Leibniz Institute for Zoo
and Wildlife Research
IN THE FORSCHUNGSVERBUND BERLIN E.V.

Freie Universität Berlin

Establishing high throughput genomic and computational methods for the real time study of retroviral endogenization and evolution

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

Date of defense: January 8th 2016

by
Pin Cui
from China

2015

This Dissertation was done in the Leibniz-Institute for Zoo and Wildlife Research in Berlin during the period 19.07.2011-30.05.2015 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. Dr. Greenwood Alex
2nd Reviewer: Prof. Dr. Heribert Hofer

Manuscripts used for this thesis and my contribution

The **chapter II** is based on the following manuscript. My main contribution to this manuscript is performing hybridization cpature experiment and I was also substantially involved in the data analysis for this work.

Tsangaras K, Siracusa MC, Nikolaidis N, Ishida Y, **Cui P**, Vielgrader H, et al. (2014) Hybridization Capture Reveals Evolution and Conservation across the Entire Koala Retrovirus Genome. PLoS ONE 9(4): e95633.

The **chapter III** is based on the following manuscript. I contributed to the manuscript by performing the whole experiment and most of data analysis except for statistic modeling. And I wrote the manuscript.

Pin Cui, Ulrike Löber, Yasuko Ishida, David E. Alquezar-Planas, Alexandre Courtiol, Peter Timms, Rebecca Johnson, Dorina Lenz, Kristofer M. Helgen, Alfred L. Roca, Stefanie Hartmann, and Alex D. Greenwood. (2015) Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without a reference genome. (In review with BMC Genomics)

The **chapter IV** is based on the following manuscript. I contributed to the manuscripte by performing the whole experiment and most of data analysis except for mitochondrial phylogenetic analysis. And I wrote the manuscript.

Pin Cui, Graham Slater, Dorina Lenz, Kyriakos Tsangaras, Bryson Voirin, Nadia de Moraes, Analía M. Forasiepi, Ross D. E. MacPhee and Alex D. Greenwood. (2015) Evolutionary relationships among extinct and extant sloths: the evidence of mitogenomes and retroviruses. (In review with Genome Biology and Evolution)

Contents

Acknowledgments 1

Zusammenfassung 2

Summary 4

Chapter I: General Introduction

1.1.1 Retrovirus and Retrovirology 7

1.1.2 Retroviral integration and biological impact 10

1.1.3 Retroviral Endogenization and Endogenous Retrovirus (EVRs) 11

1.1.4 Dating the Retroviral Endogenization 12

1.1.5 Retrovirus-host interaction 13

1.1.6 Retrovirus host co-evolution and co-speciation 15

1.1.7 Koala retrovirus (KoRV) as model to study endogenization 16

1.1.8 Sloth Endogenous Retrovirus (SloEFV) and sloths 17

1.1.9 Methods applied in ERV research 18

1.2 Study aims 20

1.3 References 20

Chapter II: Hybridization Capture Reveals Evolution and Conservation across the Entire Koala Retrovirus Genome

2.1 Summary 28

2.2 Introduction 28

2.3 Materials and Methods 30

2.4 Results 37

2.5 Discussion 48

2.6 References 51

2.7 Supplementary material 55

Chapter III: Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without a reference genome

3.1 Summary 68

3.2 Introduction 68

3.3 Materials and Methods 69

3.4 Results 80

3.5 Discussion 85

3.6 References 87

3.7 Supplementary material 92

Chapter IV: Mitogenomes and retroviral sequences from extant and extinct sloths reveal complex evolutionary trends among edentates

4.1 Summary 100

4.2 Introduction 100

4.3 Materials and Methods 103

4.4 Results 103

4.5 Discussion 115

4.6 References 119

4.7 Supplementary material 125

Chapter V: Concluding remarks 144

Curriculum vitae 147

Acknowledgments

First, I want to thank my supervisor Prof. Alex Greenwood for supervision during these years. And the opportunity for me to learn bioinformatics skills is very helpful.

I would like to offer my great gratitude to Prof. Heribert Hofer for his kind support and encouragement, he is my co-supervisor and also director of our institute (IZW).

I also want to thank Dr. Wei Chen very much, he is a member of my thesis committee and has been substantially involved in the design of my thesis. And during my work on the projects, he has offered a lot of help including letting me use some of his equipment.

It is my pleasure to perform my thesis work in Leibniz Institute for Zoo and Wildlife Research (IZW). Thanks to all colleagues in FG3 of IZW. Especially, I want to thank Katja Pohle and Nicole Dinse for excellent technical assistance. And I want to say Ulrike Löber and David Alquezar-Planas are the most helpful to me among all the scientists in FG3. I also want to thank Dr. Gudrun Wilbert and Dr. Claudia A. Szentiks for their kind support during my whole PhD years. I also would like to thank colleagues beyond FG3, especially Dorina Lenz, Dr. Beate Braun, and Prof. Jörns Fickel .

I also want to thank our colleagues in Berlin Center for Genomics in Biodiversity Research, especially, Felix Heeger, Susan Mbedi and Camila Mazzoni.

My funding agency of course deserves my gratitude. I thank the China Scholarship Council for PhD fellowship, and I also thank Leibniz Institute for Zoo and Wildlife Research for offering my very nice contract during the last few months to support me to finish my thesis work!

Without the kind help from my dear collaborators, the completion of my thesis is impossible. I offer my great gratitude to Dr. PD. Stefanie Hartmann, Jochen Singer, Prof. Knut Reinert and Wei Sun in Germany, and to Prof. Graham Slater, Dr. Alfred L. Roca, Dr. Yasuko Ishida and Prof. Ross D. E. MacPhee in U. S. A.

Finally, I dedicated this thesis to my God and Lord, Jesus Christ, his precious blood has washed away all my sin. Now I understand that he took me to Germany not only to do this PhD but to save my soul and re-shape my personality through these PhD years.

Zusammenfassung

Endogene Retroviren (ERVs) stammen von exogenen Retroviren ab, die Vorfahren infiziert haben und sich in die Keimbahn von Vertebratengenomen insertiert haben, sodass sie nach den Mendelschen Regeln vererbt werden. Bis zu 10% eines Vertebraten Genoms bestehen aus Sequenzen retroviralen Ursprungs. Im Gegensatz zu den meisten anderen ERVs, die alte Infektionen darstellen, ist der Koala Retrovirus (KoRV) in einem Übergangsstadium zwischen einem exogenen und einem endogenen Retrovirus. Somit ist KoRV ein einzigartiges Model um den Prozess retroviraler Endogenisierung zu Untersuchen. Um den Prozess der Endogenisierung eines Retroviruses zu verstehen ist es wichtig herauszufinden, wo KoRV in das Wirtsgenom integriert wird und wie spezifische provirale Integrationen in Koalapopulationen verteilt sind. Um den Verlauf der Virusintegration in das Koalagenom nachzuvollziehen, wurden Museumskoalas, welche aus den Jahren 1870-1980, untersucht, um KoRV Integrationen über eine längere Zeitspanne und die Verbreitung des Viruses zu ergründen. Die Analysen genetischen Materials von alten Exponaten ist schwierig, da die alte Erbinformation (ancient DNA - aDNA) in den Proben bereits stark beschädigt sein kann und somit konventionelle molekularbiologische Methoden wie die Polymerase Kettenreaktion (polymerase chain reaction - PCR) nicht angewand werden können. Zur Zeit existiert kein assembliertes Koalagenom, sodass die Sequenzierungsdaten aus dieser Studie nicht mit referenzbasierten Standardmethoden der Bioinformatik analysiert werden können. Aus diesem Grund müssen neue Methoden entwickelt werden um die Integration retroviraler Sequenzen in alten Proben nachzuvollziehen. Auch diese Gegebenheiten machen es nötig neue experimentelle und analytische Ansätze zu entwickeln. Das Ziel dieser Arbeit ist es eine Methode zu etablieren, in der es experimentelle und computergestützte Analysen von Hochdurchsatzsequenzierungsdaten ermöglichen, die Evolution und Endogenisierung von ERVs in Echtzeit mit Hilfe historischer Proben zu untersuchen.

In Kapitel 2 habe ich die Methode “hybridisation capture” angewandt um KoRV-Sequenzen aus DNA-Extrakten der Museumsproben von 1870-1990 und einer weiteren Probe eines modernen Koalas anzureichern. Die zusammengefassten konzentrierten Produkte wurden durch Illumina Multiplex sequenziert. Wir entwickelten eine bioinformatische Methodik, welche es ermöglicht komplette KoRV Genome von sechs Museumsproben und dem rezenten Koala zu determinieren. 138 Polymorphismen konnten bestimmt werden, von denen 72 Polymorphismen in mehr als einem Koala entdeckt wurden. Es wurde nicht ein Polymorphismus (in zwei genomischen KoRV Regionen entdeckt,) der als infektiös eingeschätzt wird. Auch Sequenzen des Wirtes, die die Integrationsstellen viralen Sequenzen flankieren, wurden erfasst; einige provirale Loci sind in mehreren Koalas detektiert worden. Zwei der derzeit beschriebenen KoRV-Varianten (KorV-B und KoRV-J) konnten in keiner der Museumsproben nachgewiesen werden, was darauf schließen lässt, dass diese Varianten erst in der heutigen Zeit auftreten.

Kapitel 3 befasst sich mit dem Vergleich und der Modifizierung dreier Techniken zur gezielten Anreicherung von DNA für die Sequenzierung mittels Illumina um KoRV Integrationsstellen der 13 Museumsproben, von Koalas zwischen 1870 und 1980, zur identifizieren und charakterisieren. Um die kurzen Integrationsstellen aus Millionen von

Illumina Sequenzen zu erfassen, habe ich eine Cluster-basierte Methodik entwickelt die unabhängig von Referenzen (Wirtsgenom) anwendbar ist. Vergleicht man die drei Anreicherungsmethoden, so zeigt sich, dass unterschiedliche Ergebnisse hervorgehen, generell kann man aber sagen, dass die zielgerichteten Methoden am besten funktioniert haben. In Verbindung mit zuvor publizierten Forschungsarbeiten zu Integrationsstellen von KoRV in modernen und alten Koalabären ist es naheliegend, dass der Anteil von KoRV-Integrationsstellen die in verschiedenen Koalapopulationen gefunden werden innerhalb der letzten 140 Jahre zugenommen hat.

Kapitel 4 behandelt die Modifizierung der “hybrid capture”-Methode zur Anwendung auf gezielte Illumina-Sequenzierung mitochondrialer Genome (Mitogenome) und Teilen des Polymerasegens des endogenen Faultier Viruses (sloth endogenous foamy virus- SloEFV) aus zwei ausgestorbenen Faultierarten und drei rezenten Faultierarten. Durch den Vergleich verschiedener informationstechnischer Methoden habe ich eine effiziente Prozedur entwickelt welche es erlaubt alte DNA Sequenzen zu charakterisieren, auch wenn nur Referenzgenome weit entfernter rezenter Arten vorhanden sind. Die mitochondrialen “hybridization-capture”-Daten ermöglichten eine komplette phylogenetische Analyse, die von der Phylogenie basierens auf morphologischen Merkmalen heute lebender Faultiere abweicht. Der Vergleich des phylogenetischen Baums des Mitogenoms und der lebenden sowie ausgestorbenen Faultiere zeigt, dass mehrfache komplexe Invasionen durch SloEFV in die Keimbahn der Vorfahren verschiedener Faultierlinien, gefolgt von anschließenden Introgressionen, stattgefunden haben.

Summary

Endogenous retroviruses (ERVs) descend from exogenous retroviruses that have infected the ancestral germ line of vertebrates becoming Mendelian traits. They make up to 10% of vertebrate genomes. Unlike most ERVs which represent ancient infections, the koala retrovirus (KoRV) is a retrovirus that is transitioning from an exogenous to endogenous state, providing a unique model to study the process of retroviral endogenization. An important feature of understanding the retroviral endogenization process is to examine where KoRV integrates and how specific proviral integrations either spread among koalas or fail to. To track the integration history of KoRV in real time, museum koalas collected from 1870s to 1980s are a potential source of understanding the spread of KoRV integrations among koalas over time. However, this is technically highly challenging because the genetic material in museum samples, generally regarded as ancient DNA (aDNA), is heavily degraded, for which conventional genetic methods like PCR cannot be used. There is no assembled koala genome available, so sequencing data of such study cannot be analyzed using standard bioinformatic approaches. Therefore, new approaches are needed to enrich and analyze retroviral integration sites and proviral sequences which can be applied to historical samples. Furthermore, ERVs have undergone co-evolution and co-divergence with their hosts both over very long periods of evolutionary history and over shorter periods accessible directly by examining aDNA from the Pleistocene. However, this also requires novel approaches at the experimental and analytical levels. The aim of this thesis was to establish high throughput sequencing based experimental and computational methods that can be used to understand the endogenization and evolution of ERVs in real time from historical samples.

In **Chapter II**, I applied hybridization capture to enrich KoRV sequences from DNA extraction of ten museum koalas sampled from 1870s to 1990s and one modern koala, and subsequently sequenced the pooled enrichment products using illumina multiplexed sequencing. The bioinformatic pipeline we established recovered full KoRV genomes from 6 museum koalas and the modern koala. And a total of 138 polymorphisms were detected, of which 72 were found in more than one koala. No polymorphism was detected within two KoRV genomic regions that are believed to affect retroviral infectivity. Host sequences flanking proviral integration sites were also captured; with few proviral loci shared among koalas. Recently described KoRV variants (KoRV-B and KoRV-J) were not detected in museum samples, suggesting that they may be of recent origin.

In **Chapter III,** I modified and compared three target enrichment techniques coupled with illumina sequencing to retrieve and to characterize KoRV integration sites from 13 museum koala samples collected between the 1870's and late 1980's. To identify and sort integration sites from tens of millions of Illumina reads, I established a sequence-clustering based reference (host genome) independent computational pipeline. Although three enrichment methods compared exhibited bias in integration sites retrieval, capture based methods performed best. The results compared to previously described integration sites from modern and museum koalas suggest that the proportion of KoRV integration sites shared among unrelated koalas has increased over the last 140 years.

In **Chapter IV,** I modified hybridization capture and applied it for Illumina targeted sequencing of full mitochondrial genomes (mitogenomes) and partial polymerase gene of SloEFV (sloth endogenous foamy virus), from two extinct and three extant sloth species. By comparing different computational methods, I established an efficient pipeline for characterization of ancient DNA sequence when only distant extant relative were available as a genomic reference. The mitochondrial hybridization capture results produced a fully resolved and strongly supported phylogeny for extinct ground and living tree sloths that conflicts with recent morphological analyses. Comparison of the retroviral gene tree to the mitochondrial phylogeny of both extant and extinct sloths demonstrates multiple complex invasions of SloEFV into the ancestral sloth germline line followed by subsequent introgressions across different sloth lineages.

Chapter I

General Introduction

Chapter I

General Introduction

1.1 Retroviruses and retrovirology

Retroviruses are important pathogens for the scientific community and general public because of their causative roles in cancer and other fatal diseases, e.g. the acquired immunodeficiency syndrome (AIDS) caused by the human immunodeficiency virus (HIV). Retroviruses comprise a large and diverse group of enveloped RNA viruses. Upon retrovirus' entry into host cells, their RNA genome is reverse transcribed into DNA by the viral reverse trancriptase protein. The viral DNA is then integrated into the genome of the host. Integrated viral DNA, namely provirus, serves as template for viral gene expression using the host cell transcription machinery (primarily RNA polymerase II) and to generate RNA copies of retroviral genome that will serve as the genome of progeny viruses (Fields et al 1996). Expressed viral genes are spliced and exported from the nucleus into the cytosol where it is translated. The viral proteins are assembled and viral RNA is packaged. Virions bud and are released from the cell membrane yielding free mature viruses. (Fig 1.)

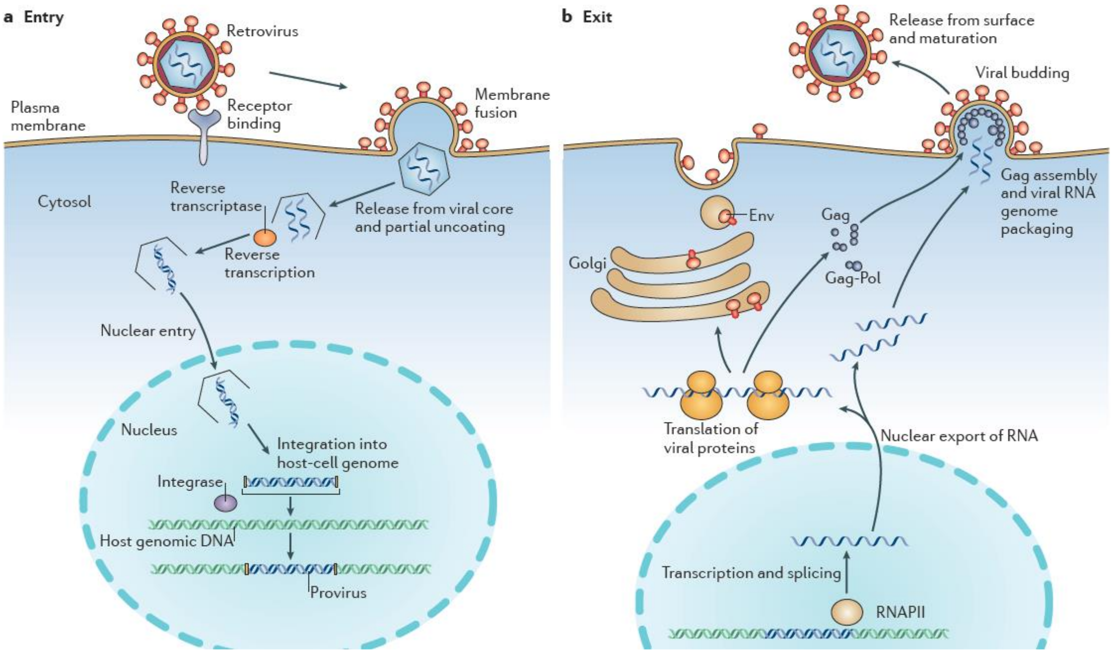
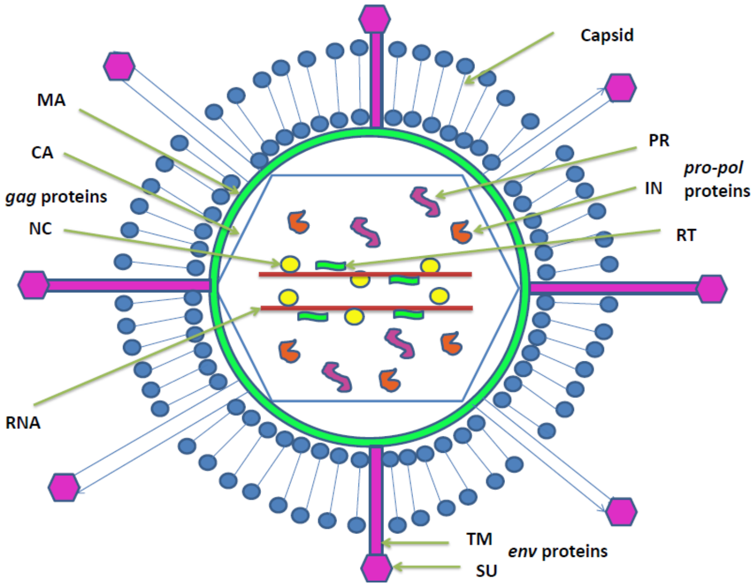


Figure 1. **Steps in the retroviral life cycle.** Different events in the life cycle of retroviruses are illustrated. **a** | Viral entry into cells involves the following steps: binding to a specific receptor on the cell surface; membrane fusion either at the plasma membrane or from endosomes (not shown); release of the viral core and partial uncoating; reverse transcription; transit through the cytoplasm and nuclear entry; and integration into cellular DNA to give a

provirus. **b** | Viral exit involves the following steps: transcription by RNA polymerase II (RNAPII); splicing and nuclear export of viral RNA; translation of viral proteins, Gag assembly and RNA packaging; budding through the cell membrane; and release from the cell surface and virus maturation. (Courtesy from Prof. Jonathan Stoye)

Retroviral virions are 80–100 nm in diameter, and their outer lipid envelope incorporates and displays viral glycoproteins (Fig. 2). The shape and location of the internal protein core are characteristic for distinguishing different retorviruses. The virions embrace two identical single-stranded RNA molecules 5-13 kb in size.

Figure 2. Schematic representation of a retrovirus particle structure. Adapted from Rodrigues et al 2011.



Retroviral genomes contain three major open reading frames which code for proteins essential for viral structure and function: *group-specific antigen (gag)*, which codes for virion proteins including the matrix, the capsid, and the nucleoprotein proteins; *polymerase* (*pol)*, which codes for the reverse transcriptase and integrase enzymes, vital for retroviral replication; and *envelope* (*env)*, which codes for the surface and transmembrane components of the viral envelope protein. Additionally, *pro*, which encodes the virion protease, is a smaller but vital gene locating between *gag* and *pol*. Retroviruses whose genome only contain the basic genes are called simple retroviruses, (Fig. 3). In contrast, complex retroviruses produce additional spliced transcripts giving rise to additional mRNAs and greater variety of gene products (up to six in addition to the *gag, pro, pol*, and *env* proteins in HIV and SIV) (Murphy et al. 1994).

These additional products exert control over cellular functions that, for simple retroviruses, are provided by the host. Through this additional control, complex viruses infect adult, immunocompetent, animals much more frequently than do simple viruses (Coffin JM 1997b).

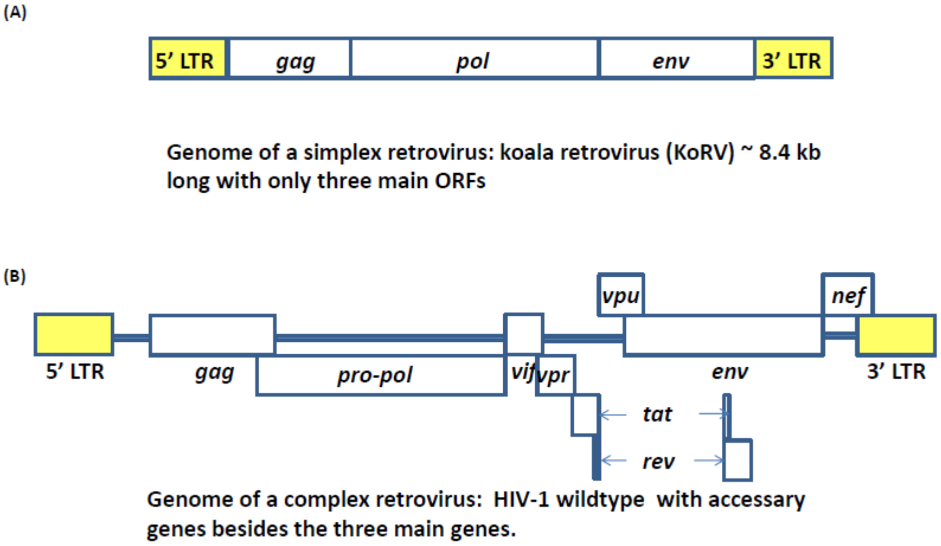

(A)
5' LTR
gag
pol
env
3' LTR
Genome of a simplex retrovirus: koala retrovirus (KoRV) ~ 8.4 kb long with only three main ORFs
(B)
vpu
nef
5' LTR
gag
pro-pol
vif
vpr
env
3' LTR
tat
rev
Genome of a complex retrovirus: HIV-1 wildtype with accessary genes besides the three main genes.

Figure 3. Retroviral genomes. Schematic representation of (A) MLV and (B) HIV-1 wild-type genomes representing simple and complex retrovirus, respectively. Adapted from Rodrigues et al 2011.

Retroviruses are classified into seven groups defined by evolutionary relatedness (Table 1). Except for the lentiviruses and spumaviruses, the remaining five groups contain retroviruses with oncogenic potential (regarded as oncoviruses). Deltaretrovirus, Lentiviruses and spuma - viruses are represent the complex retroviruses.

Table 1. Classification of retrovirus

Group	Type spieces	Virion morphology	Genome
Alpharetrovirus	Rous sarcoma virus	central, spherical core “C particles”	simple
Betaretrovirus	Mouse mammary tumor virus	eccentric, spherical core “B particles”	simple
Gammaretrovirus	murine leukemia virus	central, spherical core “C particles”	simple
Deltaretrovirus	Human T-lymphotropic virus	central, spherical core	complex
Epsilonretrovirus	Walleye dermal sarcoma virus	cylindrical core “D particles”	simple
Lentiretrovirus	human immunodeficiency virus	cone-shaped core	complex
Spumaretrovirus	human foamy virus	central, spherical core	complex

Modified from table 1 of Coffin et al 1997 http://www.ncbi.nlm.nih.gov/books/NBK19382/.

In 1911, the first oncovirus Rous sarcoma virus (RSV) was discovered in tissue tumors (sarcoma) in chickens by American virologist Peyton Rous. Cancer can be triggered by proto-oncogenes incorporated into proviral DNA or by the disruption of cellular proto-oncogenes. RSV contains the *src* gene that triggers uncontrolled growth in abnormal host cells. Thus RSV has been a model for the molecular study of cancer development. Since then, many more oncogenic retroviruses have been discovered. The lentiviruses are also pathogens though not oncogenic. A representative is human immunodeficiency virus (HIV), the causative agent of AIDS.

1.2 Retroviral integration

Retroviral integration is the stable insertion of DNA copy of retroviral genome into the host genome. It is an essential step of viral replication cycle (Williams & Wilkins, 2007) and has profound biological consequences for the host by altering the expression of genes surrounding the retroviral integration site (the position in the host where the retrovirus integrated), sometimes generating essential biological function, e.g. acquisition of the *syncytin* gene in placenta mammals (Mi et al 2000). If integration site happened to be in oncogenes, the normal host cells may be transformed into cancer cells (Hayward et al 1981). If retrovirus happens to integrate into the coding region of none oncogenes, their expression can sometimes be disrupted leading to a visible phenotypic change, eg. hairless trait in mutant mice (Coffin 2004). What is more, retrotranspositional activity may significantly add to genomic variability (Wang et al 2010) and instability of the host.

