002764.
27. **Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K.** 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol **28:**3033-3043.
28. **Coffin JM.** 2013. Virions at the gates: receptors and the host-virus arms race. PLoS Biol **11:**e1001574.
29. **Shojima T, Yoshikawa R, Hoshino S, Shimode S, Nakagawa S, Ohata T, Nakaoka R, Miyazawa T.** 2013. Identification of a novel subgroup of Koala retrovirus from Koalas in Japanese zoos. J Virol **87:**9943-9948.
30. **Xu W, Stadler CK, Gorman K, Jensen N, Kim D, Zheng H, Tang S, Switzer WM, Pye GW, Eiden MV.** 2013. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proc Natl Acad Sci U S A **110:**11547-11552.
31. **Simmons G, Clarke D, McKee J, Young P, Meers J.** 2014. Discovery of a novel retrovirus sequence in an Australian native rodent (*Melomys burtoni*): a putative link between gibbon ape leukemia virus and koala retrovirus. PLoS One **9:**e106954.

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

33. **Benveniste RE, Callahan R, Sherr CJ, Chapman V, Todaro GJ.** 1977. Two distinct endogenous type C viruses isolated from the asian rodent *Mus cervicolor*: conservation of virogene sequences in related rodent species. J Virol **21:**849-862.

34. **Callahan R, Meade C, Todaro GJ.** 1979. Isolation of an endogenous type C virus related to the infectious primate type C viruses from the Asian rodent *Vandeleuria oleracea*. J Virol **30:**124-131.

35. **Cui J, Tachedjian G, Tachedjian M, Holmes EC, Zhang S, Wang LF.** 2012. Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats. J Gen Virol **93:**2037-2045.

List of publications

This thesis is based on the following manuscripts:

1. Alfano N, Courtiol A, Vielgrader H, Timms P, Roca AL, Greenwood AD. 2015. Variation in koala microbiomes within and between individuals: effect of body region and captivity status. *Scientifc Reports*, **5**: 10189. http://dx.doi.org/10.1038/srep10189.

Niccolò Alfano and Alex D. Greenwood designed the project; Hanna Vielgrader collected and provided the samples; Niccolò Alfano performed all laboratory experiments; Niccolò Alfano performed all the bioinformatics for the analysis of the high-throughput sequencing reads; Niccolò Alfano and Alexandre Courtiol analyzed the data; Niccolò Alfano, Peter Timm, Alfred L. Roca and Alex D. Greenwood discussed the results and wrote the manuscript.

2. Alfano N, Kolokotronis SO, Tsangaras K, Roca AL, Xu W, Eiden MV, Greenwood AD. 2015. Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses. *Journal of virology*, **90**:1757-1772. http://dx.doi.org/10.1128/JVI.02745-15.

Niccolò Alfano and Alex D. Greenwood designed the project; Maribeth V. Eiden and Wenqin Xu extracted and provided the DNA from the GALV-infected cell lines; Kyriakos Tsangaras performed part of the PCRs on the GALV strains; Niccolò Alfano performed the DNA extractions, PCRs and hybridization capture experiments; Niccolò Alfano performed all the bioinformatic analyses; Niccolò Alfano and Sergios-Orestis Kolokotronis performed the phylogenetic analyses; Maribeth V. Eiden and Wenqin Xu performed the functional analyses on GALV and KoRV-B envelope proteins; Niccolò Alfano, Sergios-Orestis Kolokotronis, Alfred L. Roca, Maribeth V. Eiden and Alex D. Greenwood discussed the results and wrote the manuscript.

3. Alfano N, Michaux J, Morand S, Aplin K, Tsangaras K, Löber U, Fabre PH, Fitriana Y, Semiadi G, Ishida Y, Helgen KM, Roca AL, Eiden MV, Greenwood AD. 2016. An endogenous gibbon ape leukemia virus (GALV) identified in a rodent (*Melomys burtoni* subsp.) from Wallacea. *Journal of virology* (In review). http://dx.doi.org/10.1128/JVI.00723-16.