The retroviral DNA integration process includes three major steps: 1) Processing, integrase (IN) removes two nucleotides from the 3′ ends of the viral DNA; 2) Joining, these newly created ends are joined to staggered phosphates in the host DNA in a concerted cleavage and ligation reaction. This reaction creates an integration intermediate, called the pre-integration complex (PIC), with gaps in the flanking host DNA sequence; 3) Repair, these gaps are filled and the 5′ ends of the viral DNA are joined to the host DNA, creating a stably integrated provirus.

The PIC is in principle capable of directing integration of the viral DNA into any chromosomal locations and thus retroviral integration is not sequence specific. In this regard, all locations of host genome/chromosomes could potentially be candidates for retroviral integration (Cereseto et al 2004). The retroviral integration sites are generally defined as the host genomic sequences flanking the proviruses. However, the distribution of retroviral integration sites in host genome is found not entirely random. For example, many proviruses are preferentially detected in chromatin regions with actively transcribed genes. Such choice of the integration site make sense in the light of evolution, since integration at a transcription active site is beneficial for the proliferation of the retrovirus. It has been reported that, for the seven groups in retroviral taxonomy (Table 1), the preference of integration site selection exhibits group-specific patterns (Derse et al 2007; Cavazza 2013). Some groups show strong preference: lentiviruses (eg. HIV-1) prefer to integrate within the bodies of active genes located within gene dense regions (Schroder et al 2002) while gammaretroviruses prefer to integrate in the vicinity of strong enhancers, active gene promoters and associated CpG islands (Wu et al 2003; LaFave et al 2014; De Ravin et al 2014). In contrast, the alpharetroviruses and deltaretroviruses show a strong preference for active genes or transcription start sites (TSSs) (Narezkina et al 2004). The betaretroviruses are the least selective, displaying a random integration pattern on the genomic level (Faschinger et al 2008; Konstantoulas et al 2014).

Study of the mechanism of retroviral integration have revealed two key determinants for integration site selection (Kvaratskhelia et al 2014): the retroviral integrase (IN) protein and cognate cellular binding partners (Lewinski et al 2006; Ciuffi et al 2006). For example, HIV-1 and Moloney murine leukemia virus targeting preferences are in large part guided by integrase-interacting host factors (LEDGF/ p75 for HIV-1 and BET proteins for MoMLV) that tether viral intasomes to chromatin. In the case of lentivi-ral INs, integration site targeting is in large part guided by the cellular chromatin binding protein lens epithelium derived growth factor (LEDGF)/p75, which facilitates integration into active gene bodies (Ciuffi et al 2005; Marshall et al 2007). What is more, nucleotide preferences at integration sites by different retrovrial groups seem to be governed by the ability for the integrase protein to locally bend the DNA duplex for pairwise insertion of the viral DNA ends.

1.3 Retroviral endogenization and endogenous retrovirus (ERVs)

Although most retroviruses infect vertebrate somatic cells, occasionally retroviral infection can target germ line cells (the precursor of a sperm or egg of vertebrates), in which the integrated retroviral DNA (provirus) can be transmitted vertically from parent to offspring through Mendelian inheritance. In this way, the retrovirus become a permanently integrated part of host genome passed on from generation to generation. This permanent fossil record of ancient retroviral infection are called Endogenous retroviruses (ERVs), while the somatic cell infecting circulating form of retroviruses are exogenous. A retrovirus can integrate into a host genome multiple times through reinfection (Belshaw et al 2004). After initial integration, ERVs proliferate by generating multiple proviral copies in the host genome providing raw material for purifying selection by the host's antiviral defense. The process by which an exogenous retrovirus become ERV in the host genome is called Retroviral Endogenization (Fig. 4), although some Non-Retroviral RNA Viruses were recently found to endogenize (Horie et al 2010; Belyi et al 2010; Chiba et al 2011).

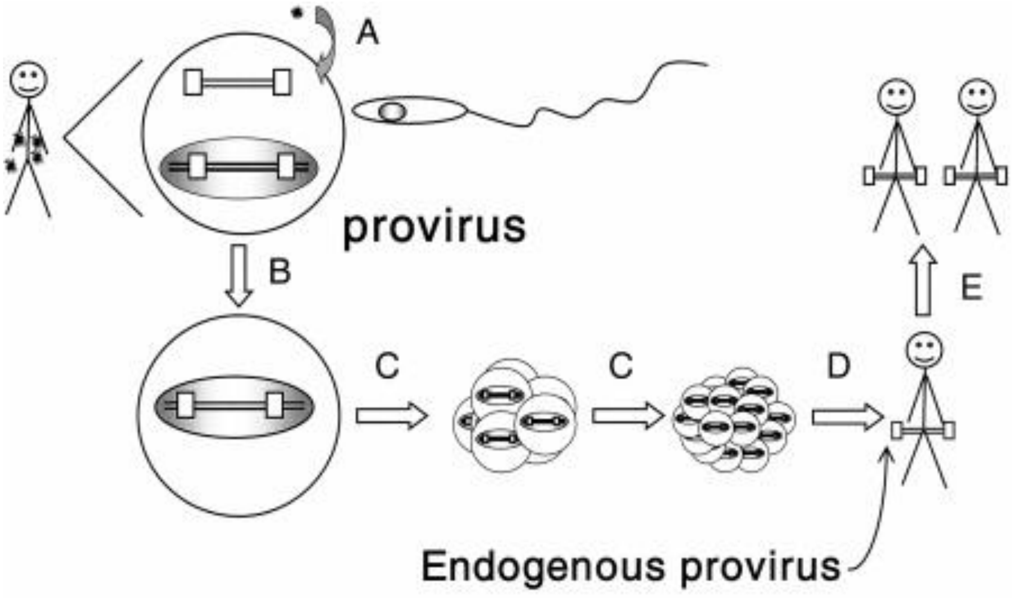


Figure 4. Formation of endogenous proviruses. When a retrovirus infects a cell (A), its genome is copied into a molecule of DNA flanked by long terminal repeats (LTRs) shown as boxes. At the time they are made, the LTRs are identical in sequence. This DNA is integrated at more or less random sites of host DNA to form a provirus. If the infected cell is in the germ line (a precursor of sperm or egg cells), then it will be found in the fertilized zygote (B) and passed on to all cells of the progeny during development (C), and found in all cells of the adult (D), from which it can be passed on to the progeny (E) as though it were a normal gene. Courtesy from Professor John Coffin.

The evolutionary history of vertebrates is accompanied by infection and endogenization of retroviruses. The proliferation of ERVs in vetebrate genome is mainly accomplished in two mechisams, reinfection form the same retrovirus and introgression of retorviral sequence from ERV-carrying host individuals to ERV-free individuals. However, it is found recently that *env*-less ERVs (ERVs that have *env* gene deleted completely) achieved highly successful proliferation through retrotransposition. Through proliferation in evolutionary history, ERVs have colonized a significant portion of host genome, now making up about 4-11% of vertebrate genomes (Lander, E.S., et al 2001; Pontius, J.U., et al. 2007). The diversity of ERVs and retroviral elements in vertebrate genomes are still being characterized, as cheaper and more powerful genome sequencing techniques and sophisticated in silico methods are being developed.

1.4 Dating retroviral endogenization

Two approaches for dating the integration of individual proviruses are commonly used owing to the two distinct properties of retroviral replication.

Upon integration, sequences of the two long terminal repeats (LTRs) of a provirus are identical. After integration, the proviral sequence evolves at the rate of its host (Johnson & Coffin 1999) and the two LTRs evolved separately (about two to four changes per 1,000 base pairs per million years). Therefore, any sequence variation between the two LTRs result from mutation postdating integration. The age of the integrated provirus can be estimated by measuring the sequence divergence between its two LTRs (Tristem 2000). However, estimation of integration date using this approach can be problematic due to differential mutation rates depending on the integration site of different proviruses (Martins et al 2011) or recombination between the LTRs of proviruses in different retroviral groups (Hughes & Coffin 2001). What is more, random mutational change makes the dating of very ancient vertebrate retroviruses inaccurate.

The second approach is based on the fact that, despite the different preference of the integration sites for different retroviral groups, integration sites distribution of multiple proviruses from the same retrovirus is still random. Thus, it is very unlikely that proviruses of the same retrovirus in two vertebrate species have integrated into the same location of the genome by chance, given that vertebrate genomes are usually billions of base pairs in size. Thus any identical integration sites found in two individuals were inherited from their common ancestor, which means one integration event predated the separation of the two species. If a retrovirus recently invaded the genome of a species, the integration sites will be mostly unique among individuals. If an older invasion, all individuals in a species will share the same fixed integration sites. Therefore, investigating the integration sites (shared vs. unique between individuals) provides evidence for the age of the retroviral invasion.

Both approaches have been used indicating most ERVs have integrated millions of years ago but that some integrated more recently and that the process is continuing (Gifford & Tristem 2003). For example, most groups of human ERVs (HERV) are found in all Old World monkeys and apes, suggesting that such ERVs integrations occurred at least 30 million years ago(Shih et al 1991). By contrast, two groups of chimpanzee ERVs are not found in human genome, implying that they entered the chimpanzee genome in the last five million years after the divergence of chimpanzee and human lineages from a common ancestor (Polavarapu et al 2006). HERV-K, recently endogenized, since members of this family were found to have repeatedly infected different lineages in primates both before and after the divergence of the human and chimpanzee lineages and for some HERV-Ks, integrations are polymorphic among humans (Barbulescu et al 1999; Turner et al 2001).

1.5 Retrovirus-host interaction

Vertebrates have developed sophisticated antiviral mechanisms to maintain their genomic integrity and cellular functions, eg. use of cytosine methylation to control ERV expression.

Through this mechanism, the retroviruses largely evolve from harmful pathogens to neutral or even beneficial genomic elements, ultimately reaching a host friendly balance (Coffin 2004).

ERVs accumulate mutations and deletions in their coding sequences with sizes ranging from single nucleotide changes to large deletions/insertions. In fact, for most ERVs, it is very rare to find intact open reading frames, except for recently integrated ERVs, eg. HERV-K. One explanation for the high number of degraded ERVs is that *env*-less retroviral elements are better at proliferating genomically than *env*-intact ones (Magiorkinis et al 2012). Therefore, there appears to be an advantage to both host and ERV when ERVs degrade.

Homologous recombination, the most common genomic recombination of ERVs, happens between the two LTRs of a provirus and results in excision of most of the provirus leaving a solitary LTR, called solo long terminal repeat (solo LTR) in the host genome at the site of the previous provirus (Coffin 2004; Stoye 2012). Retroviruses also undergo extensive recombination during the synthesis of the haploid DNA provirus (Coffin 2004). This represents an extreme form of retroviral degradation as most of the proviral genome is removed.

Beneficial biological impacts have also been observed for ERVs. Besides the retroviruses' direct biological effect upon the genes around integration sites that have been described in the section 1.2, another noteworthy point is their role in assisting host's antiviral defense. Most integrated retroviruses, except for HIV, can block infection of the host by related exogenous viruses. This is mainly achieved by blocking of cell surface molecules (receptors), which are necessary for the virus to enter a cell to start the infection (McDougall, et al. 1994). As shown in Figure 5 (Coffin 2004), vertebrates with integrated proviruses have a strong selective advantage because of their greater ability to resist infection by pathogenic (exogenous) retroviruses (Coffin 2004). This resistance can lead to elimination of the pathogenic retrovirus from the species, while the endogenous retrovirus can be passed on through generations as normal host DNA. Therefore, ERVs provide us with a fossil record of ancient retroviral infections in which ERVs actually played a role of defender for the host. A specific example is the over-expression of the HERV-K accessory protein *Rec* increases IFITM1 levels on the cell surface and inhibits viral infection (Stoye 2012).

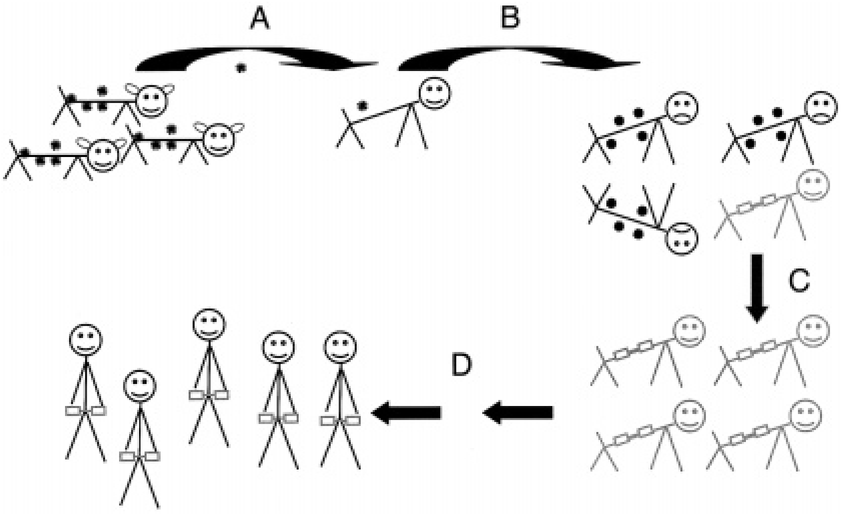


Figure 5. Effect of endogenous proviruses on host-retrovirus interaction. In the case of many retroviruses, the process of transmission and pathogenesis in new species (A.B) can be the same as in the previous figure. However, the formation of endogenous proviruses (gray figures) can block infection by the exogenous virus that gave rise to them, and contribute to their extinction in that host (C). Once fixed, the endogenous proviruses will remain in the genome through eons of evolution (D), remaining as a fossil record of the earlier epidemic. Courtesy from Professor John Coffin.

1.6 Retrovirus host co-evolution and co-speciation

The co-evolution between mammals and retroviruses is an important host–parasite interaction. To study this co-evolution, the accurate estimation of evolutionary age of retrovirus and hosts is necessary. Evolution rates of exogenous retroviruses are generally fast due to purifying selection under hosts' antiviral responses, which makes the estimation of viral age inaccurate. However, endogenous retroviruses (ERVs) integrated into the germline of their hosts and subsequently become fixed in host genome. Through mendelian inheritance, EVRs are generally confined to the host organism's neutral rate of evolution (Feschotte et al 2012), which is much slower than the evolution of their exogenous counterpart.

Following this co-evolution pattern between ERVs and host, a co-divergence scenario of retroviruses and their hosts is expected. If an ancient retrovirus infected an ancestral vertebrate and got endogenized in its genome, subsequently in evolutionary history, this ancestral vertebrate diverged into multiple species, then this EVR in the ancestral vertebrate would co-diverged with the decedent species, and the EVRs in the descendant vertebrate genomes are orthologous to each other since they originate from the a common ancester, the ancient EVR. Vice versa, if orthologous ERVs are found in two or more host species, it can be inferred that the initial infection event predates the divergence of these species (Gilbert et al 2010). At DNA sequence level, the phylogeny of these orthologous EVRs will be in congruent with the phylogeny of the hosts. A cophylogenetic analysis of a retroviral element discovered in the genome of a coelacanth gave an estimated age of 407 Myr for foamy-like viruses (Han et al 2012). Another study involved reconstructing orthologies for several non-retroviral endogenous virus elements, one of which was a virus of the family *Bornaviridae*, found to have a minimum age of 93 Myr; members of the *Parvoviridae* and *Circoviridae* were also examined, with minimum ages of 30 and 60 Myr, respectively (Katzourakis et al 2010). Furthermore, co-divergence study of retroviruses and hosts can also indicate host-retrovirus microevolutionary dynamics (Katzourakis et al 2009).

1.7 Koala retrovirus (KoRV)

Almost all identified ERVs are the result of invasion events many thousands or millions of years old and many have been subject to extensive mutation and deletion (Taruscio, D. & Mantovani, 2004). In many cases, the original exogenous virus from which the ERVs derived are extinct, making it extremely difficult to elucidate the process of retroviral invasion and subsequent inactivation.

The "koala retrovirus" (KoRV), a virus linked to leukemia and immune suppression in Australian koalas (Tarlinton, R.E., et al. 2005), was recently discovered to be spreading both horizontally and vertically among the species. The geographic distribution of KoRV in Australia and the detection of vertical transmission of its proviruses from parent to offspring strongly suggests that it is a recently emergent retrovirus (Tarlinton, R.E., et al. 2006), and it initially infected koalas in northern Australia and is still in the process of spreading to the southern koala populations(Tarlinton, R.E., et al. 2008, Simmons, et al. 2012). Recent discoveries of possibly exogenous KoRV variants in geographically isolated koalas further supported recent endogenization of KoRV(Xu et al 2013; Shojima et al 2013). Therefore, KoRV provides a unique opportunity to study retroviral endogenization. Museum koala samples collected during 19th and 20th centuriesrevealed that KoRV was already ubiquitous by 19th century (Avila-Arcos, et al. 2013) and the entire KoRV genome conserved over 130 years of evolution (Tsangaras, et al. 2014).

The hybridization capture technique described in Tsangaras et al. 2014 detected KoRV integration sites. Comparison among the results from 6 museum koalas and 7 modern koalas showed that there are less shared integrate sites than unique ones. A recent study (Ishida, et al. 2014) compared the integration sites of 39 KoRV proviral loci, and found none was shared among the 5 unrelated modern koalas (2 from the northern, 3 from the south). This suggests very few KoRV proviral integrations are widespread and supports a recent invasion of the koala genome by KoRV. However, both studies lack a comprehensive profiling of the KoRV integrate sites due to the methodology used. Therefore, an efficient method is needed to recover the KoRV endogenization in ' real time ' by intensively investigating the KoRV integration sites from our koala museum collections during the past 130 years.

1.8 Sloth Endogenous Foamy Virus (SloEFV) and sloths

Sloths used to be a diverse group of mammals (including nearly 100 species) that make up the suborder Folivora (or Phyllophaga) under the order Xenarthra. With most of the members going extinct around 10,000 years ago, their modern representatives are reduced to be only 6 species distributed in 2 families, both of which comprise only one genus, *Choloepus* and *Bradypus*, known to the public as the two groups of surviving tree sloths: the three-fingered sloth *Bradypus* (*B. tridactylus, B. variegatus, B. torquatus* and *B. pygmaeus*), and the two-fingered sloth *Choloepus* (*C. hoffmanni and C. didactylus*). The extant sloths resident in the rainforest of Central and South America, while extinct sloths include a few species of aquatic sloths and many ground sloths, some of which attained the size of elephants (eg. *Megalonyx*).

Foamy viruses are nonpathogenic complex retroviruses and form a unique group, *Spuma-retroviridae* (Linial ML 1999). They are widely distributed in all eutherian mammals and have undergone co-evolution and co-speciation with their hosts for over 100 million years (Katzourakis et al 2014). Despite their wide distribution, endogenous foamy virus-like elements have been discovered in only a limited number of mammalian genomes (Katzourakis et al 2009; Han & Worobey 2012a), suggesting that foamy viruses rarely invade the genome. Recent discoveries of endogenous foamy virus in non-mammalian species (Han & Worobey 2012b) suggests their ancient marine origin and co-evolution history with vertebrates of over 407 million years, making foamy viruses some of the oldest within the *Retroviridae* (Rethwilm & Bodem 2013). A co-evolution study foamy viruses and sloths (Katzourakis et al 2009) indicated that sloth endogenous foamy virus (SloEFV) invaded the genome of the ancestor of all sloths 39 million years ago, before the divergence of two and three finger sloths (~21 Ma) but after the anteater and sloth lineages separated (~55Ma).

However, retroviral endogenization is a complicated process in which an exogenous retrovirus initially infects a host and then undergoes continuous amplification, re-infection and re-colonization under host selection pressure before either being removed from the population by drift or becoming fixed in the host genome as an ERV (Gifford & Tristem. 2003). It is possible that different insertions of the same ERV found in a host originate from multiple independent infection or introgression events as ERV containing individuals breed with ERV free individuals. Such a process of introgression is seen in the case of the currently endogenizing retrovirus, Koala retrovirus (KoRV) (Ishida et al 2015). Furthermore, evidence of host-switching and introgression have been found both among distantly related vertebrates (Hayward et al 2013; Hayward et al 2015) and among closely related mammals (Katzourakis, et al. 2014; Jern et al, 2006). The result is that phylogenetic dissociation between host and ERVs are often seen. Examining the patterns of infection or introgression are often difficult from extant taxa because of the long time spans involved and because many hosts and exogenous retroviral counterparts within taxonomic groups are extinct. In addition, to examine cross species transmission and retroviral introgression patterns requires a well resolved host (sloths in this case) phylogeny including the information from extinct lineages. Therefore, there are many open question remaining regarding co-evolution of foamy viruses in sloths.

1.9 Methods applied in ERV research

As an important component of vertabrate genome, ERVs usually exist in high copy number ranging from hundreds to thousands and are highly diverse in sequence even for the proviruses originating from the same initial retroviral infection. Inactivation mutations through evolutionary history eventually leave ERV genomes at different stage of degradation, some proviral loci with relatively intact genome (5-13 kb in size), some with large deletions (eg, env-less ERVs are found to be widespread in Magiorkinis et al 2012) and some loci heavily degraded to be less than 1kb (Katzourakis et al 2009). Adding to the complexity of this question is the various categories of retroviral recombinations. PCR is the main stream enrichment technique for targeted sequencing of retroviruses, and for whole ERV genome investigation, long range PCR was often applied.

In contrast, PCR technique is inefficient for enriching integration sites. The 5' and 3' ends of proviral loci are LTRs, of which the sequence is known, but the host sequence flanking the proviral LTR is unknown. Therefore, normal PCR cannot be applied since primers cannot be designed in the unknown flanking region leaving only a single primer at the LTR side. Although random primers can be employed at the flanks side, this will very possibly miss the actual sequence diversity of integration sites. Conventionally, inverse PCR was used for retrieving retroviral integration sites (Ochman et al 1988). During the last decade, novel methods have been developed like RACE, ligation-mediated PCR and genome walking (Bushman et al 2005; Moalic et al 2006; Schmidt et al 2007; Ciuffi et al 2011; Kustikova et al 2009; Hüser et al 2010), and integratoin site studies have been reported using these methods.

However, for both retroviral genome and integration site studies, these PCR based methods cannot yield comprehensive result and especially do not work with degraded DNA e.g. museum samples which are needed for the real time investigation of KoRV endogenization. The genetic material in museum collections are generally considered to be ancient DNA (aDNA). aDNA research offers a unique opportunity to study evolution in real time, and has become an established field of evolutionary research. However, aDNA research is extremely difficult due to severe DNA degradation, fragmentation and contamination postmortem. The genetic material available for molecular analysis is very short, in low amounts and heavily damaged (Willerslev & Cooper 2005), especially for Pleistocene genetic samples, making conventional genetic methods, eg. PCR, inefficient or even inapplicable.

Several innovative methods have been developed for aDNA applications in the past decades. For instance, Single Primer Extenison (Brotherton et al 2007) and Primer Extension Capture (Briggs et al 2009) have shown great potential for detecting sequence diversity of unknown or variable region flanking a region with known sequence information in aDNA molecules. Recently, high throughput sequencing has opened a new era for all fields of genetic applications, and illumina platform has been widely used in aDNA studies owing to its short read length and enormously high throughput. In combination with hybridization capture (Maricic et al 2010), a highly efficient and affordable target enrichment technique, the application of illumina targeted sequencing to aDNA field has yielded unprecedented results including access to complete mitochondrial and even nuclear ancient genomes (Hagelberg et al 2015). It has also been applied in virology by identifying viral insertion sites from Formalin-fixed, paraffin-embedded tissue (Duncavage et al 2011). These aDNA oriented methods provide opportunities to real time molecular survey of retroviruses or EVRs in specific. However, these methods need to be adapted to meet the specific requirement of retroviral research in high throughput manner.

Equally problematic is the sequence data analysis, because there are currently no bioinformatic software or pipeline dedicated for aDNA based ERVs data analysis, while the conventional bioinformatic tools are inefficient for aDNA data analysis due to the special features of aDNA described above.

For identification of polymorphisms in retroviral sequences, a proper computational threshold is needed to correctly distinguish true mutations and indels from postmortem biochemical damage of aDNA. This reqiures the modification or optimization of available bioinformatic tools to increase their sensitivity and accuracy to work on aDNA data.

For characterization of retroviral integration sites, there are several recently published bioinformatic tools dedicated for such purpose, eg. SLOPE (Duncavage et al 2011), VirusFinder (Wang et al 2013) and VirusSeq (Chen et al 2013), but they all required host genome sequence. Since there is no assembled koala genome available yet, a host genome independent computational pipeline is needed for identifying KoRV integration sites.

Furthermore, correct assembly of the short aDNA sequences to construct multiple proviral loci needs more sophisticated computational manipulation. A baiting and iterative mapping approach (MITObim) (Hahn et al 2013) has been proved to be efficient for the reconstruction of complete mitochondrial genomes of non-model organisms directly from high-throughput sequencing data using distantly related mitochondrial genomes. Considering the possibly huge divergence between ancient retroviral sequences enriched from Pleistocene sloth (Mylodon) and that from modern sloths, systematic optimization of MITObim paprameters is necessary to achieve maximum recovery of ancient retroviral sequences and meanwhile to exclude the possible false positive assembly result.

2, Study aims

Despite the evolutionary and medical importance of ERVs study, the methodology for comprehensive profiling of retroviral genomic changes and integration sites are far from sufficient, especially for challenging task like retrieval of retroviral sequence from museum samples or enrichment of long pieces of EVR proviral genome. The aims of my thesis work were to modify and adapt high throughput sequencing based genomic and computational methods for the integrative study of retroviral endogenization and evolution of EVRs. In this thesis, I describe in **chapter II** the application of hybridization capture for illumina targeted sequencing of KoRV genome from ten museum koalas sampled from 1870s to 1990s and one modern koala. A high throughput data analysis pipeline is also described for investigation of the evolutionary pattern of KoRV genome and KoRV integration sites flanking proviruses. In **chapter III**, I describe the comparison of three advanced target enrichment techniques and a host referece genome independent bioinformatic pipeline for efficient identification and sorting of retroviral integration sites in automatic and high-throughput manner. In **chapter IV**, I describe a modified hybridization capture technique and an optimized MITObim based bionformatic pipeline. Application of the modified techqniue and pipeline leads to successful retrieval of mutiple retrovrial sequences from a 13000 year old Mylodon bone sample. The exeprimental techniques and the bioinformatic pipelines established in my thesis work could be applicable to retrovirological research and beyond.

1.3 References

Abel, H. J., E. J. Duncavage, et al. (2010). "SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data." Bioinformatics **26**(21): 2684-2688.

Alessia Cavazza, Arianna Moiani, and Fulvio Mavilio. Human Gene Therapy. February 2013, 24(2): 119-131. doi:10.1089/hum.2012.203.

Ávila-Arcos, M. C., S. Y. W. Ho, et al. (2013). "One Hundred Twenty Years of Koala Retrovirus Evolution Determined from Museum Skins." Molecular Biology and Evolution **30**(2): 299-304.

Barbulescu M., Turner G., Seaman M.I., Deinard A.S., Kidd K.K., Lenz J. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr. Biol. 1999;26:861–868.

Belshaw, R., V. Pereira, et al. (2004). Long-term reinfection of the human genome by endogenous retroviruses. Proceedings of the National Academy of Sciences of the United States of America. **101**(14): 4894-4899.

Belyi VA, Levine AJ, Skalka AM (2010) Unexpected Inheritance: Multiple Integrations of Ancient Bornavirus and Ebolavirus/Marburgvirus Sequences in Vertebrate Genomes. PLoS Pathog 6(7): e1001030. doi: 10.1371/journal.ppat.1001030

Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., & Pääbo, S. (2009). Primer Extension Capture: Targeted Sequence Retrieval from Heavily Degraded DNA Sources. *Journal of Visualized Experiments : JoVE,* (31), 1573. doi:10.3791/1573

Brotherton, P., Endicott, P., Sanchez, J. J., Beaumont, M., Barnett, R., Austin, J., & Cooper, A. (2007). Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of *post mortem* miscoding lesions. *Nucleic Acids Research*, *35*(17), 5717–5728.

Bushman, F., M. Lewinski, et al. (2005). "Genome-wide analysis of retroviral DNA integration." Nat Rev Micro **3**(11): 848-858.

Cereseto, A. and M. Giacca (2004). "Integration site selection by retroviruses." AIDS Reviews 2004;6:13-21.

Ciuffi A., Llano M., Poeschla E., Hoffmann C., Leipzig J., Shinn P., Ecker J.R., Bushman F. A role for LEDGF/p75 in targeting HIV DNA integration. Nat. Med. 2005;11:1287–1289.

Ciuffi A., Mitchell R.S., Hoffmann C., Leipzig J., Shinn P., Ecker J.R., Bushman F.D. Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. Mol. Ther. 2006;13:366–373.

Ciuffi, A. and S. D. Barr (2010). "Identification of HIV integration sites in infected host genomic DNA." Methods.

Chiba, S., H. Kondo, et al. (2011). Widespread Endogenization of Genome Sequences of Non-Retroviral RNA Viruses into Plant Genomes. PLoS Pathog. **7**(7): e1002146.

Chen, Y., H. Yao, et al. (2013). "VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue." Bioinformatics **29**(2): 266-267.

Chu, H., Y. Jo, et al. (2014). Evolution of endogenous non-retroviral genes integrated into plant genomes. Current Plant Biology. **1**(0): 55-59.

Coffin JM. (2004) Evolution of retroviruses: fossils in our DNA. Proc Am Philos Soc. Sep;148(3):264-80.

Daniel, R., G. Kao, et al. (2003). Evidence that the retroviral DNA integration process triggers an ATR-dependent DNA damage response. Proceedings of the National Academy of Sciences. **100**(8): 4778-4783.

Derse, D., Crise, B., Li, Y., Princler, G., Lum, N., Stewart, C. et al . (2007). Human T-Cell Leukemia Virus Type 1 Integration Target Sites in the Human Genome: Comparison with Those of Other Retroviruses . *Journal of Virology, 81*(12), 6731–6741.

Duncavage, E. J., V. Magrini, et al. (2011). Hybrid Capture and Next-Generation Sequencing Identify Viral Integration Sites from Formalin-Fixed, Paraffin-Embedded Tissue. The Journal of Molecular Diagnostics. **13**(3): 325-333.

Faschinger A., Rouault F., Sollner J., Lukas A., Salmons B., Gunzburg W.H., Indik S. Mouse mammary tumor virus integration site selection in human and mouse genomes. J. Virol. 2008;82:1360–1367.

Feschotte C & Gilbert C (2012) Endogenous viruses: insight into viral evolution and impact on host biology. Nature Reviews Genetics 13: 283-296.

Gifford R, Tristem M. (2003) The evolution, distribution and diversity of endogenous retroviruses. Virus Genes. May;26(3):291-315.

Gilbert C, Feschotte C. (2010) Genomic fossils calibrate the long-term evolution of hepadnaviruses. PLoS Biol. 8:e1000495.

Guan-Zhu Han and Michael Worobey. An Endogenous Foamy Virus in the Aye-Aye (Daubentonia madagascariensis). J. Virol. July 2012a 86:14 7696-7698.

Hahn, C., L. Bachmann, et al. (2013). "Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach." Nucleic Acids Research **41**(13): e129.

Han G-Z, Worobey M (2012b) An Endogenous Foamy-like Viral Element in the Coelacanth Genome. PLoS Pathog 8(6): e1002790.

Hagelberg, E., Hofreiter, M., & Keyser, C. (2015). Ancient DNA: the first three decades. Philosophical Transactions of the Royal Society B: Biological Sciences, *370*(1660), 20130371.

Hayward, A., M. Grabherr, et al. (2013). "Broad-scale phylogenomics provides insights into retrovirus–host evolution." Proceedings of the National Academy of Sciences **110**(50): 20146-20151.

Hayward, A., C. K. Cornwallis, et al. (2015). "Pan-vertebrate comparative genomics unmasks retrovirus macroevolution." Proceedings of the National Academy of Sciences **112**(2): 464-469.

Hayward, W. S., B. G. Neel, et al. (1981). "Activation of a cellular onc gene by promoter insertion in ALV-induced lymphoid leukosis." Nature **290**(5806): 475-480.

Hüser D, Gogol-Döring A, Lutter T, Weger S, Winter K, Hammer E-M, et al. (2010) Integration Preferences of Wildtype AAV-2 for Consensus Rep-Binding Sites at Numerous Loci in the Human Genome. PLoS Pathog 6(7): e1000985.

Horie, M., T. Honda, et al. (2010). "Endogenous non-retroviral RNA virus elements in mammalian genomes." Nature **463**(7277): 84-87.

Hughes, J. F. and J. M. Coffin (2001). "Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution." Nat Genet **29**(4): 487-489.

Ishida,Y., Zhao,K., Greenwood,A.D. and Roca,A.L. (2015). Proliferation of Endogenous Retroviruses in the Early Stages of a Host Germ Line Invasion. *Mol. Biol. Evol.*, 32(1): 109-120.

Jern P, Sperber GO, Blomberg J. 2006. Divergent patterns of recent retroviral integrations in the human and chimpanzee genomes: probable transmissions between other primates and chimpanzees. J. Virol. 80:1367–75.

John M Coffin, Stephen H Hughes, and Harold E Varmus. (1997) Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. ISBN-10: 0-87969-571-4.

Johnson, W. E. and J. M. Coffin (1999). "Constructing primate phylogenies from ancient retrovirus sequences." Proceedings of the National Academy of Sciences **96**(18): 10254-10260.

D. M. Knipe, Peter M. Howley, D. E. Griffin, (Hrsg.): *Fields Virology.* 5. Auflage, Lippincott Williams & Wilkins, Philadelphia 2007, ISBN 978-0-7817-6060-7.

Katzourakis A, Gifford RJ. (2010) Endogenous viral elements in animal genomes. PLoS Genet. 6:e1001191.

Katzourakis A, Gifford RJ, Tristem M, Gilbert MTP, Pybus OG. *(*2009) Macroevolution of complex retroviruses. Science.325:1512.

Katzourakis A**,** Aiewsakun P**,** Jia H**,** Wolfe N**,** et al. **(**2014) Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. Retrovirology. **11**:61.

Kustikova, O., U. Modlich, et al. (2009). Retroviral Insertion Site Analysis in Dominant Haematopoietic Clones. Genetic Modification of Hematopoietic Stem Cells. C. Baum, Humana Press. **506:** 373-390.

Kvaratskhelia, M., A. Sharma, et al. (2014). Molecular mechanisms of retroviral integration site selection. Nucleic Acids Research.

LaFave, M. C., Varshney, G. K., Gildea, D. E., Wolfsberg, T. G., Baxevanis, A. D., & Burgess, S. M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Research, *42*(7), 4257–4269.

Lander et al (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lewinski MK, Yamashita M, Emerman M, Ciuffi A, Marshall H, Crawford G, et al. (2006) Retroviral DNA Integration: Viral and Cellular Determinants of Target-Site Selection. PLoS Pathog 2(6): e60.

Linial, M. L. (1999). "Foamy Viruses Are Unconventional Retroviruses." Journal of Virology **73**(3): 1747-1755.

Martin-Serrano, J. and S. J. D. Neil (2011). "Host factors involved in retroviral budding and release." Nat Rev Micro **9**(7): 519-531.

Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. PLoS ONE 5(11): e14004.

Marshall H., Ronen K., Berry C., Llano M., Sutherland H., Saenz D., Bickmore W., Poeschla E., Bushman F. Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. PLoS One. 2007;2:e1340. doi: 10.1371/journal.pone.0001340.

Martins, H. and P. Villesen (2011). Improved Integration Time Estimation of Endogenous Retroviruses with Phylogenetic Data. PLoS ONE. **6**(3): e14745.

Magiorkinis, G., R. J. Gifford, et al. (2012). Env-less endogenous retroviruses are genomic superspreaders. Proceedings of the National Academy of Sciences**. 109**(19): 7385-7390.

McDougall, A. S., A. Terry, T. Tzavaras, C. Cheney, J. Rojko, and J. C. Neil. 1994. Defective endogenous proviruses are expressed in feline lymphoid cells: evidence for a role in natural resistance to subgroup B feline leukemia viruses. Journal of Virology. 68(4): 2151–2160.

Mi, S., X. Lee, et al. (2000). "Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis." Nature **403**(6771): 785-789.

Murphy F.A., Fauquet C.M., Bishop D.H.L., Ghabrial S.A., Jarvis A.W., Martelli G.P., Mayo M.A., Summers M.D. 1994 Virus taxonomy: The classification and nomenclature of viruses, Retroviridae Springer-Verlag, Vienna.

Narezkina A., Taganov K.D., Litwin S., Stoyanova R., Hayashi J., Seeger C., Skalka A.M., Katz R.A. Genome-wide analyses of avian sarcoma virus integration sites. J. Virol. 2004;78:11656–11663.

Ochman, H., A. S. Gerber, et al. (1988). "Genetic applications of an inverse polymerase chain reaction." Genetics **1**Moalic, Y., Y. Blanchard, et al. (2006). "Porcine Endogenous Retrovirus Integration Sites in the Human Genome: Features in Common with Those of Murine Leukemia Virus." Journal of Virology **80**(22): 10980-10988.**20**(3): 621-623.

Polavarapu, N., Bowen, N. J., & McDonald, J. F. (2006). Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biology*, *7*(6), R51.

Pontius JU, *Mullikin JC, Smith DR,* et al*: Initial sequence and comparative analysis of the cat genome. Genome Res 17:1675,* 2007.

De Ravin, S. S., L. Su, et al. (2014). "Enhancers Are Major Targets for Murine Leukemia Virus Vector Integration." Journal of Virology **88**(8): 4504-4513.

Rethwilm, A., & Bodem, J. (2013). Evolution of Foamy Viruses: The Most Ancient of All Retroviruses. *Viruses*, *5*(10), 2349–2374.

Schmidt, M., K. Schwarzwaelder, et al. (2007). "High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR)." Nat Meth **4**(12): 1051-1057.

Shih, C. K., Rose, J. M., Hansen, G. L., Wu, J. C., Bacolla, A., & Griffin, J. A. (1991). Chimeric human immunodeficiency virus type 1/type 2 reverse transcriptases display reversed sensitivity to nonnucleoside analog inhibitors. Proceedings of the National Academy of Sciences of the United States of America, *88*(21), 9878–9882.

Schröder, A. R. W., P. Shinn, et al. "HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots." Cell **110**(4): 521-529.

Shojima, T., R. Yoshikawa, S. Hoshino, S. Shimode, S. Nakagawa, T. Ohata, R. Nakaoka, and T. Miyazawa. 2013. Identification of a novel subgroup of Koala retrovirus from koalas in Japanese zoos. Journal of Virology 87(17):9943–9948.

Simmons, G. S., P. R. Young, J. J. Hanger, K. Jones, D. T. W. Clarke, J. J. McKee, and J. Meers. 2012. Prevalence of koala retrovirus in geographically diverse populations in Australia. Australian Veterinary Journal 90(10): 404–409.

Stoye, J. P. (2012). "Studies of endogenous retroviruses reveal a continuing evolutionary saga." Nat Rev Micro **10**(6): 395-406.

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. 2005;86:783–787.

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

Tsangaras,K, Siracusa,MC, Nikolaidis,N, Ishida,Y, Cui,P, et al. (2014) Hybridization Capture Reveals Evolution and Conservation across the Entire Koala Retrovirus Genome. *PLoS ONE,* 9(4): e95633.

Turner G., Barbulescu M., Su M., Jensen-Seaman M.I., Kidd K.K., Lenz J. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr. Biol. 2001;11:1531–1535.

Tristem, M. (2000). "Identification and Characterization of Novel Human Endogenous Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database." Journal of Virology **74**(8): 3715-3730.

Wang Q, Jia P, Zhao Z (2013) VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. PLoS ONE 8(5): e64465.

Wang, Y., F. Liška, et al. (2010). A novel active endogenous retrovirus family contributes to genome variability in rat inbred strains. Genome Research. **20**(1): 19-27.

Welkin E. Johnson and John M. Coffin. Constructing primate phylogenies from ancient retrovirus sequences. (1999) PNAS. 96 (18) 10254-10260.

Willerslev, E., & Cooper, A. (2005). Ancient DNA. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1558), 3–16.

Xu, W., C. K. Stadler, K. Gorman, N. Jensen, D. Kim, H. Zheng, S. Tang, W. M. Switzer, G. W. Pye, and M. V. Eiden. (2013) An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proceedings of the National Academy of Sciences, USA. 110(28): 11547–11552.

Wu, X., Y. Li, et al. (2003). "Transcription Start Regions in the Human Genome Are Favored Targets for MLV Integration." Science **300**(5626): 1749-1751.

Chapter II

Hybridization Capture Reveals Evolution and Conservation across the Entire Koala Retrovirus Genome

2.1 Summary

The koala retrovirus (KoRV) is the only retrovirus known to be in the midst of invading the germ line of its host species. Hybridization capture and next generation sequencing were used on modern and museum DNA samples of koala (Phascolarctos cinereus) to examine ca. 130 years of evolution across the full KoRV genome. Overall, the entire proviral genome appeared to be conserved across time in sequence, protein structure and transcriptional binding sites. A total of 138 polymorphisms were detected, of which 72 were found in more than one individual. At every polymorphic site in the museum koalas, one of the character states matched that of modern KoRV. Among non-synonymous polymorphisms, radical substitutions involving large physiochemical differences between amino acids were elevated in env, potentially

reflecting anti-viral immune pressure or avoidance of receptor interference. Polymorphisms were not detected within two functional regions believed to affect infectivity. Host sequences flanking proviral integration sites were also captured; with few proviral loci shared among koalas. Recently described variants of KoRV, designated KoRV-B and KoRV-J, were not detected in museum samples, suggesting that these variants may be of recent origin.

2.2 Introduction

Endogenous retrovirus-like elements (ERVs) are common in the genomes of vertebrates, comprising 8% of the human genome [1]. ERVs derive from retroviruses that invaded the germ line of ancestral host organisms, becoming permanent genomic elements in the host lineage. Although most ERVs have adapted to become non-pathogenic and non-functional in their host, a role in human health and disease has been established for some ERVs [2,3]. One ERV in the human germ line has been co-opted as a functional

gene, syncytin, which is critical for normal development of the human placenta [4]. Recently, another human ERV has been found to play a critical role in the progression of Hodgkin’s lymphoma [5]. Despite their biomedical importance, the process by which ERVs invade their host germ lines has been difficult to

study, given that almost all known ERVs are many thousands or millions of years old. The only retrovirus known to be in the midst of transitioning from an exogenous to an endogenous form is the koala retrovirus

(KoRV). KoRV is currently invading the germ line of its host species, the koala (Phascolarctos cinereus), but is not found in the genomes of all koalas [3,6–8]. KoRV is ubiquitous among northern Australian koalas, but is less common in southern Australian mainland and island populations [8–10]. PCR and sequencing of KoRV env genes in museum specimens of koalas from the late 1800s revealed that KoRV was already ubiquitous among northern Australian koalas at that time [6]. While env has been examined in historical samples, little is known about the historical variability or stability of the rest of the KoRV genome or

changes in integration site diversity over time.

Two protein motifs, one in Gag and another in Env, have been associated with reduced infectivity of KoRV relative to the closely related gibbon ape leukemia virus (GALV). A CETTG motif in GALV Env is highly conserved across gammaretroviruses, while SRLPIY in GALV Gag is associated with promoting viral release [3]. Both protein motifs differ between KoRV and GALV, and these differences are believed to lower the relative infectivity of KoRV [3]. In historical samples of koalas, both motifs matched that of modern koalas, with no differences or polymorphisms detected in koala samples going back to the late 1800s [6]. The

reduced virulence of KoRV relative to GALV, and the lack of historical polymorphisms, has led to a hypothesis that the changes to these two protein domains may have both preceded and enabled the invasion of the koala germ line by KoRV.

Several laboratories have recently reported novel variants of KoRV [11,12]. One variant has been designated KoRV-B, with the originally identified KoRV labeled KoRV-A [12]. KoRV-B has greater virulence than KoRV-A, and has been isolated only from a subset of the koalas housed at the Los Angeles Zoo, and not from wild koalas. The KoRV-B long terminal repeat (LTR) U3 region includes 4 repeats of a core enhancer element, whereas KoRV-A has only one. The KoRV-B Env also has a different receptor-binding domain [12]. KoRV-B has the CETTG motif that is present in other infectious gammaretroviruses, but that has the sequence CETAG in KoRV-A. While KoRV-A uses the sodium dependent phosphate transporter membrane protein (PiT-1 or SLC20A1) as a receptor for viral entry, KoRV-B uses the thiamine transporter protein 1 (THTR1 or SLC19A2) [12]. Another recently identified variant, designated KoRV-J, also utilizes the THTR1 receptor for viral entry although KoRV-J does not have the CETTG motif of KoRV-B [11]. KoRV-J has been detected in zoo koalas [11]. Both KoRV-J and KoRV-B may be recently arisen variants, differing from KoRV-A in the LTR and env sequences, although they have not been examined in historical samples.

KoRV variants Such as KoRV-B show differences in regions beyond env and thus, it would be of interest to characterize polymorphisms, not just for env, but also across gag, pol, LTRs, and the koala genomic sequences flanking KoRV proviral loci. However, PCR based methods are labor intensive and often

unSuccessful when applied to historical samples. To examine KoRV evolution, we here applied a hybridization capture method to modern and ancient koala DNA, including multiple koala specimens in a single next generation sequencing run, in order to capture DNA sequences spanning the full length of the KoRV proviral genome. Recently developed solution hybridization capture methods allow for the specific enrichment of target sequences from genomic libraries, using PCR amplicons as ‘‘bait’’ to which target DNA hybridizes [13,14]. Even when the target sequences is divergent, both long (200–500 nt) and short (,30

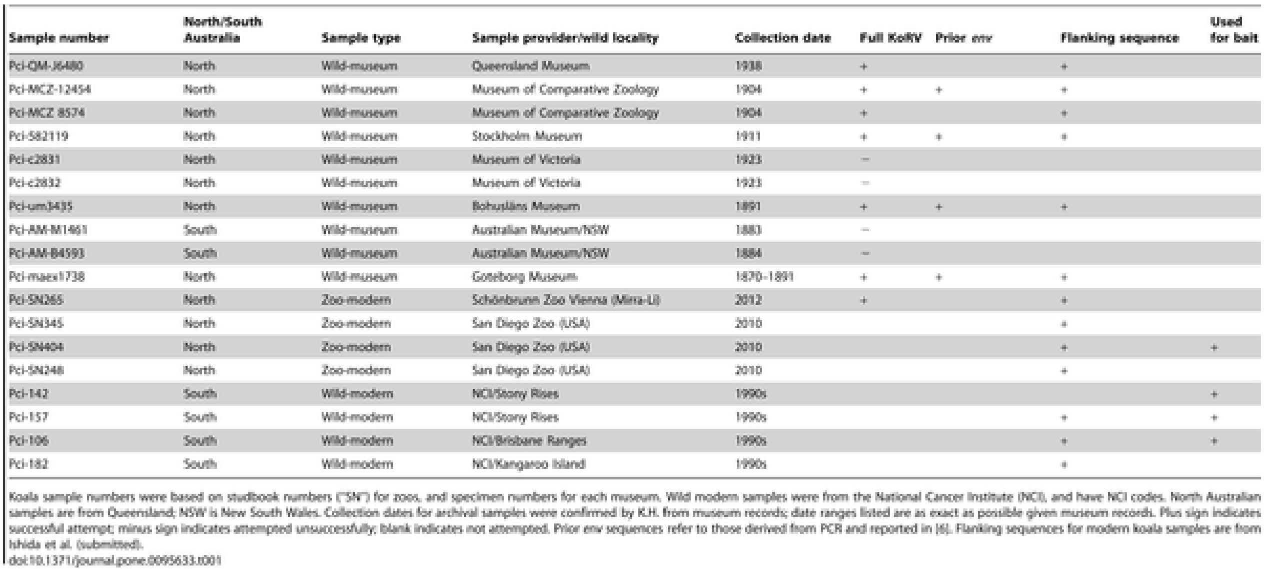
nucleotide) DNA fragments can be captured and sequenced efficiently [15], allowing use of the method with both modern and ancient DNA. This enabled us to characterize polymorphisms across the entire KoRV genome and koala genomic sequences flanking KoRV proviral loci. Polymorphisms were analyzed, and used to model potential changes to protein structure, or to identify potential changes to transcription factor binding sites in the LTRs. The flanking sequence data was used to identify integration sites common to more than one koala, identifying endogenous loci. Hybridization capture also allowed us to investigate whether KoRV-B, KoRV-J, and other recently described variants [11,12] were present in a modern deep sequenced koala, or in historical samples.

2.3 Materials and Methods

2.3.1 Koala Samples and DNA Extraction

Archival and modern samples are described in Table 1. All archival samples were extracted in a dedicated ancient DNA laboratory in the Department of Wildlife Diseases of the Leibniz Institute for Zoo and Wildlife Research under plexiglass UV hoods dedicated to DNA extraction. The ancient DNA laboratory was never used for molecular or genetic work on modern samples, and followed procedures designed to minimize the possibility of contamination, Such as wearing protective clothing during extractions to avoid contamination from the researchers. Each extraction involved approximately 250 mg of dried skin, and used the Geneclean Ancient DNA extraction kit from MP Biomedicals, USA, following the manufacturer’s protocol. Mock extractions were performed for each set of museum specimens as controls for potential contamination during the extraction process. Each DNA extract was further purified using Qiaquick spin columns (Qiagen) as described previously [16]. DNA extraction from a blood sample of modern koala Pci-SN265 (zoo koalas in North America and Europe are included in the North American regional studbook, and are here designated by studbook number, ‘‘SN’’) was performed in a separate laboratory in a different floor of the Leibniz Institute for Zoo and Wildlife Research. This extraction was performed using the Qiagen DNeasy Blood & Tissue Kit following the manufacturer’s protocol. The extracted DNA was then fragmented using a Covaris-S220 to generate 150 bp fragments.

Table 1. Koala sample information.