Niccolò Alfano and Alex D. Greenwood designed the project; Johan Michaux, Serge Morand, Ken Aplin, Pierre-Henri Fabre, Yuli Fitriana, Gono Semiadi and Kristofer M. Helgen collected and provided the rodent samples; Kyriakos Tsangaras performed a preliminary study where he used hybridization capture to screen 10 Southeast Asian rodent samples for the presence of KoRV and GALV sequences; Yasuko Ishida generated the PCR products covering the genome of KoRV to be used as hybridization capture baits; Maribeth V. Eiden provided the DNA from the GALV-infected cell lines; Niccolò Alfano performed the DNA extractions, PCRs and hybridization capture experiments; Ulrike Löber wrote the script to automatize some of the bioinformatic analyses; Niccolò Alfano performed all the bioinformatic and phylogenetic analyses; Pierre-Henri Fabre, Ken Aplin, Kristofer M. Helgen and Johan Michaux provided important information regarding the distribution of the rodent species; Niccolò Alfano, Pierre-Henri Fabre, Alfred L. Roca, Maribeth V. Eiden and Alex D. Greenwood discussed the results and wrote the manuscript.

Acknowledgements

I would like to thank all those people who really helped me during my PhD and contributed to make these years a lot easier and more enjoyable. First of all, I would like to thank Prof. Alex Greenwood for choosing me for this PhD position, for being always present during these 4 years and finding always the time to talk or review my work and for his guidance during these years of growth as a scientist. I would like to thank Prof. Heribert Hofer as well for giving me the possibility to work at the IZW, which was a very lively and friendly environment to work every day and do science, and for his advices and support as second supervisor and during the thesis committee meetings.

I am also grateful to all the members of the Wildlife Diseases Department at the IZW. If I was happy to come to work every day it was especially because of the friendly and cosy atmosphere I found in this group. A special thank to Gabor Czirjak, whose office door was always open for me, for being always ready to listen, give advices and hearten during the stressful moments. You are a great PhD students advisor, Gabor one of us! Another special thank to Kyriakos Tsangaras who really helped me a lot at the beginning of my PhD and was an invaluable guide to start my work at the IZW. Thanks also to Alexander Hecht for sharing many years of PhD and therefore many chats, troubles and good moments. A big thank and many many hugs go to all my other mates during this PhD trip, the ones who were there from the beginning, Zaida, Luis, Marie-Louise, Olia, Ximena, Marina, Jundong, Azza and Pin, and the ones I met along the way, Sanatana, David, Ulrike, Peter, Daniela, Paula, John, Anisha, Saskia, Sonia, Renata, Tanja, and all the other students I met during this time. Thank you guys: you have been not only great colleagues, but also good friends, and I have been happy to share the nice and fun time with you. I really hope to keep most of you as friends, wherever we will be. Thanks also to Karin Hoening and Katja Pohle for their technical support. I am thankful also to Stephanie Vollberg for being always so available in helping me in several bureaucratic issues.

I am also grateful to all the co-authors on my papers, especially Alfred L. Roca and Maribeth Eiden, for their helpful comments and contribution in reviewing and improving the manuscripts, Alexandre Courtiol for his comments and the great help on the microbiome paper, and Sergios-Orestis Kolokotronis for his contribution in the phylogenetic analyses. I wish also to thank ZIBI graduate school and the International Max Planck Research School for Infectious Diseases and Immunology which gave me the support to attend very interesting and helpful courses, workshops and conferences. Thanks also to my closest ZIBI mates, Gianna and Georg for sharing many good

moments during these years and their friendship. The start in Berlin was great thanks to you guys.

Finally a special thank to my parents, who supported me all the time, for being always ready to listen, care and help me in every aspect of my life and for their endless love.

Curriculum Vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

For reasons of data protection, the Curriculum vitae is not published in the online version.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Dissertation selbständig, ohne unzulässige fremde Hilfe und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

I hereby confirm that I have made this work autonomously. I assure that I have read and used only the specified sources claimed in this work.

Berlin, 2nd June, 2016
Niccolò Alfano