Sample number	North/South Australia	Sample type	Sample provider/wild locality	Collection date	Full KoRV	Prior *env*	Flanking sequence	Used for bait
Pci-QM-J6480	North	Wild-museum	Queensland Museum	1938	+		+	
Pci-MCZ-12454	North	Wild-museum	Museum of Comparative Zoology	1904	+	+	+	
Pci-MCZ 8574	North	Wild-museum	Museum of Comparative Zoology	1904	+		+	
Pci-582119	North	Wild-museum	Stockholm Museum	1911	+	+	+	
Pci-c2831	North	Wild-museum	Museum of Victoria	1923	−			
Pci-c2832	North	Wild-museum	Museum of Victoria	1923	−			
Pci-um3435	North	Wild-museum	Bohusläns Museum	1891	+	+	+	
Pci-AM-M1461	South	Wild-museum	Australian Museum/NSW	1883	−			
Pci-AM-B4593	South	Wild-museum	Australian Museum/NSW	1884	−			
Pci-maex1738	North	Wild-museum	Goteborg Museum	1870–1891	+	+	+	
Pci-SN265	North	Zoo-modern	Schönbrunn Zoo Vienna (Mirra-Li)	2012	+		+	
Pci-SN345	North	Zoo-modern	San Diego Zoo (USA)	2010			+	
Pci-SN404	North	Zoo-modern	San Diego Zoo (USA)	2010			+	+
Pci-SN248	North	Zoo-modern	San Diego Zoo (USA)	2010			+	
Pci-142	South	Wild-modern	NCI/Stony Rises	1990s				+
Pci-157	South	Wild-modern	NCI/Stony Rises	1990s			+	+
Pci-106	South	Wild-modern	NCI/Brisbane Ranges	1990s			+	+
Pci-182	South	Wild-modern	NCI/Kangaroo Island	1990s			+	

Koala sample numbers were based on studbook numbers ("SN") for zoos, and specimen numbers for each museum. Wild modern samples were from the National Cancer Institute (NCI), and have NCI codes. North Australian samples are from Queensland; NSW is New South Wales. Collection dates for archival samples were confirmed by K.H. from museum records; date ranges listed are as exact as possible given museum records. Plus sign indicates successful attempt; minus sign indicates attempted unsuccessfully; blank indicates not attempted. Prior env sequences refer to those derived from PCR and reported in [6]. Flanking sequences for modern koala samples are from Ishida et al. (submitted).
doi:10.1371/journal.pone.0095633.t001

Blood samples of San Diego Zoo koalas were collected during routine physical exams and genomic DNA was isolated from buffy coat using the Qiagen DNeasy Blood & Tissue Kit following the manufacturer’s protocol. DNA from blood samples of wild koalas had been extracted using a phenol-chloroform method. These samples were used to generate baits.

2.3.2 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 01-01-2013. Work involving other modern koala samples was conducted at the University of Illinois at Urbana-Champaign (UI°C), under IAC°C approval number 12040.

2.3.3 Polymerase Chain Reaction

All museum specimen were initially screened for a KoRV pol fragment by PCR (Table 1) performed in a volume of 34 ml using 5.5 ml of extract, 10 nm of primers, 0.5 U Platinum HiFi supermix (Invitrogen), 1ml of bovine serum albumin (Fermentas), and 1 ml of primers P1aF 5'-TTGGAGGAGGAATACCGATTACAC-3' with P1aR 5'-GCCAGTCCCATACCTGCCTT-3' [8]. Cycling conditions were: 94°C for 4 min; 60 cycles at 94°C for 30 s, 55°C for 30 s, 72°C for 30 s; and 72°C for 10 min, with the samples then held at 4°C [17]. The high cycle number (60) PCR was only used for screening museum koala samples for the presence of

KoRV and not for polymorphism analyses. The modern sample was screened by PCR amplification performed in a volume of 34 ml using 1 ml (26.7 ng/ml) of extract, 10 nM of each primer, 0.5 U of Platinum HiFi supermix (Invitrogen). Cycling condition were: 94°C for 4 min; 35 cycles at 94°C for 30 s, 55°C for 30 s, 72°C for 30 s; and 72°C for 5 min, with the samples then held at 4°C. PCR products were visualized on a 3% gel. All gels used GelRed nucleic acid gel stain by Biotium. PCR products were purified using the nucleoSpin Gel and PCR Clean up kit (Macharey-Nagel). PCR products were commercially sequenced

by the Sanger method using the forward and reverse PCR primers (StarSeq, Germany). Primers used in this study are listed in Table S1. The Sanger sequences were not included in the hybridization capture alignments but were only used to establish the presence of KoRV in museum and modern samples.

2.3.4 Illumina Library Preparation

Aliquots from each DNA extract were used in generating Illumina libraries. Archival extract libraries were generated in the ancient DNA facility in a library-dedicated plexiglass PCR UV hood, while the modern koala library was generated in a modern DNA laboratory in a different part of the Institute. Libraries were

generated as described in Mayer et al. [18]. 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 any contamination introduced during the experiment. Indexes were added by PCR using Amplitaq

Gold DNA polymerase (Applied Biosystems [ABI]) in 100 ml reactions. Cycling condition were: 94°C for 5 min; 10 cycles at 94°C for 30 s, 55°C for 30 s, 72°C for 30 s; and 72°C for 5 min; the samples were then held at 4°C. After indexing, the samples would

effectively be at little or no risk from cross contamination either from the other libraries or from laboratory DNA. Quantitative PCR (qPCR) was performed after index PCR with a standard that was developed using 100 bp PCR product with Illumina primer binding sites ligated at the 59 and 39 ends as described in Mayer et al. [18]. The qPCR standard curve was obtained using a series dilution of the standard. The assay was performed in a Stratagene MxPro 3000p qPCR system using Brilliant III Ultra-Fast SyBr Green qPCR master mix (Agilent) with Illumina bridge primers P5 and P7 [18] to determine the number of molecules in each sample. Additional amplification was followed using Herculase II DNA polymerase (Agilent) with P5 and P7 Illumina library outer primers with the same cycling conditions. DNA products were purified using Minelute columns (Qiagen) after each amplification step. Final quantification was performed on an Agilent 2200 tape station D1K tape.

2.3.5 Primer Design and Preparation of Baits

PCR products used as ''bait'' for capturing sequences from the Illumina libraries were generated at the University of Illinois to limit the amount of koala and KoRV amplicons present in the laboratories in Berlin. DNA of one northern koala, Pci-SN404 (see above) and three southern koalas (PCI-157 and PCI-142 from the Stony Rises and PCI-106 from the Brisbane Ranges of southern Australia) were used in preparing the bait. Primers were newly designed to cover the complete KoRV genome outside the envelope region. For the envelope region, previously designed primers were used [6] but with primer combinations that would yield amplicon sizes of approximately 500 bp. For the other KoRV regions, novel primers based on the published KoRV sequence (GenBank: AF151794) [7] were designed using Primer3 (http://fokker.wi.mit.edu/primer3/ input.htm) [19] to yield amplicons of approximately 500 bp. The KoRV genome was amplified in thirty-eight 500 bp overlapping products using the primers shown in Table S1. The PCR mix consisted of 1 X PCR Buffer II (ABI), 1.5 mM $MgCl_2$ (ABI), 0.4 mM of final concentration of each primer, 200 mM of each dNTP (ABI), with 0.04 units/ml final concentration of AmpliTaq Gold DNA Polymerase (ABI). The PCR algorithm

consisted of an initial 95°C for 10 min; with cycles of 15 sec at 95°C; followed by 30 sec at 60°C, 58°C, 56°C, 54°C, 52°C (2 cycles at each temperature) or 50°C (last 30 cycles); and 1 min at 72°C; with a final extension of 7 min at 72°C. An aliquot of each PCR product was visualized on a 1% agarose gel with ethidium bromide. PCR products were enzyme-purified [20] and Sanger-sequenced to verify that the target region had been amplified. The PCR products were purified using Qiaquick columns (Qiagen) and then quantified using a NanoDrop ND-1000 (Thermo-Scientific). KoRV amplicons were then blunt-ended, ligated to a biotin adapter, and immobilized on streptavidin magnetic beads in equimolar amounts of 1.3 mg as described previously [14].

2.3.6 Hybridization Capture

Mixtures of blocking agent, blocking oligos, and indexed koala libraries were heated to 95°C to separate the DNA strands [14]. One aliquot from each index library was mixed with streptavidin beads bound with biotinylated KoRV PCR products. Samples were incubated for 48 hours at 65°C under rotation in a Labnet

mini incubator. After 48 hours the beads were washed and the hybridized libraries eluted by heating. The DNA concentration was measured by quantitative PCR (qPCR), and the eluted libraries were further amplified accordingly using P5 and P7 Illumina outer primers. The products were then pooled at equimolar concentrations for paired-end sequencing on an Illumina MiSeq platform at the National High-Throughput

DNA Sequencing Center, University of Copenhagen.

2.3.7 Sequence Assembly, Identification of Polymorphisms and Integration Site Analysis

Sequences were separated based on their index sequence at the National High-Throughput DNA Sequencing Center, University of Copenhagen, Denmark. The programs cutadapt v1.2 and trimmomatic [21,22], respectively, were used to remove adaptor sequences and poorly sequenced reads. After trimming, reads that

were shorter than 20 bp were excluded from further analyses. Reads were mapped to the KoRV full genome reference sequence (NCBI: AF151794) using BWA version 0.6.2 [23] with default parameters. The resulting SAM files were further processed with samtools [24] and picard (http://picard.sourceforge.net) for sorting and removal of clonality, respectively. The Perl script mapDamage was run on the museum data using the default settings to determine the percentage of DNA damage present, before SNP calling [25]. Variant call analysis was performed using VarScan 2.2.3 with the following settings -min-coverage 8, -minvar-freq 0.01, and -p-value 5e-02 [26]. The resulting variants were further curated using

Geneious 6.0.4 for visualization. Negative control reads were also compared to the reference KoRV sequence. The 59 and 39 LTRs were distinguished from each other by examining sequences adjacent to the LTR sequence for genomic flank sequences or for KoRV sequence (gag leader or env). The 5' LTR is preceded by a koala genomic flank and followed by a KoRV gag leader, while the 3' LTR is preceded by KoRV env and followed by a koala genomic flank. Where possible, LTR sequences that also included a KoRV non-LTR sequence or a koala genomic flanking sequence were used to distinguish between 5' and 3' LTR polymorphisms (Figure 1). Consensus sequences generated were deposited in GenBank (accession numbers KF786280–KF786286). Illumina reads mapping to KoRV for each koala were deposited in the NIH Short Read Archive (SRP03960187947).

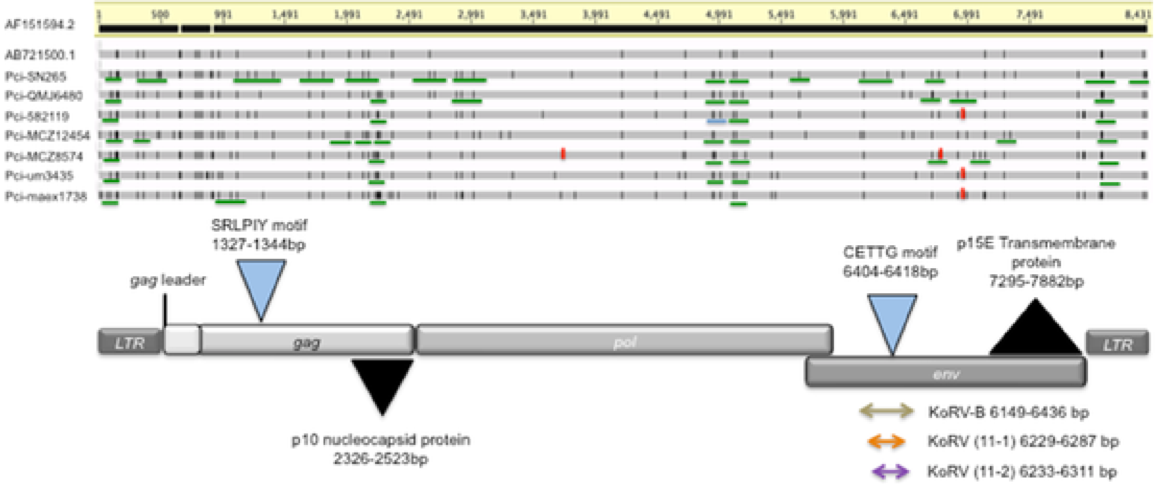


Figure 1. Alignment of modern and museum koala retrovirus sequences, showing positions of proviral genes and proteins. Upper Panel: Character states matching the reference sequence (AF151794) are indicated in light grey, while mismatches (position 312) or polymorphisms (all other positions) are shown as black hatch marks. The infectious clone KV522 (AB721500) is the first sequence below the reference. The aligned sequences all display open reading frames for viral gag, pol, and env regions, except that polymorphisms at three positions in the museum samples code for a stop codon that would disrupt an open reading frame; these are indicated by red hatch marks. Green lines represent olymorphisms that could be placed in phase in overlapping sequence reads. Lower Panel: The coded proteins are indicated, following the divisions proposed by Hanger et al. (2000) relative to the polymorphism alignment. The positions of the SRLPIY domain potentially involved in viral infectivity of the GAG protein and p10 domain (Gag assembly and nuclear export signal, respectively) are indicated. Likewise, the Env motif CETTG, and p15E transmembrane envelope protein are indicated. Regions known to be divergent in Japanese isolates, KoRV-B and KoRV- C, and KoRV-D [12,40], are indicated by orange, purple, and green arrows, respectively.

Integration sites were identified in sequence reads that contained 59 or 39 LTR sequences extending into non-KoRV sequences. To examine whether proviral integration sites identified in the ancient samples were present among modern koalas, koala genomic sequences flanking the integration sites found by hybridization were queried against sequences flanking the integration sites of six modern koalas (three northern and three southern koalas, Table 1) that had been generated using a different method (Ishida et al. manuscript in preparation). Integration site sequences were also queried against a koala (Pci-SN404) whole

genome sequence generated using one-sixteenth of a PicoTiter-Plate of 454 GS-FLX+ Technology (Roche Applied Science) following standard protocols.

2.3.8 Statistical Analyses and Tests of Selection

For non-synonymous polymorphisms, a ‘‘radical’’ change was defined as a mutation that produces a negative score in both BLOSUM62 and BLOSUM90 substitution matrices. Associations between variables were examined using a 262 contingency table, testing for significance using Fisher’s exact test implemented in

GraphPad (graphpad.com/quickcalcs/contingency1.cfm). The number of synonymous and non-synonymous substitutions was determined, and the Nei-Gojobori method [27] was used to determine the proportion of synonymous substitutions per synonymous sites and the proportion of non-synonymous substitutions per non-synonymous sites. MEGA, version 5.2 [28] was used to estimate Tajima’s D, and to implement the codon-based Z-test for selection and the codon-based Fisher’s exact test of selection. These were determined for the concatenated KoRV codons of gag, pol, and env, and for each of the three separately. Bonferroni correction for multiple hypothesis testing divided a p value of 0.05 by the number of hypotheses tested.

The dN/dS ratio provides an indicator of the selective pressures that acted upon a gene, with low values indicating purifying selection and increases in values indicating relaxation of constraint or positive selection. To account for the different phase of polymorphisms at the same site we generated an individual

sequence for each different phase of a polymorphism and analyzed all of the individual sequences. For the modern koala, available sequences were long enough for phase to be determined for many (but not all) polymorphisms (Figure 1, positions underlined in green). For historical samples, sequence lengths were short, and the phase of polymorphisms could only rarely be determined. To test for this signature of selection in this dataset, we calculated dN/dS using two different approaches: the GA-Branch and FUBAR methods [29]. In the first case estimates were obtained using a fixed tree topology generated by the Neighbor-joining method. The nucleotide model was specified as GTR; otherwise, the default GA-Branch configuration was used. This dataset, which compares all identified polymorphisms (even if they are not present in the consensus sequence) against the modern sequence, was also analyzed using the Z-test for selection and Tajima's D.

2.3.9 Identification of Protein Domains and Functional Residues, and Protein Modeling

The corresponding amino acid sequences were subjected to domain identification analysis using the Conserved Domains Database (CDD) from NCBI. We also examined whether any of the observed polymorphisms alter amino acid residues of known function using the Conserved Features/Sites option of the CDD database.

To examine the structural characteristics of KoRV variants, we predicted their three-dimensional structures using the SWISSMODEL server [30]. Only models with high statistical support (high reliability score as defined by QMEAN4 values) [31] were considered for further analyses. Using this strategy we were able to reliably model several regions corresponding to different domains of all three viral polypeptides (Gag, Pol, and Env). Pairwise structural alignments and structural superimposition were performed using the DaliLite server [32]. Models and Figures were drawn using Pymol (DeLano Scientific).

2.3.10 Transcription Binding Factor Site Analysis

The long terminal repeat polymorphic sequences for each koala were analyzed for putative transcription binding domains using MatInspectror software (Genomatix, Munich). The default core similarity and matrix similarity greater than 0.8 were employed as the selection criteria.

2.4 Results

2.4.1 Hybridization Capture and Sequencing of KoRV

DNA was extracted in an ancient DNA dedicated facility from 10 museum skins from southern (n =2) or northern (n=8) Australian koalas, which had been collected as long as 130 years ago (Table 1). In separate facilities, modern DNA was extracted from blood samples of zoo and free-ranging wild koalas (Table 1).

Illumina libraries were prepared from all museum koala DNA extracts (in an ancient DNA facility), and from one sample of modern DNA from an adult northern koala 14 years old, Pci-SN265 (‘‘Mirra-Li’’, studbook number 265 from the Zoo Vienna, Austria). In order to process all samples in a single next-generation sequencing run, each library was tagged with a distinct index sequence. Baits for hybridization were generated covering the entire KoRV genome, which was amplified in thirty-eight fragments, each ca. 500 bp, from four koalas representing northern (Pci-SN404) and southern (Pci-SN106, 142, 157) koala KoRV diversity (Table 1). Equimolar amounts of index samples were pooled and applied to KoRV baits bound to streptavidin beads for in-solution hybridization capture. The enriched koala libraries were then sequenced using an Illumina MiSeq. After sequencing, a bioinformatics routine used the distinct index sequence tags to separate sequences by individual. Sequences were screened for quality and reliability before being aligned to KoRV reference genomes (GenBank accession number AF151794, AB721500).

Full coverage of the KoRV genome was obtained from six of the northern Australian museum specimens and for Pci-SN265 (Figure 1); museum specimens of two additional northern and two southern koalas were not successful. Among the six successful historical samples, KoRV-specific sequences represented 2.5% to 41% of the total number of reads, comparable to enrichment rates previously reported for ancient DNA. Negative controls demonstrated only sporadic matches to KoRV (Figure 2). Such sporadic

reads are observed in hybridization capture experiments [33] and may reflect index sequence errors (misassignment) or PCR jumping causing exchange between sample index and control [34]. However, the profile of coverage was randomly dispersed and the number of reads marginal in the negative control.

Coverage was consistently far higher at every position for the modern koala (Pci-SN265) than for any of the museum specimens (Figure 2). The historical samples yielded only 20–80% of the coverage of the KoRV genome obtained for the modern sample. Among the museum koalas, the earliest collected sample (Pcimaex1738) had the poorest coverage whereas the most recently collected sample (Pci-QMJ6480) had the highest relative coverage. There was otherwise no obvious correlation between the number of reads obtained and the year the sample was collected.

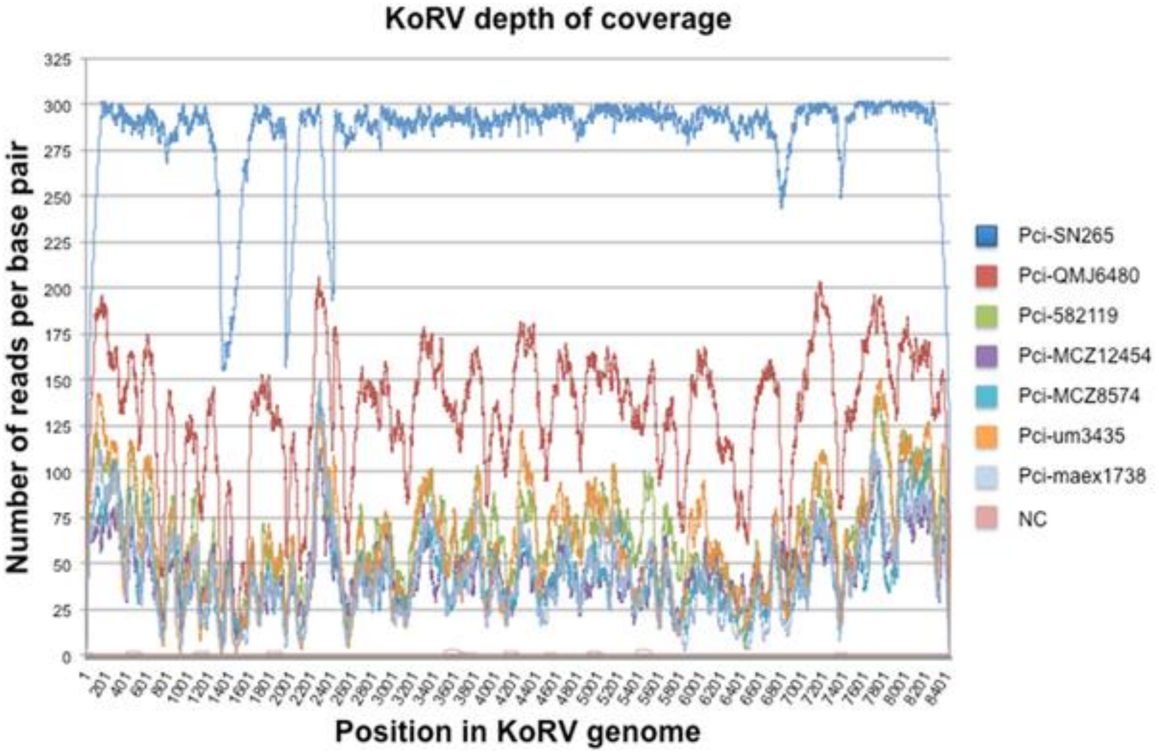


Figure 2. Hybridization capture sequence coverage across the KoRV genome for modern and museum koala samples. The sequence coverage is shown for each nucleotide position numbered as in the KoRV reference genome (AB721500). Results are shown for 1 modern (Pci-SN265) and 6 museum koala samples. Mapping of results for a negative control (NC) are also shown. Each sample is color-coded.

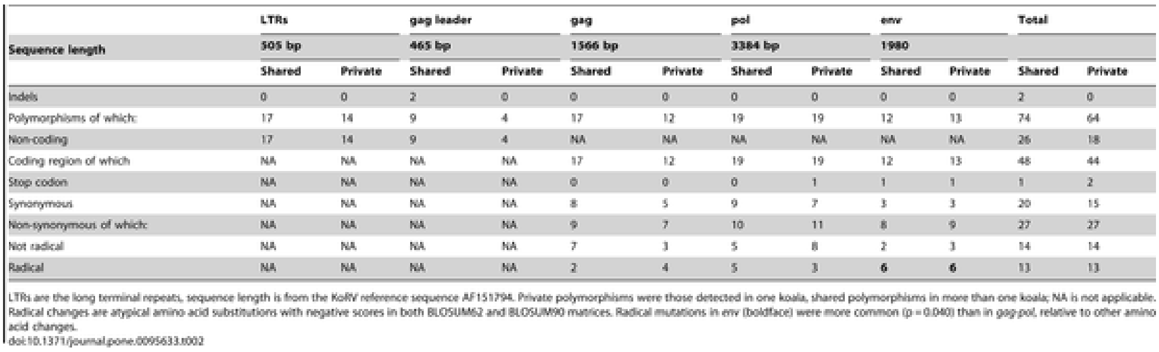
2.4.2 KoRV Polymorphisms

For the museum samples, the average read length was ca. 90 bp. This is similar to read lengths reported previously for DNA from archival specimens, which may be degraded as a result of environmental, bacterial, and enzymatic damage [35–37], and was shorter than the ca. 135 bp read length for modern sample Pci-

SN265. Although similar analyses were conducted on the historical and on the modern koala sequences, prior to assembly the museum specimen datasets were processed using the map-Damage Perl script [25], to account for DNA damage present in ancient DNA. The mapDamage results identified the expected nucleotide misincorporation patterns of cytosine to thymine and guanine to adenine on the 5' and 3' end termini, respectively (not shown). However, damage occurred only at a very low frequency of 0.02 to 0.08%, indicating that the damage present would have negligible effects on polymorphism scoring or other analyses. After assembling the reads to the KoRV reference sequence AF151794, polymorphisms were scored if they occurred at a position in 8% or more of the reads for an individual koala [38]. For the env gene, four of 20 env polymorphisms that had been previously detected by PCR from museum samples were also found in the current dataset [6]. Of the remaining 16, seven could be identified but were not present above the cutoff employed when identifying polymorphisms by the current study. The other nine could not be identified from the data, likely due to insufficient coverage in some koalas for those regions of env (Figure 2). Fourteen novel polymorphic sites in the env region were identified by hybridization capture that had not been identified in the same museum koalas when previously examined by PCR.

At position 312 a fixed difference as opposed to a polymorphism was present in all koalas relative to the reference AF151794 (Figure 1). Across the modern and archival koalas, a total of 138 KoRV polymorphisms were detected. At each of these polymorphisms, one of the character states matched that of the KoRV reference sequence AF151794. Considering only the character states that differed from the reference, seventy-one of the polymorphic sites were detected in two or more koalas (shared alleles) and sixty-seven were detected only in one koala (private alleles) (Table 2, Table S2). Of 92 polymorphisms in the coding regions, 3 would result in stop codons, of which one was shared across individuals (Figure 1, Table 2 KoRV). Of the remaining coding region polymorphisms, 35 were synonymous and 54 were non-synonymous (Table 2, Table S2).

Table 2. Types of KoRV polymorphisms detected across 6 museum specimens of koalas.



	LTRs		gag leader		gag		pol		env		Total	
Sequence length	505 bp		465 bp		1566 bp		3384 bp		1980			
	Shared	Private	Shared	Private	Shared	Private	Shared	Private	Shared	Private	Shared	Private
Indels	0	0	2	0	0	0	0	0	0	0	2	0
Polymorphisms of which:	17	14	9	4	17	12	19	19	12	13	74	64
Non-coding	17	14	9	4	NA	NA	NA	NA	NA	NA	26	18
Coding region of which	NA	NA	NA	NA	17	12	19	19	12	13	48	44
Stop codon	NA	NA	NA	NA	0	0	0	1	1	1	1	2
Synonymous	NA	NA	NA	NA	8	5	9	7	3	3	20	15
Non-synonymous of which:	NA	NA	NA	NA	9	7	10	11	8	9	27	27
Not radical	NA	NA	NA	NA	7	3	5	8	2	3	14	14
Radical	NA	NA	NA	NA	2	4	5	3	**6**	**6**	13	13

LTRs are the long terminal repeats, sequence length is from the KoRV reference sequence AF151794. Private polymorphisms were those detected in one koala, shared polymorphisms in more than one koala; NA is not applicable. Radical changes are atypical amino acid substitutions with negative scores in both BLOSUM62 and BLOSUM90 matrices. Radical mutations in env (boldface) were more common (p = 0.040) than in gag-pol, relative to other amino acid changes.
doi:10.1371/journal.pone.0095633.t002

Functional regions in the viral sequence reported to reduce the infectivity of KoRV when compared to that of GALV were also examined [3]. The CETAG motif in KoRV (CETTG in other gammaretroviruses) is believed to be responsible for viral fusion activity while the gag L-domain is believed to affect the release of mature virus. Across KoRVs in the newly sequenced koalas, there were no polymorphic sites in either of these regions (Figure 1). The immunosuppressive domain of the p15E transmembrane protein of retroviral Env exhibited only a single polymorphism, present in Pci-SN265 and Pci-MCZ12454 (Table S2).

In the museum specimens, multiple non-synonymous polymorphisms were detected in the nucleocapsid protein region (p10) of the gag gene. Using the Conserved Features/Sites function of the CDD database we also determined that none of the amino acid residues of known or inferred function, e.g., DNA binding site of the reverse transcriptase domain in POL or the homotrimer interface in ENV, are polymorphic (data not shown).

Amino acid substitution matrices have been generated by comparing large numbers of proteins to identify non-synonymous mutations that are only rarely observed empirically. These rare amino acid substitutions, termed ‘‘radical’’, typically involve major physiochemical differences between the two amino acids. We defined a radical change as a mutation that produces a negative score in both BLOSUM62 and BLOSUM90 substitution matrices. Among the non-synonymous polymorphisms observed in the koala, 48% (26/54) of substitutions were defined as radical. The proportion of radical non-synonymous mutations appeared to be higher in env than in gag or pol (Table 2), and this difference was confirmed as significant using Fisher’s exact test (p= 0.0397) comparing radical vs. non-radical non-synonymous substitutions in env to those in gag-pol. This suggested that selective constraints on env may differ from those affecting the other two KoRV coding regions. Across the three coding regions, no other pattern suggestive of an association across variables was evident in the dataset for private vs. shared polymorphisms, non-synonymous vs.

synonymous polymorphisms, or radical vs. non-radical amino acid changes (Table 2).

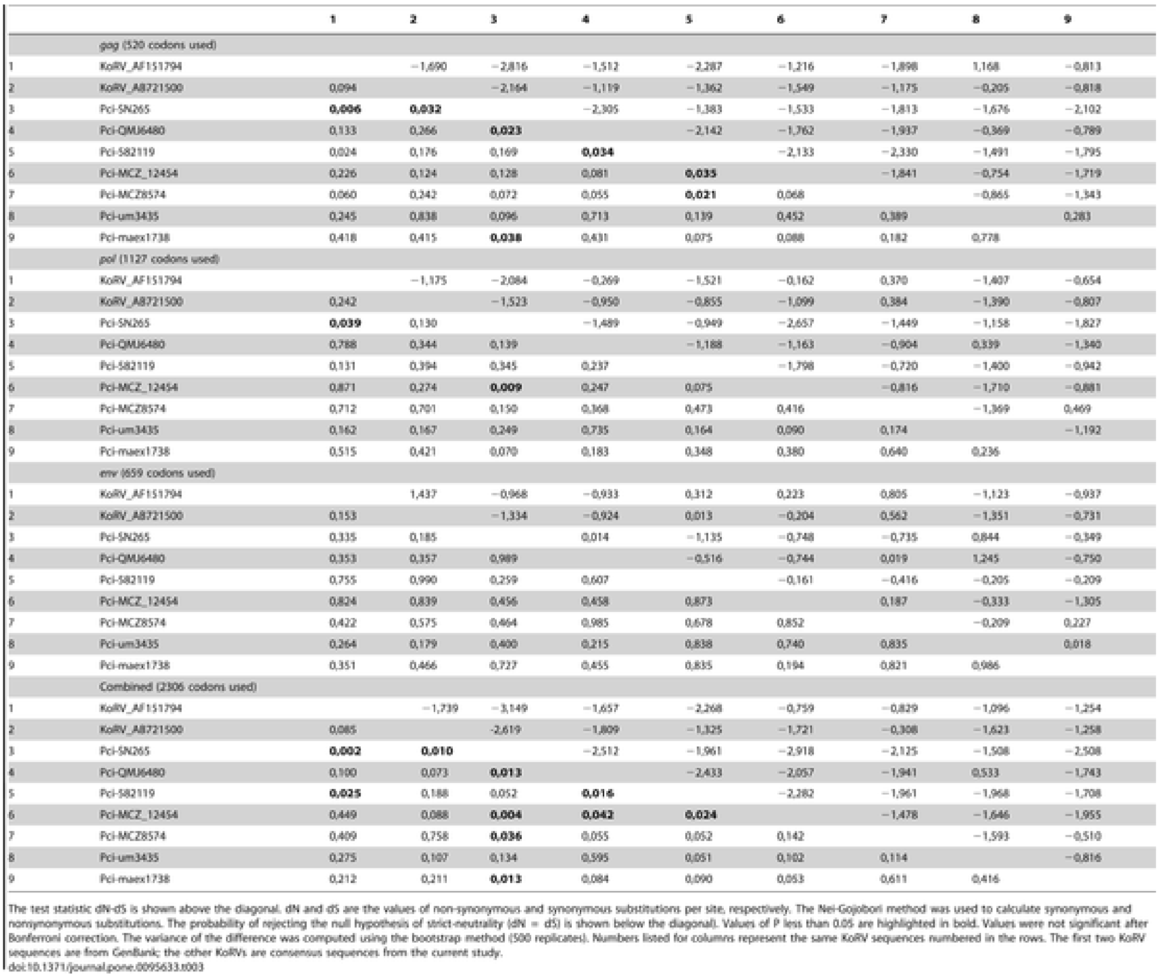
The selective pressure variation among all branches of the KoRV tree estimated by the GA-Branch method suggested that several branches in gag (more than 70%) and fewer in pol and env (60 and 17%, respectively) are under purifying selection (not shown). FUBAR implemented in HyPhy also suggested that only a few codons deviate neutrality (not shown). Similarly, the total distance estimates of dN/dS using the Nei-Gojobori method suggested stronger purifying selection in gag than in pol and env (Table S3). The same trend was observed by the Z-tests for selection and the Tajima’s test of neutrality, with gag showing multiple significantly negative dN-dS values and the lowest negative values, respectively (Table S4 and Table S5).

2.4.3 Comparisons of KoRV Consensus Sequences

The nucleotide consensus sequence (majority character state at every position in an alignment of sequences) was generated for each of the seven Successful KoRV-positive koalas. These were compared to the first reported KoRV sequence (AF151794) [7] and to the infectious clone KV522 (AB721500) [39], which themselves are 0.5% divergent, generating an alignment of 9 sequences. Each of the newly generated consensus sequences was more similar to the infectious clone KV522 (99.2–99.5% similarity) versus for AF151794 (99.0–99.2% similarity). All of the koala retroviral consensus sequences from the current study included a 6-bp insertion in the non-coding gag leader region, position 651, which is also present in KV522 (Figure 1, Table S2). The six archival sample consensus sequences also shared a 3-bp insertion at position 905 in the gag leader region (Table S2) also present in KV522 (AB721500) [39]. The 3 bp insertion could be found in the modern koala (Pci-SN265) as a minority sequence, thus the consensus for this animal lacked the 3 bp insertion (Figure 1). Thus, in contrast to the museum samples, the modern koala had an underrepresentation (16.8%) of the 3 bp insertion variant. In addition, the deletion itself is polymorphic representing 1–3 bp deletions though the 3 bp deletion is the most common and therefore represented in the consensus sequence generated.

The alignment of nine sequences was also examined for signatures of selection (or neutrality), for gag, pol, and env, and for all three codon sets concatenated. The Nei-Gojobori method was used to estimate synonymous and non-synonymous mutation rates for each pair of sequences. Codon-based Fisher’s exact tests of selection found no evidence of positive selection in any of the pairwise comparisons for any of the coding regions (not shown). Codon based Z-tests of selection (Table 3) suggested that among the coding regions purifying selection may be more pronounced in gag, with significant purifying selection detected in eight of the pairwise comparisons (although these would not be significant after Bonferroni correction). The pol coding region appeared to be under weaker purifying selection, with negative values significant (before Bonferroni correction) for only 2 comparisons, while env comparisons yielded both positive and negative estimates consistent with neutrality (none significant). Tajima’s D was calculated using the same nine KoRV sequences (Table 4), for each coding region separately or all three combined. A negative value would be consistent with purifying selection. However, although gag had the most negative value, none of the values for Tajima’s D were extreme, thus there were no significant deviations from neutrality. The consensus sequence and total polymorphism data yielded consistent results with respect to selective pressures on KoRV.

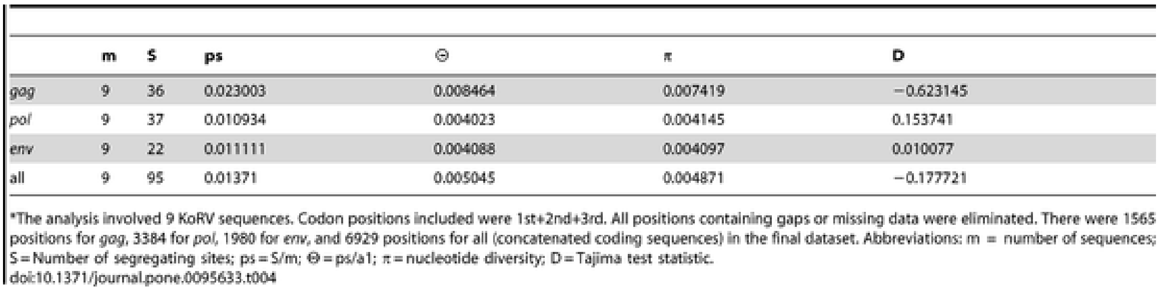
Table 3. Codon based Z tests of selection.



		1	2	3	4	5	6	7	8	9
	gag (520 codons used)									
1	KoRV_AF151794		−1,690	−2,816	−1,512	−2,287	−1,216	−1,898	1,168	−0,813
2	KoRV_AB721500	0,094		−2,164	−1,119	−1,362	−1,549	−1,175	−0,205	−0,818
3	Pci-SN265	**0,006**	**0,032**		−2,305	−1,383	−1,533	−1,813	−1,676	−2,102
4	Pci-QMJ6480	0,133	0,266	**0,023**		−2,142	−1,762	−1,937	−0,369	−0,789
5	Pci-582119	0,024	0,176	0,169	**0,034**		−2,133	−2,330	−1,491	−1,795
6	Pci-MCZ_12454	0,226	0,124	0,128	0,081	**0,035**		−1,841	−0,754	−1,719
7	Pci-MCZ8574	0,060	0,242	0,072	0,055	**0,021**	0,068		−0,865	−1,343
8	Pci-um3435	0,245	0,838	0,096	0,713	0,139	0,452	0,389		0,283
9	Pci-maex1738	0,418	0,415	**0,038**	0,431	0,075	0,088	0,182	0,778	
	pol (1127 codons used)									
1	KoRV_AF151794		−1,175	−2,084	−0,269	−1,521	−0,162	0,370	−1,407	−0,654
2	KoRV_AB721500	0,242		−1,523	−0,950	−0,855	−1,099	0,384	−1,390	−0,807
3	Pci-SN265	**0,039**	0,130		−1,489	−0,949	−2,657	−1,449	−1,158	−1,827
4	Pci-QMJ6480	0,788	0,344	0,139		−1,188	−1,163	−0,904	0,339	−1,340
5	Pci-582119	0,131	0,394	0,345	0,237		−1,798	−0,720	−1,400	−0,942
6	Pci-MCZ_12454	0,871	0,274	**0,009**	0,247	0,075		−0,816	−1,710	−0,881
7	Pci-MCZ8574	0,712	0,701	0,150	0,368	0,473	0,416		−1,369	0,469
8	Pci-um3435	0,162	0,167	0,249	0,735	0,164	0,090	0,174		−1,192
9	Pci-maex1738	0,515	0,421	0,070	0,183	0,348	0,380	0,640	0,236	
	env (659 codons used)									
1	KoRV_AF151794		1,437	−0,968	−0,933	0,312	0,223	0,805	−1,123	−0,937
2	KoRV_AB721500	0,153		−1,334	−0,924	0,013	−0,204	0,562	−1,351	−0,731
3	Pci-SN265	0,335	0,185		0,014	−1,135	−0,748	−0,735	0,844	−0,349
4	Pci-QMJ6480	0,353	0,357	0,989		−0,516	−0,744	0,019	1,245	−0,750
5	Pci-582119	0,755	0,990	0,259	0,607		−0,161	−0,416	−0,205	−0,209
6	Pci-MCZ_12454	0,824	0,839	0,456	0,458	0,873		0,187	−0,333	−1,305
7	Pci-MCZ8574	0,422	0,575	0,464	0,985	0,678	0,852		−0,209	0,227
8	Pci-um3435	0,264	0,179	0,400	0,215	0,838	0,740	0,835		0,018
9	Pci-maex1738	0,351	0,466	0,727	0,455	0,835	0,194	0,821	0,986	
	Combined (2306 codons used)									
1	KoRV_AF151794		−1,739	−3,149	−1,657	−2,268	−0,759	−0,829	−1,096	−1,254
2	KoRV_AB721500	0,085		-2,619	−1,809	−1,325	−1,721	−0,308	−1,623	−1,258
3	Pci-SN265	**0,002**	**0,010**		−2,512	−1,961	−2,918	−2,125	−1,508	−2,508
4	Pci-QMJ6480	0,100	0,073	**0,013**		−2,433	−2,057	−1,941	0,533	−1,743
5	Pci-582119	**0,025**	0,188	0,052	**0,016**		−2,282	−1,961	−1,968	−1,708
6	Pci-MCZ_12454	0,449	0,088	**0,004**	**0,042**	**0,024**		−1,478	−1,646	−1,955
7	Pci-MCZ8574	0,409	0,758	**0,036**	0,055	0,052	0,142		−1,593	−0,510
8	Pci-um3435	0,275	0,107	0,134	0,595	0,051	0,102	0,114		−0,816
9	Pci-maex1738	0,212	0,211	**0,013**	0,084	0,090	0,053	0,611	0,416	

The test statistic dN-dS is shown above the diagonal. dN and dS are the values of non-synonymous and synonymous substitutions per site, respectively. The Nei-Gojobori method was used to calculate synonymous and nonsynonymous substitutions. The probability of rejecting the null hypothesis of strict-neutrality (dN = dS) is shown below the diagonal). Values of P less than 0.05 are highlighted in bold. Values were not significant after Bonferroni correction. The variance of the difference was computed using the bootstrap method (500 replicates). Numbers listed for columns represent the same KoRV sequences numbered in the rows. The first two KoRV sequences are from GenBank; the other KoRVs are consensus sequences from the current study.
doi:10.1371/journal.pone.0095633.t003

Table 4. Estimates of Tajima's D*.



	m	S	ps	Θ	π	D
gag	9	36	0.023003	0.008464	0.007419	−0.623145
pol	9	37	0.010934	0.004023	0.004145	0.153741
env	9	22	0.011111	0.004088	0.004097	0.010077
all	9	95	0.01371	0.005045	0.004871	−0.177721

*The analysis involved 9 KoRV sequences. Codon positions included were 1st+2nd+3rd. All positions containing gaps or missing data were eliminated. There were 1565 positions for *gag*, 3384 for *pol*, 1980 for *env*, and 6929 positions for all (concatenated coding sequences) in the final dataset. Abbreviations: m = number of sequences; S = Number of segregating sites; ps = S/m; Θ = ps/a1; π = nucleotide diversity; D = Tajima test statistic.
doi:10.1371/journal.pone.0095633.t004

2.4.4 KoRV-B and J

A recent study of koalas from Los Angeles Zoos identified a KoRV variant, designated KoRV-B, which has greater virulence than previously characterized KoRV, and which was present in a subset of zoo koalas [12]. Hybridization capture should enrich sequences, Such as those of KoRV-B, that are somewhat divergent from the KoRV sequence used as bait. We therefore screened the novel next generation sequencing data, searching for the sequence of KoRV-B at the junction where KoRV-B diverges from the KoRV reference sequences in the env region. The divergent region of KoRV-B env was not detected in any of the ancient koalas, suggesting that KoRV-B may have evolved recently as a variant of KoRV. However in the modern koala Pci-SN265, sequences matching KoRV-B were detected for some (but not all) of the regions within the env gene that characterize KoRV-B (Figure S1). Three other variants of KoRV have recently been described among zoo koalas in Japan, tagged as clones 11-1, 11-2, and 11-4 [11,40]. Clone 11-1 has been designated KoRV-D, clone 11-2 has been designated KoRV-C, and clone 11-4 has been designated KoRV-J. The three recently identified KoRV clones differ mainly in variable region A of the env gene that is involved in retroviral

receptor determination and recognition [11,41]. Our novel sequences were screened for each of the KoRV variants reported in the Japanese zoo koalas. Sequences similar to those of KoRV-C and KoRV-D were identified in the modern koala Pci-SN265 but not in any of the museum samples (Figure S2 A and B) [40].

Sequences related to KoRV-J were not identified in any of the novel reads, whether from the modern or museum samples.

2.4.5 Potential Effects of KoRV Polymorphisms on Protein Structure

Variants present below a cutoff of 8% of relevant Illumina reads were not considered to represent confirmable polymorphisms. Those that appeared at a higher frequency than this cutoff likely represent common variants rather than mutations within a single provirus. We examined the effects of the non-synonymous polymorphisms on the protein structure of KoRV by generating three-dimensional models for Gag, Pol, and Env protein fragments. First, we sought to identify whether amino acid differences present across modern sequences of KoRV led to major structural differences. Sequences included in the comparisons of modern KoRV were the original KoRV isolate AF151794 and infectious clone KV522 (AB721500), which differ by 0.5% at the nucleotide. In these comparisons, our consensus KoRV sequence from the modern koala Pci-SN265 served as the reference (the amino acid residue in Pci-SN265 is always the first listed in each substitution). When superimposed on the structure of Pci-SN265, the structures of AF151794 and KV522 showed minor localized changes affecting the polarity, charge, or local protein conformation (Figure 3). Specifically, in the Gag protein mutations K47E and S464F alter the local charge and the local protein conformation, respectively (Figure 3). In the Pol protein, mutations P6S, A124V, K764R, R771G, and N924D between the Pci-SN265 and AF151794 had only minor effects on the overall topology of the structure. In Pol, mutations I19V, A822T, and S829P altered the local conformation of the Pci-SN265 and KV522 relative to the AF151794 structure by changing a surface residue to a buried one (I19V) and changing two partially buried residues to surface ones (A822T and S829P) (Figure 3). Only two positions in the Env protein could be structurally modeled (P147S, D187G). Both of these were radical substitutions that changed buried amino acids to exposed ones (Figure 3). Both of these changes were located away from the putative receptor-binding region, as this has been defined in [6].

Second, non-synonymous polymorphisms in the historical KoRV sequences were examined for predicted changes to the protein structure as compared to the modern consensus sequence Pci-SN265, which again served as a reference sequence (the amino acid residue in Pci-SN265 is always the first listed in each substitution). The effects on polypeptide structures of the nonsynonymous polymorphisms present in KoRV for each koala were examined using a composite sequence that contained all of the amino acid differences versus Pci-SN265. This composite sequence would necessarily combine polymorphisms present on different proviral loci. Nonetheless, this composite sequence would be useful in identifying all of the potential disruptions to predicted structure, when the effects of each mutation are considered individually. Several ancient variants were predicted to cause small local fluctuations of the KoRV structure. Specifically positions G33E, K421E, Q429K, and S464Y are predicted to alter the local charge at the Gag surface resulting in deviated conformations (Figure 4A, Figure S3). In the Pol protein, positions S514R, F396Y, A685S, and Y676N exchange a buried residue for a surface one, resulting in topological differences (Figure 4B, Figure S4). Additionally, in the Pol protein major conformational changes were predicted to occur at ancient variants R853Q, P933T, V939E, and T1014I. Lastly, two ancient variants S75F and R214W found in the Env protein were predicted to have major structural effects (Figure 4C, Figure S5). Both of these changes are located away from the putative receptor binding region as this has been defined in [6]. It is important to note that the other character state found at each of the polymorphic sites in the ancient koalas matched the character state present in the reference sequence. Thus, despite the presence of these polymorphisms and their modeled effects on proteins, KoRV overall has remained stable in sequence and structure over time.

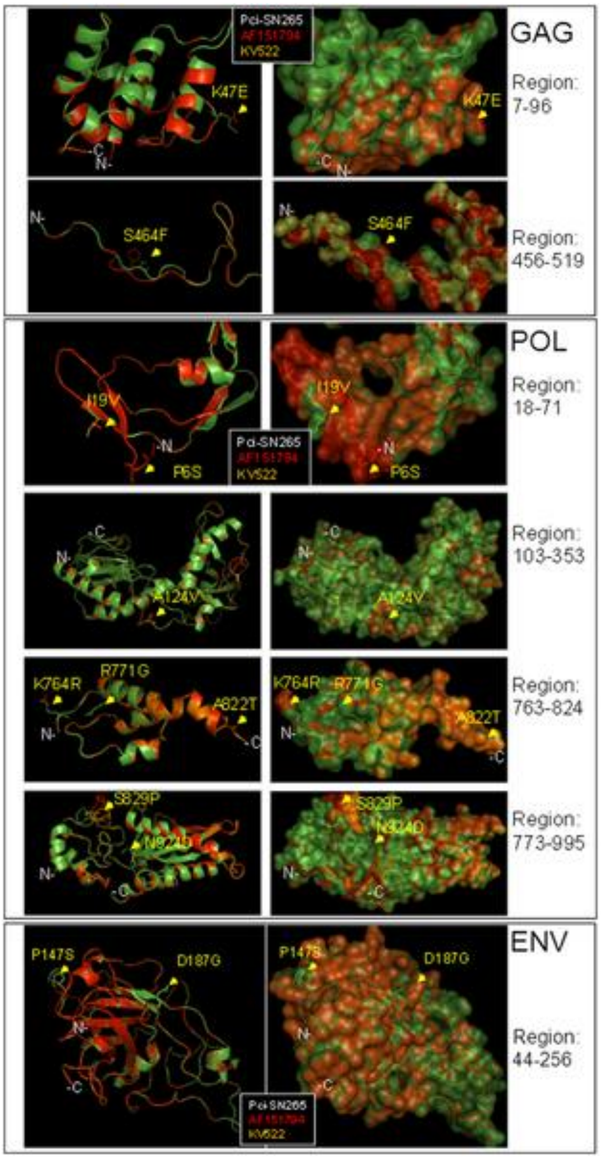


Figure 3. Structural superimpositions of Pci-SN265 (green), AF151794 (red), and KV522 (gold) KoRV Gag, Pol, and Env protein structures, demonstrating the overall similarity of the structures. Amino acid variations across these three sequences are mapped on the protein models (arrows). The structural differences predicted are attributed to changes in the polarity, charge, and atom conformations. The models are shown in cartoon (left panels) and semitransparent surface (right panels) representations. The atoms of the variable amino acid residues are shown in line representations to view the side chains. In all comparisons the Pci-SN265 consensus sequence was used as the reference.

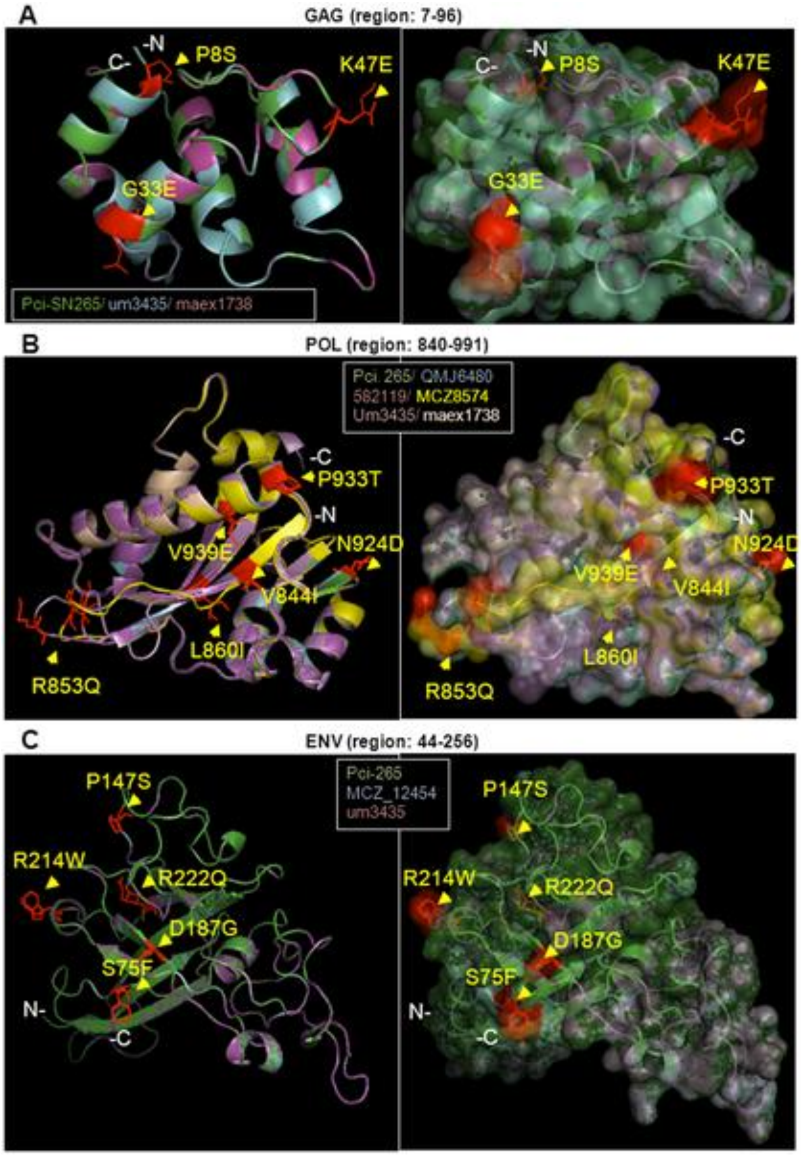


Figure 4. The effects of historical KoRV polymorphisms on protein structure. Superimpositions are shown between the present day consensus KoRV (Pci-SN265) protein structure and ancient KoRV variants. Amino acid variations between these sequences mapped on the protein models are shown in red and with arrows. The models are shown in cartoon ribbon representations (left panels) and as semitransparent surfaces (right panels). The atoms of the variable amino acid residues are in line representations to view the side chains. In all comparisons the Pci-SN265 consensus was used as the reference sequence. (A) The model of the Pci-SN265 Gag protein is superimposed with the models of variants found in archival koalas um3435 and maex1738. (B) The model of the Pci-SN265 Pol protein is superimposed

with variants found in QMJ6480, 582119, MCZ8574, Um3435, and maex1738. (C) The model of the Pci-SN265 Env protein is superimposed with the model of variants found in MCZ_12454 and um3435. For all three polypeptides, the structural differences predicted are attributed to changes in the polarity, charge, and atom conformations and are largely localized onto flexible loop regions.

2.4.6 LTR and Integration Site Diversity

Multiple polymorphisms were observed in the long terminal repeats (LTRs) that serve as promoter and terminator of the retroviral transcription process (Figure 1). MatInspector (Genomatix) was used to examine U3, R, and U5 regions of the LTRs for sequences matching transcription factor binding sites (TFBs) and for disruptions of TFBs by polymorphisms. Polymorphic sites from the next generation sequencing data were placed in phase when possible (Figure 1). This analysis of the LTRs revealed the presence of 82 putative TFBs (Table S3). Six of these had been referred to by Hanger et al. (2000), including a TATA box, CCAAT retroviral signal, and C-type poly-A signal (Table S6). None of the polymorphisms detected would have disrupted the previously predicted TFBs. However 20 additional predicted TFBs would be generated by the various polymorphisms.

Sequences at the 5' and 3' end of the KoRV genome often extended beyond the proviral integration sites into the host genomic flank. Shared integration sites among koalas would be strong evidence that a given locus represents an endogenized retrovirus, since the chance that two proviruses would independently integrate into the same locus is minuscule [42]. Four hundred twenty nine 5' flanks and three hundred thirty one 3' flanks were identified across all the koalas tested (Table S7). Thirty-two 5' and twenty-three 3' flanking sequences were shared by two or more koalas, representing 7.5% and 7% of the total respectively (Table S4).

The sequences flanking the integration sites were queried against flanks found in six modern koalas, which had been detected by other methods (Ishida et al., in preparation). Eight of the integration sites found by hybridization capture were also identified in one or more of six modern koalas tested (Table S7). The sequences flanking KoRV were also queried against whole genome sequences from a single koala, generated from onesixteenth plate GS-FLX shotgun sequence (Ishida et al., in preparation). Two flanking sequences had matches among the GS-FLX results. One of the flanking sequences matched 45 of the next-generation sequences, suggesting that this KoRV provirus had integrated in a repetitive element of the koala genome. The other matching flank sequence was detected only once. This sequence was KoRV negative at the locus, with the host genomic flank on the other side of the provirus evident in the sequence.

2.5 Discussion

Hybridization capture using archival samples has been used to efficiently sequence mitogenomes [15], bacterial genes [43] and low copy number genomically integrated viruses [44]. Here we use hybridization capture to generate sequences at high coverage across the full length of KoRV from both museum samples and modern genomic DNA. Information on both the provirus and its integration sites was obtained simultaneously, providing information on ca. 130 years of KoRV evolution. Limited variation was detected across the entire proviral genome including the LTRs. A previous study had examined env from several of the same samples used in this study. Using PCR and GS FLX sequencing, 20 polymorphisms and one fixed difference had been reported for env between museum samples and reference sequence AF151794 [6]. Of these 20, only 4 polymorphisms were also identified by the current study. However, 7 of the remaining 16 polymorphisms were also detectable in the current dataset, but at levels below the 8% threshold used to screen for polymorphisms. This may reflect the bias introduced by PCR based approaches to ancient DNA,

where molecules amplified in the earlier cycles (of which there are few to begin with) may come to dominate in the pool of sequences. This is particularly true for historical DNA where 60 or more cycles have been used to generate templates. In contrast, hybridization capture does not initially rely on PCR in enriching the target from the library. Library primers are used post enrichment to generate sufficient template for sequencing.

However, all templates have the sequences targeted by the primers, and a lower cycle number is used (7–30 cycles), which should yield a less biased data set. The drawback, common to PCR and hybridization capture, is that very low frequency polymorphisms may not be scored above background error and DNA damage levels, though this is anticipated to be a lesser problem with modern DNA than ancient DNA (which has a lower number of templates). Variation in coverage also influenced polymorphism scoring in the hybridization capture data; this had a larger impact on the historical samples that generally have lower

coverage (Fig. 2). For example 6 of the env polymorphisms not identified were likely due to low coverage in env for the poorest performing sample, Pci-maex1738. However, hybridization capture of all samples identified 14 novel polymorphisms not previously detected by PCR, including two novel polymorphisms

in the poorest performing sample Pci-maex1738. Increasing the depth of coverage is possible with hybridization capture, whereas removing bias from PCR based approaches is not. Thus, the ease, coverage and lower expense of hybridization capture provide advantages over PCR based approaches.

The polymorphisms in the gag, pol and env coding regions did not display any evident differences in the proportions of private versus shared alleles and/or in terms of synonymous versus nonsynonymous mutations. However, the number of radical versus non-radical amino acid was significantly different across the three coding regions. The relative number of radical mutations, those corresponding to large physiochemical differences between the amino acids, was significantly elevated in the env coding region

when compared to gag-pol coding regions. The higher proportion of radical changes in Env could potentially reflect either anti-viral immune pressure on the exposed portions of the Env proteins or avoidance of receptor interference [45–49].

However, none of the non-synonymous substitutions altered functional regions of the respective proteins reported as being critical for infection or replication previously reported [3] or altered any of the residues that have been functionally characterized in other viruses based on the CDD database (NCBI). The latter results imply that negative selection could be responsible for the conservation of these sites although statistically, deviation from neutral evolution was not observed.

Most tests of selection suggested that the evolution of KoRV does not deviate greatly from neutrality. An alignment of two KoRV sequences from GenBank with seven KoRV consensus sequences derived from the novel data did suggest a trend for purifying selection to have played a stronger role in gag and to have

a reduced role in env. A weak trend was evident in calculations of Tajima’s D, and was also suggested by codon-based Z-tests, although these were not significant after Bonferroni correction. An elevated number of non-synonymous changes in the Gag protein may potentially suggest that anti-viral proteins Such as TRIM5alpha are acting on KoRV. Evidence for Such selective pressure has been studied for TRIM5alpha itself [50]. Although these analyses indicate relaxation of constraints overall, purifying selection may

have shaped and conserved particular structural and functional elements.

The LTR region enrichment also resulted in retrieval of viral integration sites. Only ca. 7% of the integration sites were found in two or more koalas which suggests fixation of specific KoRV integrations is not very advanced in koalas even where KoRV has been present the longest Such as Queensland. The large number of unique integrations is consistent with previously reported results for Southern blot hybridization based on pol and env genes, which suggested that KoRV integration sites were quite variable across individual koalas [51]. Although the hybridization capture method would potentially capture both endogenous and exogenous proviruses, the presence of proviruses at the same locus in more than one koala would indicate that at least some of the sequences obtained are from endogenous retroviruses.

We found no evidence for KoRV-B in any of the historical koalas, although partial KoRV-B receptor sequences were identified in the modern koala Pci-SN265. The modern koala was born in the Houston Zoo, Texas and had a complex pedigree and transfer history including transfer among American and European institutions. Thus, exposure to KoRV-B infected individuals may have been possible although the exact source of infection cannot be determined. Alternatively, KoRV-B may be more widespread in captive koalas than previously estimated. However, the absence of KoRV-B in the historical datasets would be consistent with a recent emergence of this variant. By contrast, two of three KoRV variants described from koalas in Japanese zoos were also detected in the modern koala Pci-SN265. The absence of KoRV-J sequences in the museum koalas is consistent with a recent origin of these sequences.

Overall, our results suggest that for ca. 130 years, the majority of KoRV proviruses have remained conserved with one of the character states at each ancient polymorphism matching that of modern KoRV. Considering the potential pathological effects of modern KoRV, its historical genomic and structural stability

suggests that koalas have suffered long term negative health impacts in populations where KoRV has occurred. It also suggests that fitness may eventually decrease in koala populations in southern Australia where KoRV appears to be emerging.

2.6 References

1. Tarlinton R, Meers J, Young P (2008) Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci 65: 3413–3421.

2. Hunter P (2010) The missing link. Viruses revise evolutionary theory. EMBO Rep 11: 28–31.

3. 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.

4. Cornelis G, Heidmann O, Degrelle SA, Vernochet C, Lavialle C, et al. (2013) Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. Proc Natl Acad Sci U S A 110: 828–837.

5. Kewitz S, Staege MS (2013) Expression and Regulation of the Endogenous Retrovirus 3 in Hodgkin’s Lymphoma Cells. Front Oncol 3: 179.

6. Avila-Arcos MC, Ho SY, Ishida Y, Nikolaidis N, Tsangaras K, et al. (2013) One hundred twenty years of koala retrovirus evolution determined from museum skins. Mol Biol Evol 30: 299–304.

7. 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.

8. Tarlinton R, Meers J, Hanger JJ, 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. Journal of general virology 86: 783–787.

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

10. Stoye JP (2006) Koala retrovirus: a genome invasion in real time. Genome Biol 7: 241.

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

12. Xu W, Stadler CK, Gorman K, Jensen N, Kim D, et al. (2013) An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proc Natl Acad Sci U S A 110: 11547–11552.

13. Briggs AW, Good JM, Green RE, Krause J, Maricic T, et al. (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325: 318–321.

14. Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One 5(11):e14004.

15. 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. Gen Res 21: 1695–1704.

16. Tsangaras K, Avila-Arcos MC, Ishida Y, Helgen KM, Roca AL, et al. (2012) Historically low mitochondrial DNA diversity in koalas (Phascolarctos cinereus). BMC Genet 13: 92.

17. Wyatt KB, Campos PF, Gilbert MT, Kolokotronis SO, Hynes WH, et al. (2008) Historical mammal extinction on Christmas Island (Indian Ocean) correlates with introduced infectious disease. PLoS One 3(11): e3602.

18. Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols doi:10.1101/pdb.prot5448.

19. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365–386.

20. Hanke M, Wink M (1994) Direct DNA sequencing of PCR-amplified vector inserts following enzymatic degradation of primer and dNTPs. Biotechniques 17: 858–860.

21. Lindgreen S (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes 5: 337.

22. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 17: 10–12.

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

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

25. Ginolhac A, Rasmussen M, Gilbert MT, Willerslev E, Orlando L (2011) mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 27: 2153–2155.

26. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568–576.

27. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426.

28. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731–2739.

29. Pond SLK, Frost SDW (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure (vol 22, pg 478, 2005). Molecular Biology and Evolution 22: 1157–1157.

30. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22: 195–201.

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

32. Holm L, Park J (2000) DaliLite workbench for protein structure comparison. Bioinformatics 16: 566–567.

33. Horn S (2012) Case study: enrichment of ancient mitochondrial DNA by hybridization capture. Methods Mol Biol 840: 189–195.

34. Kircher M (2012) Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol 840: 197–228.

35. Burger J, Hummel S, Hermann B, Henke W (1999) DNA preservation: a microsatellite-DNA study on ancient skeletal remains. Electrophoresis 20: 1722–1728.

36. Capelli C, Tschentscher F, Pascali VL (2003) ‘‘Ancient’’ protocols for the crime scene?: Similarities and differences between forensic genetics and ancient DNA analysis. Forensic Sci Int 131: 59–64.

37. Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, et al. (2004) Genetic analyses from ancient DNA. Annu Rev Genet 38: 645–679.

38. Bull RA, Eden JS, L°Ciani F, McElroy K, Rawlinson WD, et al. (2012) Contribution of intra- and interhost dynamics to norovirus evolution. J Virol 86:3219–3229.

39. Shojima T, Hoshino S, Abe M, Yasuda J, Shogen H, et al. (2013) Construction and characterization of an infectious molecular clone of Koala retrovirus. J Virol 87: 5081–5088.

40. Shimode S, Nakagawa S, Yoshikawa R, Shojima T, Miyazawa T (2014) Heterogeneity of koala retrovirus isolates. FEBS Lett 588: 41–46.

41. Han JY, Zhao Y, Anderson WF, Cannon PM (1998) Role of variable regions A and B in receptor binding domain of amphotropic murine leukemia virus envelope protein. J Virol 72: 9101–9108.

42. Stoye JP (2001) Endogenous retroviruses: Still active after all these years? Current Biology 11: 914–916.

43. Schuenemann VJ, Bos K, DeWitte S, Schmedes S, Jamieson J, et al. (2011) Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. Proc Natl Acad Sci U S A 108:746–752.

44. Duncavage EJ, Magrini V, Becker N, Armstrong JR, Demeter RT, et al. (2011) Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. J Mol Diagn 13: 325–333.

45. Yan Y, B°Ckler-White A, Wollenberg K, Kozak CA (2009) Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene Fv1 in the genus Mus. Proc Natl Acad Sci U S A 106: 3259–3263.

46. Hu Y, Tan PT, Tan TW, August JT, Khan AM (2013) Dissecting the dynamics of HIV-1 protein sequence diversity. PLoS One 8(4): e59994.

47. De Feo CJ, Weiss CD (2012) Escape from human immunodeficiency virus type 1 (HIV-1) entry inhibitors. Viruses 4: 3859–3911.

48. Melder DC, Pankratz VS, Federspiel MJ (2003) Evolutionary pressure of a receptor competitor selects different subgroup a avian leukosis virus escape variants with altered receptor interactions. J Virol 77: 10504–10514.

49. Tsibris AM, Korber B, Arnaout R, Russ C, Lo CC, et al. (2009) Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. PLoS One 4(5): e5683.

50. Ortiz M, Bleiber G, Martinez R, Kaessmann H, Telenti A (2006) Patterns of evolution of host proteins involved in retroviral pathogenesis. Retrovirology: 3:11.

51. Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442: 79–81.

2.7 Supplementary material

Fig. S1. Alignment of hybridization capture sequences to koala retrovirus B (KoRV-B) isolate Br2-1CETTG. KoRV-B sequences are shown and used as a reference. Individual reads from koala Pci-SN265 were aligned with dots indicating a match, dashes indicating

indels, and mismatches indicated by the appropriate DNA base. The KoRV-A sequence AB721500 (Shojima et al. 2013) is included to highlight the differences between KoRV-A and KoRV-B. The sequences shown also correspond to positions 6149-6436 of the reference sequence AF151794 (Hanger et al. 2000).

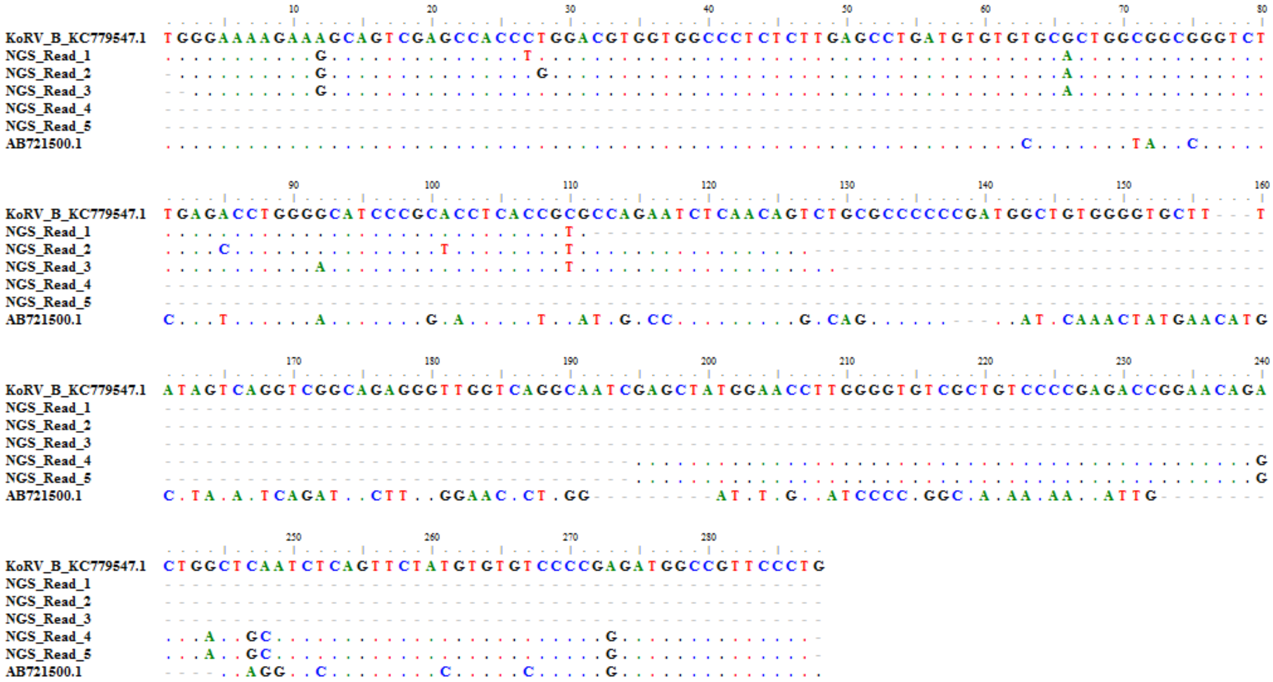
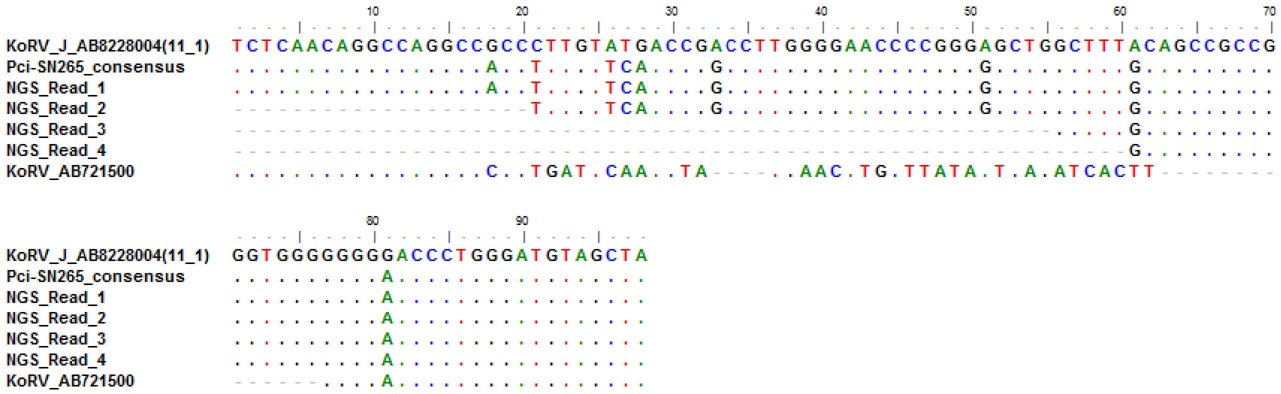
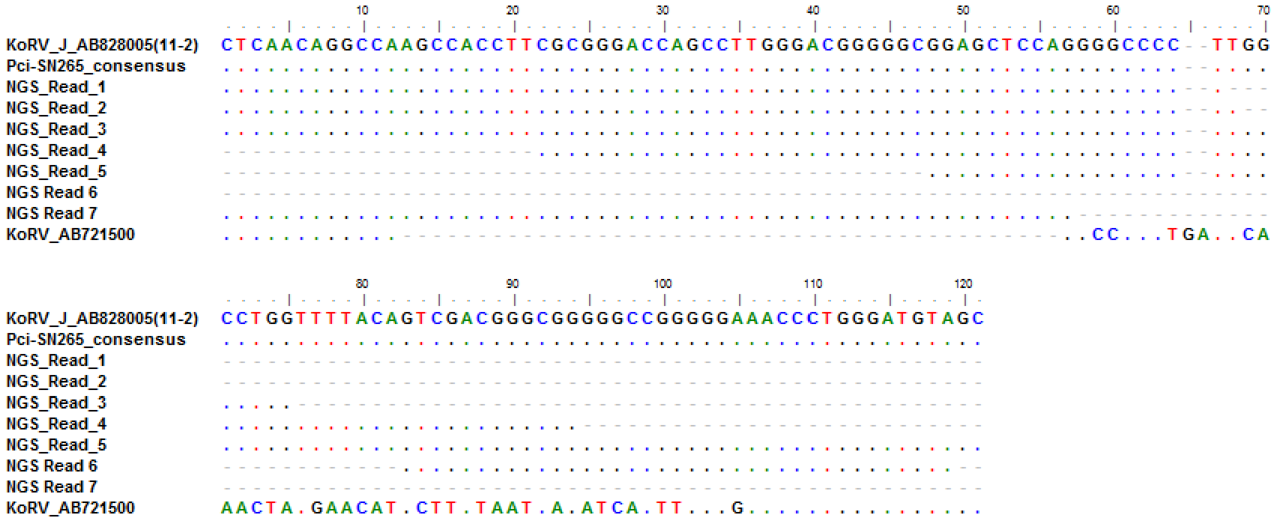


Fig. S2. Alignment of hybridization capture sequences to KoRV isolates identified in Japanese zoo koalas: clone 11-1 (panel A) and 11-2 (Panel B). Clone 11-1 and 11-2 sequences correspond respectively to positions 6205-6303 and 6204-6324 of AF151794 (Hanger et al. 2000). The sequence from each clone is shown as the references to which ale aligned the Pci-SN265 consensus sequence, as well as individual next-generation reads. The sequence of KoRV AB721500 (Shojima et al. 2013) was also aligned to highlight differences versus the two Japanese isolates. Identity to the reference is indicated by a dot, differences by the corresponding base, and indels by a dash.

A



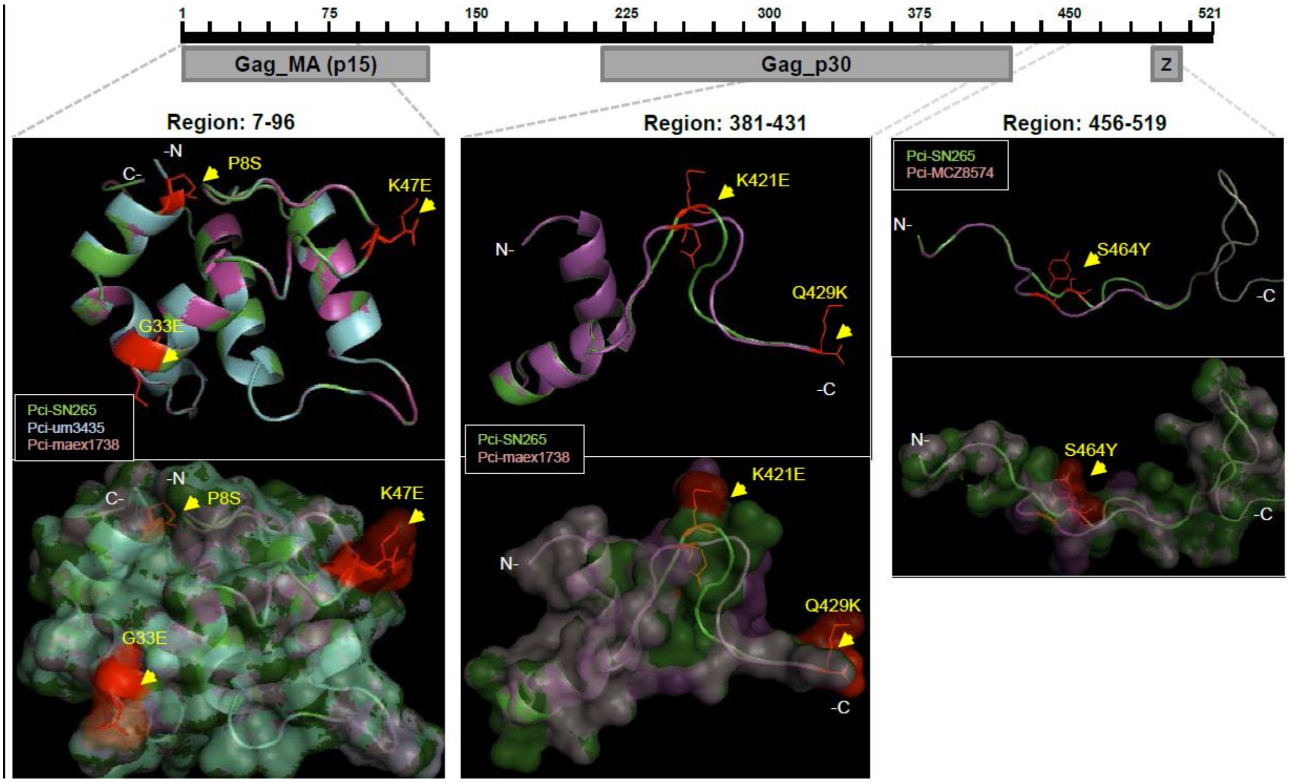
B



```
                                   10        20        30        40        50        60        70
                          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
KoRV_J_AB828005(11-2)     CTCAACAGGCCAAGCCACCTTCGCGGGACCAGCCTTGGGACGGGGGCGGAGCTCCAGGGGCCCC--TTGG
Pci-SN265_consensus       ................................................................--....
NGS_Read_1                ................................................................--.---
NGS_Read_2                ................................................................--..--
NGS_Read_3                ................................................................--....
NGS_Read_4                ---------------------...........................................--....
NGS_Read_5                -----------------------------------------------.................--....
NGS Read 6                ----------------------------------------------------------------------
NGS Read 7                .........................................................-------------
KoRV_AB721500             ............--------------------------------------------..CC...TGA..CA

                                   80        90        100       110       120
                          ....|....|....|....|....|....|....|....|....|....|.
KoRV_J_AB828005(11-2)     CCTGGTTTTACAGTCGACGGGCGGGGGCCGGGGGAAACCCTGGGATGTAGC
Pci-SN265_consensus       ...................................................
NGS_Read_1                ---------------------------------------------------
NGS_Read_2                ---------------------------------------------------
NGS_Read_3                .....----------------------------------------------
NGS_Read_4                ........................---------------------------
NGS_Read_5                ...................................................
NGS Read 6                ------------.....................................--
NGS Read 7                ---------------------------------------------------
KoRV_AB721500             AACTA.GAACAT.CTT.TAAT.A.ATCA.TT...G................
```

Fig S3. Superimpositions between the modern KorV (Pci-SN265-green) protein structure and its historical variants show the overall similarity of the structures of the Gag protein. Amino acid variations (red) between these sequences are mapped on the protein models (arrows). The structural differences predicted are attributed to changes in the polarity, charge, and atom conformations. These differences are largely localized onto flexible loop regions. The models are shown in cartoon representations and the atoms of the variable amino acid residues in line representations to view the side chains. In all comparisons the consensus sequence of Pci-SN265 was used as the reference. The domain organization (as depicted at the CDD database) and the location of the modeled structure of the protein are shown at the top of the figure. Gag_MA: Gag Matrix protein; Gag_p30: core shell protein; Z: zinc finger.



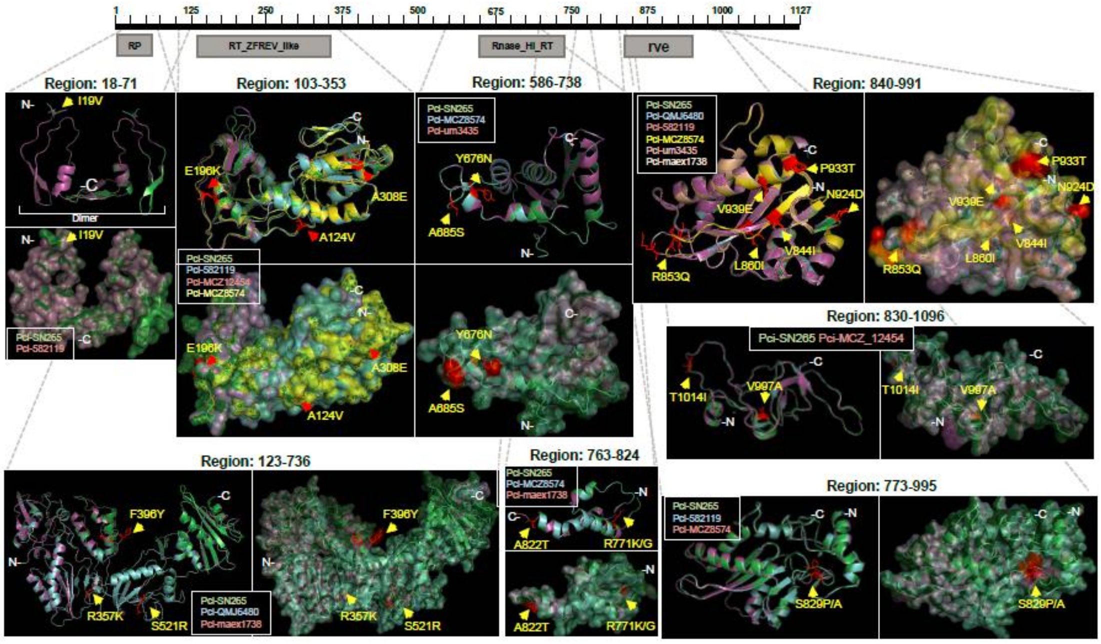


Fig S4. Superimpositions between the modern KoRV (Pci-SN265-green) protein structure and its historical variants show the overall similarity of the structures of the Pol protein. Amino acid variations (red) between these sequences are mapped on the protein models (arrows). The structural differences predicted are attributed to changes in the polarity, charge, and atom conformations. These differences are largely localized onto flexible loop regions. The models are shown in cartoon representations and the atoms of the variable amino acid residues in line representations to view the side chains. In all comparisons the Pci-SN265 consensus sequence was used as the reference. The domain organization (as depicted at the CDD database) and the location of the modeled structure of the protein are shown at the top of the figure. RP: Retropepsin of the RTVL_H family of human endogenous retrovirus-like elements; RT_ZFREV_like: Reverse transcriptase subfamily found in sequences similar to the intact endogenous retrovirus from zebrafish and Moloney murine leukemia virus; Rnase_HI_RT: Bel/Pao family of RNase HI in long-term repeat retroelements ; rve: Integrase core domain.

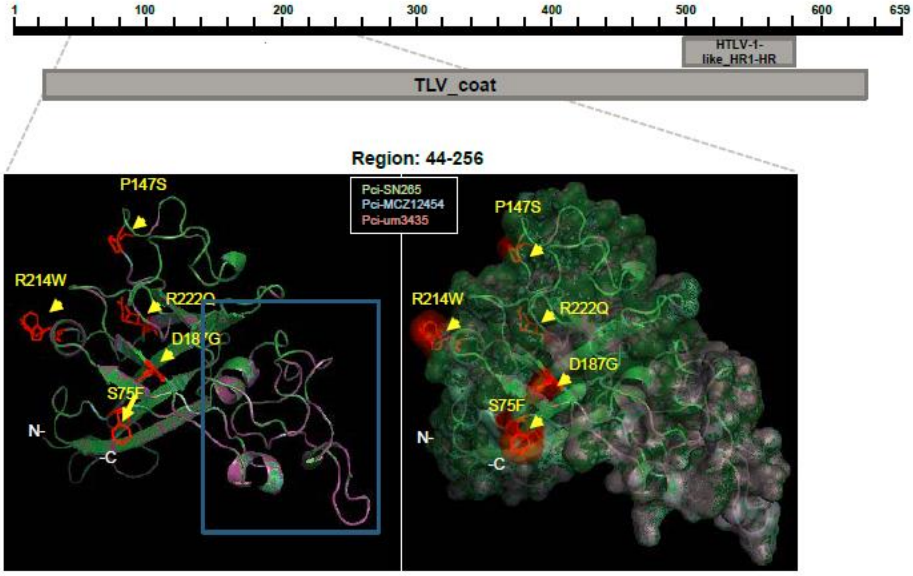


Fig S5. Superimpositions between the modern KoRV (Pci-SN265-green) protein structure and its historical variants show the overall similarity of the structures of the Env protein. Amino acid variations (red) between these sequences are mapped on the protein models (arrows). The structural differences predicted are attributed to changes in the polarity, charge, and atom conformations. These differences are largely localized onto flexible loop regions. The models are shown in cartoon representations and the atoms of the variable amino acid residues in line representations to view the side chains. In all comparisons the Pci-SN265 sequence was used as the reference sequence. The domain organization (as depicted at the CDD database) and the location of the modeled structure of the protein are shown on the top of the figure.

Table S1: Koala retrovirus (KoRV) primers

Primer Name	Primer Sequence
PCI-KoRV-F1	5'-AGGAGGCAGAAATCATGAGG-3'
PCI-KoRV-R1	5'-AGAAACCCTCCCAGGATCAA-3'
PCI-KoRV-F2	5'-TCGTAAGTTCAATAAACCTCTTGC-3
PCI-KoRV-R2	5'-ACGTATATTAAAAAGACAGGAAAA
PCI-KoRV-F3	5'-GAGATTCCCACCCAAGGAC-3'
PCI-KoRV-R3	5'-CAGTGATCTAGTGTAAGAGAGAGA
PCI-KoRV-F4	5'-CCTGTCTTTTTAATATACGTCTACG
PCI-KoRV-R4	5'-CGACTTTCGCCCGTTATC-3'
PCI-KoRV-F5	5'-GGGTGAGTCGACCCCTCT-3'
PCI-KoRV-R5	5'-GAGTCCCTCAGCCATTAGGC-3'
PCI-KoRV-F6	5'-AAGATCGCCGTTGCCTCT-3'
PCI-KoRV-R6	5'-AGGAGCTGCTGGCAATCG-3'
PCI-KoRV-F7	5'-ATCCAACGTCCCCTCCAC-3'
PCI-KoRV-R7	5'-GATCCCACTGAGGTCGATT-3'
PCI-KoRV-F8	5'-TTTTCCCACCAGCCTACTTG-3'
PCI-KoRV-R8	5'-AGATCCTGCAAGGAATGGTC-3'
PCI-KoRV-F9	5'-GCCCCTACACAACTCGAGAA-3'
PCI-KoRV-R9	5'-TTTCTCCTGGCGCCTGTC-3'
PCI-KoRV-F10	5'-TTACAAAGGCTGGAAGGACTC-3'
PCI-KoRV-R10	5'-GCTTGGTCAATACTGAATGTTCG-3'
PCI-KoRV-F11	5'-GAGACAGAGGAAAGAGAGAGACG-
PCI-KoRV-R11	5'-AAGAATGAGTGGGTCACTTGC-3'
PCI-KoRV-F12	5'-CTGAGTTTTTGGTTGATACCG-3'
PCI-KoRV-R12	5'-TTGCTCATTGGGTACTGTCG-3'
PCI-KoRV-F13	5'-CAAGAGACTTTTGAAAATTGGACA-
PCI-KoRV-R13	5'-CGATAGTCATTGGTTCCAGGT-3'
PCI-KoRV-F14	5'-TGAAGTCAGATGCCTCACCA-3'
PCI-KoRV-R14	5'-GTTGAGAGCCCTGAAGGATG-3'
PCI-KoRV-F15	5'-CCTGGAACACCCCTTTGTTA-3'
PCI-KoRV-R15	5'-TTGGCCGACACTCGGTAT-3'
PCI-KoRV-F16	5'-ACTCTCCCACCCTCTTCGAT-3'
PCI-KoRV-R16	5'-GCCTCTTTTATACGGCCAAA-3'
PCI-KoRV-F17	5'-GGGACACGAAGGCTCTTACA-3'
PCI-KoRV-R17	5'-GGCCACCGGATCTAATTTTT-3'
PCI-KoRV-F18	5'-TCCCTTTACCTGGACTGAGG-3'
PCI-KoRV-R18	5'-ATTGGGGTGTCGTCTGACTC-3'
PCI-KoRV-F19	5'-CCCGGTAGCTTACCTGTCAA-3'
PCI-KoRV-R19	5'-GTTCCCTCTGGCAGGTTG-3'
PCI-KoRV-F20	5'-CGGCCATTCTGAATCCTG-3'
PCI-KoRV-R20	5'-CGGGGCAATGGATGATAG-3'
PCI-KoRV-F21	5'-GCCATTGTGGACAACAAGC-3'
PCI-KoRV-R21	5'-GATGCAGCCGTTGAATGAAT-3'
PCI-KoRV-F22	5'-TGCTAGAGGCCATCCATCTC-3'
PCI-KoRV-R22	5'-GCTTGTCTGGCCCTAAGTG-3'
PCI-KoRV-F23	5'-CTACACGGGGGAAGATCAAG-3'
PCI-KoRV-R23	5'-TGTCGGACCCGAGTACCTTA-3'

Table S2: KoRV variable sites in modern and historic koalas

Position	20	70	93	111	136	145	147	157	312	357	470	477	492	495	602	604	783	804	806	827
AF151794	A	G	G	G	G	G	G	G	C	T	A	A	T	G	G	C	T	T	C	G
AB721500	A	G	G	G	G	A	A	G	T	C	A	A	T	G	G	C	C	T	T	G
Vienna Zoo	A	G	G/A	G	G	G/A	G/A	G/A	T	C	A	A/T	C/T	G/A	G	C	C/T	C/T	T	G/A
QMJ6480	A	G	G/A	G/A	G	G/A	G/A	G/A	T	C	A	A/T	T	G	G	C	C/T	C/T	T	G/A
521198	A	G	G/A	G	G	G/A	G/A	G	T	C/T	A	A/T	T	G	G	C	C/T	T	C/T	G/A
MCZ12454	A	G	G/A	G	G/A	G/A	G/A	G	T	C/T	A	A/T/G	C/T	G	C/G	C	C/T	T	T	G/A
MCZ8574	A	G	G/A	G	G/A	G/A	G/A	G	T	C	G/A	A/T	T	G	G	C	C/T	T	T/A	G/A
um3435	A	G	G/A	G	G	G/A	G/A	G	T	C/T	A	A/T	T	G	G	C	C/T	T	T	G/A
maex1738	G/A	G/A	G/A	G	G	G/A	G/A	G	T	C	A	A/T	T	G	C/G	C/T	C/T	T	T	G/A
Amino Acid	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Position	866	874	900	909	910	916	964	991	1002	1065	1067	1108	1113	1197	1263	1301	1325	1404	1689	1809
AF151794	G	A	T	C	C	A	C	C	T	G	G	A	G	G	C	G	C	G	T	A
AB721500	G	A	T	C	C	A	A	C	T	G	G	A	G	G	C	G	C	G	C	G
Vienna Zoo	G	A	A/T	C/T	C/T	A	A/C	C	T	G	G	A/G	G	G/A	C/T	G	C	G/A	T/C	A/G
QMJ6480	G	A	T	C/T/G	C/T/G	A	A/C	C	T	G	G	A/G	G/A	G	C	G/A	C	G	T/C	A/G
521198	G	A	T	C/T	C/T	A	A/C	C	T	G	G	A/G	G	G/A	C/T	G	C	G	T/C	A/G
MCZ12454	G	A	T	C/T	C/T	A	A/C	C	T/C	G	G	A/G	G	G	C/T	G	C	G/A	T/C	A/G
MCZ8574	G	A	T	C/T	C/T	A/T	C	C	T	G/A	G	A/G	G	G	C	G	C	G/A	T/C	A/G
um3435	G/T	G/A	T	C/T	C/T	A	A/C	C/T	T	G	G	A/G	G	G	C	G	C	G	T/C	A/G
maex1738	G	A	T	C/T	C/T	A/T	A/C	C	T	G	G/A	A/G	G	G	C	G	C/G	G	T/C	A/G
Amino Acid	-	-	-	-	-	-	-	-	-	-	-	E/K	V	P	P	G/E	P/R	T	L	K

Position	1993	2019	2130	2133	2154	2166	2229	2230	2240	2244	2245	2248	2254	2261	2263	2276	2360	2412	2665	2704
AF151794	G	A	G	G	G	G	T	A	C	G	G	G	C	G	G	C	T	G	T	G
AB721500	G	C	A	G	G	A	T	A	C	G	G	G	C	G	G	C	C	G	C	G
Vienna Zoo	G	A/C	A	G/A	G	G/A	T	A	C	G	G	G	C	G	G	C	C	G	C/T	G/A
QMJ6480	G	A/C	A	G	G	G/A	T	A	C/G	G	G	G	C	G	G	C/T	C	G	C	G/A
521198	G	A/C	A	G	G	G/A	T	A	C/G	G/A	G	G	A/C	G	G	C	C	G	C	G
MCZ12454	G/A	A/C	G/A	G	G/A	G/A	T	A	C/G	G	G/A	G/A	A/C	G	G/A	C	C	G	C	G
MCZ8574	G	A/C	A	G	G	G/A	T	A/G	C/G/A	G	G	G	A/C	G	G	C	C/A	G	C/A	G
um3435	G	A/C	G/A	G	G	G/A	A/T	A	C/G/A	G	G/A	G/A	A/C	G	G	C	C	G	C/T	G
maex1738	G	A/C	A	G	G	G/A	T	A/G	C/G	G	G	G	A/C	G/A	G/A	C	C/T	G/A	C	G/A
Amino Acid	V/I	T	G	Q	K	R	H/Q	K/E	T/R/K	E	E/K	E/K	Q/K	K/R	E/K	T/I	S/T/Y	R	S/T/P	V/I

Position	2880	2922	2926	3020	3069	3235	3306	3336	3572	3672	3717	3719	3736	3810	3836	4212	4485	4675	4702	4725
AF151794	C	A	G	T	T	G	G	C	C	C	G	G	C	A	T	C	A	T	G	A
AB721500	C	A	G	T	T	G	G	C	C	C	G	G	C	A	T	T	A	T	G	G
Vienna Zoo	C/T	G/A	G	T/C	T	G	G	C/T	C	C	G	G	C	G/A	T	T	A/G	T	G	A/G
QMJ6480	C/T	G/A	G	T/C	T	G	G	C	C	C	G	G	C	A	T	T/A	A/G	T	G	A/G
521198	C/T	G/A	G	T/C	T	G	G	C	C/A	C	G	G	C	A	T	T	A/G	T	G	A/G
MCZ12454	C/T	A	G	T/C	T/C	G/A	G/T	C	C	C	G	G	C	A	T	T	A	T	G	A/G
MCZ8574	C	G/A	G	T/C	T	G	G	C	C	C	G	G	C/T	A	T	T/C	A	T	G/T	A/G
um3435	C/T	G/A	G	T/C	T	G	G	C/T	C	C/T	G	G	C	A	T	T/A	A/G	T/A	G	A/G
maex1738	C/T	G/A	G/A	T/C	T	G	G/T	C	C	C	G/A	G/A	C	A	T/A	T/C	A/G	T	G	A/G
Amino Acid	P	L	E/K	A/V	V	E/K	P	D	A	C	K	R	R/*	G	F/Y	S/R	R	Y	A/S	K

Position	4940	4960	5004	5113	5133	5134	5179	5207	5227	5419	5446	5465	5639	5646	5690	6138	6157	6353	6474	6541
AF151794	G	G	G	G	G	C	G	G	C	G	C	T	C	G	C	C	C	T	G	C
AB721500	G	G	A	A	A	T	G	G	C	A	C	T	C	G	C	C	C	T	G	C
Vienna Zoo	G/A	G/A	G/A	G/A	G/A	T	G	G	C	G/A	C	T	C/T	G/A	C	C	C/T	T/C	G/A	C
QMJ6480	G/A	G/A	G/A	G/A	G/A	C/T	G/A	G	C	G/A	C	T	C/T	G	C	C	C/T	T	G/A	C/T
521198	G/A	G/A	G/A	G/A	G/A	C/T	G	G	C	G/A	C	T	C	G/A	C	C	C/T	T	G	C
MCZ12454	G/A	G/A	A	G/A	G/A	T/G	G	G	C	G/A	C	T	C	G	C/T	C/T	C/T	T	G/A	C
MCZ8574	G/A	G/A	A	G/A	G/A	T/G	G	G	C	G/A	C	T	C	G	C	C	C/T	T	G	C
um3435	G/A	G/A	G/A	G/A	G/A	C/T	G	G/A	C/A	G/A	C	T/A	C	G	C	C	C/T	T	G	C
maex1738	G/A	G/A	G/A	G/A	G/A	T/A	G	G	C	G/A	C/A	T	C	G	C	C	C/T	T	G	C
Amino Acid	R/K	R/G	G	A/T	E	P/S	V/I	R/Q	L/I	N/D	P/T	V/E	A/V	G	T/I	S/F	C	S/P	G/D	T

Position	6554	6579	6710	6717	6765	6779	6922	6964	6993	7046	7097	7137	7140	7287	7289	7336	7381	7891	7925	7933
AF151794	C	G	A	C	C	C	T	T	G	G	A	G	G	A	T	C	A	T	G	G
AB721500	C	G	A	C	C	C	T	T	G	G	A	A	G	A	C	C	A	T	G	G
Vienna Zoo	C	G	A/G	C	C/T	C	T/C	T	G	G	A	A	G	A	C	C	A/G	T/C	G	G
QMJ6480	C	G	A/G	C/T	C	C	T/C	T	G/A	G	A	A	G/A	A	T/C	C	A	T/C	G	G
521198	C	G	A/G	C	C/T	C	T/C	T/A	G	G	A	A	G	A	T/C/A	C	A	T/C	G/A	G/A
MCZ12454	C/T	G	A/G	C	C	C	T/C	T	G/A	G	A	A	G	A/T	C	C/T	A	T/C	G	G
MCZ8574	C	G	A/G	C	C/T	C/T	T/C	T	G	G/A	A/G	A/C	G	A	C	C	A	T/C	G/A	G/A
um3435	C	G/A	A/G	C	C	C	T/C	T/A	G	G	A	A/C	G	A	T/C	C	A	T/C	G	G
maex1738	C	G	A/G	C	C	C	T/C	T/A	G/A	G	A	A/T	G	A/T	T/C	C	A	T/C	G	G
Amino Acid	R/W	R/Q	K/E	P/L	P/L	Q/*	P	Y/*	G/E	G/R	N/D	R/H/P/K	S/N	H/P	P/S	L	L	L	-	-

Position	7938	8071	8076	8080	8091	8092	8405	8410	8412	8426	8430	8439	8440
AF151794	T	G	A	G	G	G	T	A	A	A	G	A	T
AB721500	T	G	G	A	G	G	T	G	A	A	G	A	T
Vienna Zoo	T	G/A	A/G	G/A	G/A	G	T/C	A/G	A/T	A/T	G/A	A	T/A
QMJ6480	T	G	A/G	G/A	G	G/A	T/C	G	A/T	A	G	A	T/G
521198	T	G	A/G	G/A	G	G	T/C	A/G	A/T	A	G	A	T/A
MCZ12454	T	G	A/G	G/A	G	G	T/C	A/G	A/T	A	G	A	T/C
MCZ8574	T	G	A/G	G/A	G	G	T/C	A/G	A/T	A	G	A	T
um3435	T	G	A/G	G/A	G	G	T/C	G	A/T	A	G	A	T/A
maex1738	T/G	G	A/G	G/A	G	G	T/C	G	A/T	A	G	A/T	G
Amino Acid	-	-	-	-	-	-	-	-	-	-	-	-	-

-Non coding regions

Position numbers follow those of the KoRV reference genome, GenBank accession AF151794

The first two sequences are from Genbank; the "Vienna Zoo" koala is PCI-SN265; for other koala specimens see Table 1

Variation present in more than two koalas (shared)

Variation present in one koala (private)

Fixed difference

The first amino acid corresponds to AF151794; the second or subsequent amino acids are in order of appearance in the rows. Asterisk (*) indicates a stop codon

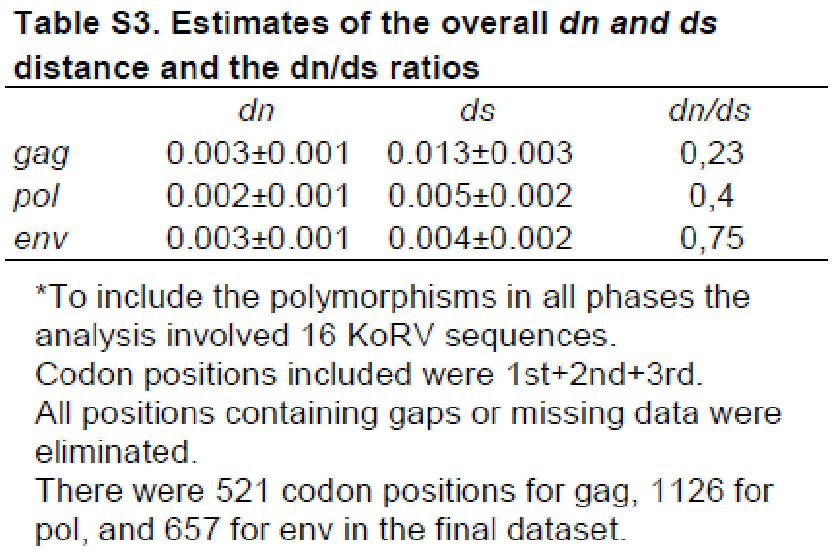


Table S3. Estimates of the overall *dn and ds* distance and the dn/ds ratios

	dn	*ds*	*dn/ds*
gag	0.003±0.001	0.013±0.003	0,23
pol	0.002±0.001	0.005±0.002	0,4
env	0.003±0.001	0.004±0.002	0,75

*To include the polymorphisms in all phases the analysis involved 16 KoRV sequences.
Codon positions included were 1st+2nd+3rd.
All positions containing gaps or missing data were eliminated.
There were 521 codon positions for gag, 1126 for pol, and 657 for env in the final dataset.

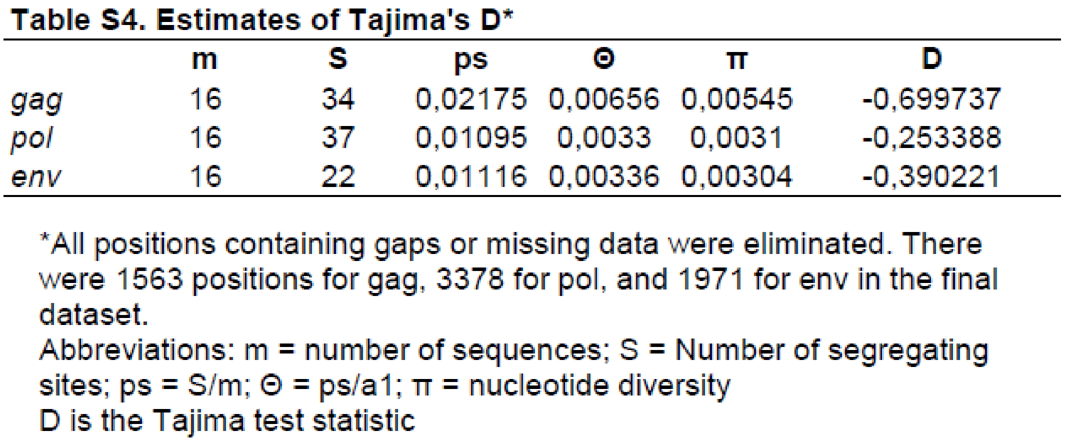


Table S4. Estimates of Tajima's D*

	m	S	ps	Θ	π	D
gag	16	34	0,02175	0,00656	0,00545	-0,699737
pol	16	37	0,01095	0,0033	0,0031	-0,253388
env	16	22	0,01116	0,00336	0,00304	-0,390221

*All positions containing gaps or missing data were eliminated. There were 1563 positions for gag, 3378 for pol, and 1971 for env in the final dataset.
Abbreviations: m = number of sequences; S = Number of segregating sites; ps = S/m; Θ = ps/a1; π = nucleotide diversity
D is the Tajima test statistic

Table S5. Codon based Z tests of selection

gag (521 codons)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pci_SN265		0	0	-1,049	-0,65	-0,662	-1,015	-0,677	-1,817	-2,716	-1,775	-2,011	-2,221	-2,138	-0,794	-1,318
Pci_QMJ6480	1		0	-1,049	-0,65	-0,662	-1,015	-0,677	-1,817	-2,716	-1,775	-2,011	-2,221	-2,138	-0,794	-1,318
Pci_582119	1	1		-1,049	-0,65	-0,662	-1,015	-0,677	-1,817	-2,716	-1,775	-2,011	-2,221	-2,138	-0,794	-1,318
Pci_MCZ_12454	0,296	0,296	0,296		-1,17	0,988	-1,499	1,017	-2,068	-2,902	-2,002	-2,24	-2,425	-2,327	-1,14	-1,584
Pci_MCZ8574	0,517	0,517	0,517	0,244		-1,456	-1,145	-0,939	-1,856	-2,751	-1,998	-2,219	-2,411	-2,086	-1,269	-1,577
Pci_um3435	0,509	0,509	0,509	0,325	0,148		-1,218	1,441	-1,886	-2,749	-2,004	-2,242	-2,427	-2,484	-1,298	-1,586
Pci_maex1738	0,312	0,312	0,312	0,137	0,254	0,225		-1,225	-2,099	-2,914	-2,023	-2,249	-2,464	-2,373	-1,153	-1,003
1Pci_SN265	0,5	0,5	0,5	0,311	0,35	0,152	0,223		-1,91	-2,787	-1,88	-2,106	-2,315	-2,211	-1,001	-1,462
1Pci_QMJ6480	0,072	0,072	0,072	0,041	0,066	0,062	0,038	0,059		-2,09	-0,682	-0,948	-1,331	-1,154	2,63	0,173
1Pci_582119	0,008	0,008	0,008	0,004	0,007	0,007	0,004	0,006	0,039		-2,026	-0,913	-1,306	-1,793	-0,989	-1,489
1Pci_MCZ_12454	0,078	0,078	0,078	0,048	0,048	0,047	0,045	0,063	0,496	0,045		-1,624	-1,837	-1,55	0,421	-0,688
1Pci_MCZ8574	0,047	0,047	0,047	0,027	0,028	0,027	0,026	0,037	0,345	0,363	0,107		-1,565	-1,94	-0,249	-1,165
1Pci_um3435	0,028	0,028	0,028	0,017	0,017	0,017	0,015	0,022	0,186	0,194	0,069	0,12		-1,655	-1,46	-1,728
1Pci_maex1738	0,035	0,035	0,035	0,022	0,039	0,014	0,019	0,029	0,251	0,076	0,124	0,055	0,1		-0,766	-1,488
KoRV_AF151794.2_Hanger	0,429	0,429	0,429	0,256	0,207	0,197	0,251	0,319	0,01	0,324	0,675	0,804	0,147	0,445		0,74
KoRV_AB721500	0,19	0,19	0,19	0,116	0,117	0,115	0,318	0,146	0,863	0,139	0,493	0,246	0,087	0,139	0,461	

pol (1126 codons)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pci_SN265		-0,293	-0,293	-0,618	-0,618	-1,161	-0,618	-1,175	-1,461	-1,828	-1,422	-1,924	-1,202	-0,277	-1,906	-1,681
Pci_QMJ6480	0,77		0	1,023	1,47	-0,614	1,023	-0,657	-1,137	-2,017	-1,146	-1,707	-0,989	0,013	-1,66	-1,431
Pci_582119	0,77	1		1,023	1,47	-0,614	1,023	-0,657	-1,137	-2,017	-1,146	-1,707	-0,989	0,013	-1,66	-1,431
Pci_MCZ_12454	0,538	0,308	0,308		1,03	-0,283	0	-0,308	-1,344	-2,142	-1,294	-1,85	-0,989	0,013	-1,791	-1,564
Pci_MCZ8574	0,538	0,144	0,144	0,305		-0,283	1,03	-0,308	-1,13	-2,007	-1,153	-1,712	-0,847	0,18	-1,671	-1,444
Pci_um3435	0,248	0,541	0,541	0,778	0,778		-0,283	-1,342	-1,322	-1,886	-1,009	-1,573	-0,847	-0,285	-1,542	-1,314
Pci_maex1738	0,538	0,308	0,308	1	0,305	0,778		-0,308	-1,344	-2,142	-1,294	-1,85	-0,989	0,013	-1,791	-1,564
1Pci_SN265	0,242	0,513	0,513	0,759	0,759	0,182	0,759		-1,323	-2,092	-1,284	-1,816	-1,18	-0,292	-1,772	-1,548
1Pci_QMJ6480	0,147	0,258	0,258	0,182	0,261	0,189	0,182	0,188		-1,68	-0,514	-1,269	-0,977	0,193	-1,239	-0,95
1Pci_582119	0,07	0,046	0,046	0,034	0,047	0,062	0,034	0,039	0,096		-1,473	-0,635	-2,112	-1,556	-0,715	-1,35
1Pci_MCZ_12454	0,158	0,254	0,254	0,198	0,251	0,315	0,198	0,202	0,608	0,143		0,286	-1,232	-0,333	0,009	0,011
1Pci_MCZ8574	0,057	0,09	0,09	0,067	0,089	0,118	0,067	0,072	0,207	0,527	0,775		-1,841	-1,126	-0,962	-0,496
1Pci_um3435	0,232	0,324	0,324	0,324	0,399	0,399	0,324	0,24	0,331	0,037	0,22	0,068		-1,132	-1,822	-0,937
1Pci_maex1738	0,782	0,989	0,989	0,989	0,858	0,776	0,989	0,771	0,848	0,122	0,74	0,263	0,26		-1,121	-0,818
KoRV_AF151794.2_Hanger	0,059	0,1	0,1	0,076	0,097	0,126	0,076	0,079	0,218	0,476	0,993	0,338	0,071	0,264		-0,607
KoRV_AB721500	0,095	0,155	0,155	0,12	0,151	0,191	0,12	0,124	0,344	0,18	0,992	0,621	0,351	0,415	0,545	

env (657 codons)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pci_SN265		-0,202	0,012	0,012	0,583	0,584	0,018	-1,29	-1,108	-1,108	-1,102	-1,29	-1,108	-1,436	-0,923	-1,29
Pci_QMJ6480	0,84		0,834	-0,2	1,354	0,861	-0,199	-1,089	-0,912	-0,912	-0,913	-1,089	-0,912	-1,24	-0,743	-1,089
Pci_582119	0,991	0,406		0,992	2,165	2,142	0,582	-1,12	-1,12	-1,12	-0,905	-1,12	-1,12	-1,477	-0,899	-1,12
Pci_MCZ_12454	0,991	0,841	0,323		1,434	1,011	-0,422	-0,9	-0,737	-0,737	-1,068	-0,9	-0,737	-1,089	-0,573	-0,9
Pci_MCZ8574	0,561	0,178	0,032	0,154		2,052	0,565	-0,698	-0,511	-0,511	-0,509	-0,698	-0,511	-0,919	-0,51	-0,698
Pci_um3435	0,56	0,391	0,034	0,314	0,042		0,013	-1,091	-0,888	-0,888	-0,886	-1,091	-0,888	-1,266	-0,887	-1,091
Pci_maex1738	0,986	0,843	0,562	0,674	0,573	0,99		-1,273	-1,099	-1,099	-1,085	-1,273	-1,099	-0,699	-1,098	-1,273
1Pci_SN265	0,199	0,278	0,265	0,37	0,486	0,277	0,205		1,02	1,02	1,036	0	1,02	-0,664	1,449	0
1Pci_QMJ6480	0,27	0,364	0,265	0,463	0,61	0,376	0,274	0,31		0	1,38	1,02	0	-1,045	1,067	1,02
1Pci_582119	0,27	0,364	0,265	0,463	0,61	0,376	0,274	0,31	1		1,38	1,02	0	-1,045	1,067	1,02
1Pci_MCZ_12454	0,272	0,363	0,367	0,288	0,612	0,378	0,28	0,302	0,17	0,17		1,036	1,38	-0,3	1,723	1,036
1Pci_MCZ8574	0,199	0,278	0,265	0,37	0,486	0,277	0,205	1	0,31	0,31	0,302		1,02	-0,664	1,449	0
1Pci_um3435	0,27	0,364	0,265	0,463	0,61	0,376	0,274	0,31	1	1	0,17	0,31		-1,045	1,067	1,02
1Pci_maex1738	0,154	0,217	0,142	0,278	0,36	0,208	0,486	0,508	0,298	0,298	0,765	0,508	0,298		-0,651	-0,664
KoRV_AF151794.2_Hanger	0,358	0,459	0,37	0,568	0,611	0,377	0,275	0,15	0,288	0,288	0,088	0,15	0,288	0,516		1,449
KoRV_AB721500	0,199	0,278	0,265	0,37	0,486	0,277	0,205	1	0,31	0,31	0,302	1	0,31	0,508	0,15	

The test statistic dN-dS is shown above the diagonal. dN and dS are the values of non-synonymous and synonymous substitutions per site, respectively. The Nei-Gojobori method was used to calculate synonymous and nonsynonymous substitutions.
The probability of rejecting the null hypothesis of strict-neutrality (dN = dS) is shown below the diagonal. Values of P less than 0.05 are highlighted in gray. The variance of the difference was computed using the bootstrap method (500 replicates). Numbers listed for columns represent the same KoRV sequences numbered in the rows. The last two KoRV sequences are from GenBank; the other KoRVs are sequences from the current study.

Table S6: Variation in putative transcription factor binding sites in KoRV LTRs

	Samples								
Family	AF151794.2	AB721500.1	Pci-SN265	Pci-QMJ648	Pci-582119	Pci-MCZ12454	Pci-MCZ8574	Pci-um3435	Pci-maex1738
Core promoter initiator elements	1	1	1	1	1	1	1	1	1
Vertebrate TATA binding protein factor	1	1	1	1	1	1	1	1	1
Abdominal-B type homeodomain transcription factors	3	3	3	3	3	3	3	3	3
MAF and AP1 related factors	1	1	1	1	1	1	1	1	1
Bicoid-like homeodomain transcription factors	1	1	1	1	1	1	1	1	1
POZ domain zinc finger expressed in B-Cells	3	3	3	3	3	3	3	3	3
Brn-5 POU domain factors	2	2	2	2	2	2	2	2	2
CCAAT binding factors	1	2	1	2	1	1	2	1	1
Calcium-response elements	1	1	1	1	1	1	1	1	1
Cell cycle regulators	1	1	1	1	1	1	1	1	1
Vertebrate caudal related homeodomain protein	2	2	2	2	2	2	2	2	2
Ccaat/Enhancer Binding Protein	2	2	2	3	2	2	2	2	2
Cell cycle regulators: Cell cycle homology element	1	1	1	1	1	1	1	1	1
CLOX and CLOX homology (CDP) factors	1	1	1	1	1	1	1	1	1
CP2-erythrocyte Factor related to drosophila Elf1	0	0	1	0	0	0	1	0	0
CTCF and BORIS gene family	1	1	1	1	1	1	1	1	1
Cyclin D binding myb-like transcription factor	1	1	1	1	1	1	1	1	1
E2F-myc activator/cell cycle regulator	4	4	4	4	4	4	4	4	4
E-box binding factors	2	2	2	2	2	2	2	2	2
Estrogen response elements	2	2	2	2	2	2	2	2	2
Human and murine ETS1 factors	6	5	5	5	5	5	5	5	5
EVI1-myleoid transforming protein	0	0	0	1	0	0	0	0	1
FAST-1 SMAD interacting proteins	1	1	1	1	1	1	1	1	1
Fork head domain factors	4	5	4	5	4	4	5	4	4
Farnesoid X - activated receptor response elements	1	1	1	1	1	1	1	1	1
Growth factor independence transcriptional repressor	1	1	1	1	1	1	1	1	1
Glucocorticoid responsive and related element	1	1	2	2	1	1	1	1	1
Human acute myelogenous leukemia factors	1	1	1	1	1	1	1	1	1
Twist subfamily of class B bHLH transcription factors	3	2	2	2	2	2	3	2	2
Heat shock factors	5	5	5	5	5	5	5	5	5
Vertebrate homologues of enhancer of split complex	3	3	3	3	3	3	3	3	3
Krueppel-like C2H2 zinc finger factors hypermethylated in cancer	1	1	1	1	1	1	1	1	1
Human muscle-specific Mt binding site	1	1	1	1	1	1	1	1	1
Onecut homeodomain factor HNF6	1	1	1	1	1	1	1	1	1
HOX - PBX complexes	2	2	2	2	2	2	2	2	2
HOX - MEIS1 heterodimers	0	0	0	0	0	0	0	0	1
Ikaros zinc finger family	1	1	1	1	1	1	1	1	1
Interferon regulatory factors	1	1	1	1	1	1	1	1	1
Krueppel like transcription factors	2	2	2	2	2	2	2	2	2
LEF1/TCF	2	2	2	2	2	2	2	2	2
Myc associated zinc fingers	1	1	1	1	1	1	1	1	1
MEF3 binding sites	1	1	1	1	1	1	1	1	1
Myc-interacting Zn finger protein 1	1	1	1	1	1	1	1	1	1
Cellular and viral myb-like transcriptional regulators	3	2	3	2	2	3	2	2	2
Myoblast determining factors	1	1	1	1	1	1	1	1	1
MYT1 C2HC zinc finger protein	0	0	0	0	0	0	0	0	1
Myeloid zinc finger 1 factors	1	1	1	1	1	1	1	1	1
NGFI-B response elements, nur subfamily of nuclear receptors	1	1	1	1	1	1	1	1	1
Nuclear factor 1	0	1	1	1	1	1	1	1	1
Nuclear factor of activated T-cells	1	1	1	1	1	1	1	1	1
Nuclear factor kappa B/c-rel	1	1	1	1	1	1	1	1	1
Octamer binding protein	5	5	5	6	5	5	5	5	1
Odd-skipped related factors	0	1	1	1	1	1	1	1	1
OVO homolog-like transcription factors	1	1	1	1	1	1	1	1	5
PAR/bZIP family	3	3	3	4	3	3	3	3	3
PAX-4/PAX-6 paired domain binding sites	1	1	1	1	1	1	1	1	3
Peroxisome proliferator-activated receptor	1	1	1	1	1	1	1	1	1
Pleomorphic adenoma gene	3	3	3	3	3	3	3	3	1
v-ERB and RAR-related orphan receptor alpha	1	1	1	1	1	1	1	1	2
SWI/SNF related nucleophosphoproteins with a RING finger DNA binding motif	1	1	1	1	1	1	1	1	3
RXR heterodimer binding sites	2	2	2	2	2	2	2	2	2
Spalt-like transcription factor 2	1	1	1	1	1	1	1	1	1
Vertebrate steroidogenic factor	1	1	1	1	1	1	1	1	3
Sine oculis (SIX) homeodomain factors	1	1	1	1	1	1	1	1	1
Vertebrate SMAD family of transcription factors	0	0	1	0	0	0	1	0	1
SOX/SRY-sex/testis determinig and related HMG box factors	2	2	2	2	2	2	2	2	2
GC-Box factors SP1/GC	1	1	1	1	1	1	1	1	1
Testis-specific bHLH-Zip transcription factors	1	1	1	1	1	1	1	1	2
Signal transducer and activator of transcription	2	2	2	2	2	2	2	2	1
Motif composed of binding sites for pluripotency or stem cell factors	3	3	4	3	3	3	3	3	1
TEA/ATTS DNA binding domain factors	1	1	1	1	1	1	1	1	2
X-box binding factors	1	1	1	1	1	1	1	1	3
Y-box binding transcription factors,	1	1	1	1	1	1	1	1	1
C2H2 zinc finger transcription factors 2	1	1	1	1	1	1	1	1	1
C2H2 zinc finger transcription factors 3	0	0	1	0	0	1	0	0	0
C2H2 zinc finger transcription factors 10	1	1	1	1	1	1	1	1	1
ZF5 POZ domain zinc finger	0	1	1	1	0	1	0	0	1
Two-handed zinc finger homeodomain transcription factors	2	2	2	2	2	2	2	2	2
Retroviral CCAAT binding factors	2	2	2	2	2	2	2	2	2
Retroviral PolyA Downstream signal	1	1	1	1	1	1	1	1	1
Retroviral PolyA signal	1	1	1	1	1	1	1	1	1
Retroviral upstream element	2	2	2	2	2	2	2	2	2

Grey shading indicates sequences with differing numbers of predicted TFBs for a given domain

Table S7: KoRV flanking sequences common to two or more koalas

	Sequence	Total	Pci-SN265	Pci-QMJ6480	Pci-582119	Pci-MCZ12454	Pci-MCZ8574	Pci-um3435	Pci-maex1738	Pci-SN404(WG (1/16))	Pci-SN404	Pci-SN248	Pci-SN345	Pci-157	Pci-106	Pci-182
' flanking sequences																
PCI -5'Flank-1	5'-GAAGGAAGGAAGGAAGGAAGGAAGGAA-3'	5		+		+	+		+	+						
PCI -5'Flank-2	5'-GTTCCCTGTCTTTACAAACTGYAAAAGAAAGAAAACGAGGAA-3'	3	+	+				+								
PCI -5'Flank-3	5'-CTGCAAAAGAAAGAAAATGGGGGAA-3'	2						+	+							
PCI -5'Flank-4	5'-AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAATGGGGGAA-3'	2				+		+								
PCI -5'Flank-5	5'-ATTCAGAAGAAAAAATGGGGGAA-3'	2			+		+									
PCI -5'Flank-6	5'-AGCAATGMAARGARAKASAAGAAAATGGGGGAA-3'	2					+	+								
PCI -5'Flank-7	5'-AGAAGGAAATGAGAGTCSCMTGKAGCAAAGAAAATGGGGGAA-3'	2		+				+								
PCI -5'Flank-8	5'-GAGTTGGAGAAGGAAATGGGGGAA-3'	2			+	+										
PCI -5'Flank-9	5'-AGACAGGATTGGGAGGAATGAACAGATGGGGGRRCCAAAGAAAATGGGGGAA-3'	2		+					+							
PCI -5'Flank-10	5'-TGCTCTTCCGATCTTACCTTTCCAATTATACATGTTGRAGATTCAAAGAAAATGGGGGAA	3	+	+					+							
PCI -5'Flank-11	5'-AGTAACTCAGCAGGCAATTAAAGGAAATTAGAGGCAGTCTTCAAAGAAAATGGGGGAA-	3	+					+	+							
PCI -5'Flank-12	5'-ACGATCCCATTTGGGGTTTTCTTGGCAAAGAT-3'	3	+	+		+										
PCI -5'Flank-13	5'-GTCAAAAGAGAAAATAGAATAAATGGGGGAA-3'	2				+		+								
PCI -5'Flank-14	5'-AGGAAAGGATCCATGTGCAGCAAAGAACTCGGAA-3'	2			+				+							
PCI -5'Flank-15	5'-GCTGTGGGAAGACAGGGATACTAGTGCATTGTTGGTGGAGCTATGAATCAGTACAAC-3'	2	+		+											
PCI -5'Flank-16	5'-ACCAAAACCCTTTGGGCCCTGATTGACTCAGAACAAATGTAAATAGGAATTGTTT-3'	2	+		+											
PCI -5'Flank-17	5'-TACATTATTATAAGCTGTCACTATTGCACTCATGATGCTATATATAACTATCAATCTTGG/	3	+		+	+										
PCI -5'Flank-18	5'-TTGTTCAGGACTGGATTAGACGTGTGCTCTTCGATCTRTGGGGGAA-3'	2			+				+							
PCI -5'Flank-19	5'-AAATATGAGTCAGTCCCAGGCCTTGGAAGAGCTCAAAAGGGATTTTGAGG-3'	2	+	+												
PCI -5'Flank-20	5'-TTATCCCAAAGGGAGCCTGGAAAGATATCCCAACCTTGTTAATATCAGGACTTGTCTCCA	2	+		+											
PCI -5'Flank-21	5'-CAGAAACCTTATTTGTAAAAAATTCACTTTTCTCATGGATGAACAAAGCTCTTCTGACAC	2	+	+												
PCI -5'Flank-22	5'-TGGACAGGAATTTCAGCCTTACAATTTAAAACGCAAAAATCTACCCCAGAAACAAGGAA'	2	+		+											
PCI -5'Flank-23	5'-TTTGTTTATTTTAGGAAGCCACAGTAAGTCATAAAAGGGTGCAGC-3'	2	+						+							
PCI -5'Flank-24	5'-AGTAACCCTAGATCAACTTAACCCCTTGTTTTATAT-3'	4	+		+		+				+					
PCI -5'Flank-25	5'-CTCCGTAACAGTGATGATCATCTCTAGTGAGCATATATCTCCCAGTTTTGGCCTTGTCTG/	2			+				+							
PCI -5'Flank-26	5'-ATTCTTAGAATACCTGGCTTCCTTCAAGGTAAGCCCCCTTCCTATTTCTACATGAAGT-3'	3	+	+	+											
PCI -5'Flank-27	5'-CTTCAGCTGGTACGCTAGGCTTTGGAGATTTAACATGAGAAGGGTGAGTCAGTAAGGT-3	3	+					+			+					
PCI -5'Flank-28	5'-ACATGGTCTTTTCCTTTTAGGGGTTCAGATGCCATCCCCATCATACCCACTGCAACACCT/	2	+		+											
PCI -5'Flank-29	5'-TTGATCTGAGGTCCTCCTGACTTCAGGGCTGGGGCTCTATCCACTGCACCACCTGGCTGC	2	+		+											
PCI -5'Flank-30	5'-CCCCCATCCTCCTCCTCCTGGTCCTTGCCTAGTGTCAGCTTCCAAACACTCTAGTAGAGT(	2	+		+											
PCI -5'Flank-31	5'-ACATGAGAGCAGCCTGGTTTGAACTTCCTCTGGTTCTTTTCAGCTTCAGTGCAGGAA-3'	2			+		+									
PCI -5'Flank-32	5'-CTAATATGACAAAAAAAGAAAATGGGGGAA-3'	2			+				+							
' flanking sequences																
PCI -3'Flank-1	5'-TGTGAACCCTGAGCAAATCACTTAACCCCATTGCCTAGCCAAAAAAAAAGCAAAACAAAA.	4	+	+	+				+							
PCI -3'Flank-2	5'-GTCTGTACTCTGGACTCCATTCTCTATACTTTCACCATTTACCCCTGGGCCGGGCTCGTA(	2	+			+										
PCI -3'Flank-3	5'-AGACCATAACGAGAGGGGAGTTCAGGGGAACAAGAACAGAAGTGGGGTTCTGGTGGTG.	5	+	+				+	+				+			
PCI -3'Flank-4	5'-CTATATTTCAAAAGATCTTTATTTTCACCAACAATGTATACATCCTACAGAAATCAATGT(	3	+	+	+											
PCI -3'Flank-5	5'-TCADATHGWTTCATTATACTCTGTTTCTGCGCTCAGGATTCTCTCCTTTCAACAGGCCG-3	2	+					+								
PCI -3'Flank-6	5'-CTGAGAAAGGTGGAGCTAAGCAGTATCTGTAGTTTTTCTCTTTTAGCAGTGAAAAAA-3'	2	+			+										
PCI -3'Flank-7	5'-TTTTCATTTACTCAGATTTCAGTTTCCTATTCTCGTAATCATTGTGCATATTTTTCCT-3'	2	+				+									
PCI -3'Flank-8	5'-AGACCAAACACTTGACACTTAGTAGCTGTGTGACCCTGGGCAAGTCA-3'	3		+					+	+						
PCI -3'Flank-9	5'-GAAGCAGCATTTGTGGGATCCTTTGGAAAATACTGTGAGAATAAC-3'	2	+		+											
PCI -3'Flank-10	5'-CAACCAAGCAATTAAAGCCATCAACAGCAAGSSAGAGCATGTGG-3'	4	+		+			+	+							
PCI -3'Flank-11	5'-CGAGGAGTTTGTTTGCTAAAGGATAGGAGGTGAGGGGCACA-3'	2	+	+												
PCI -3'Flank-12	5'-GTGCCTCTTATATGACAGGACTATGTAGGCACCATGTCTTTATTCCTGCTTGCAATTATA1	5	+	+				+	+			+				
PCI -3'Flank-13	5'-CTGTGAATGACCTAGCTATTCTCAGCAATGCAATGATCCAAGACCATTCTGAAGAACTCA	2	+	+												
PCI -3'Flank-14	5'-CTACATTCTGACATTTCTAGTCTTCTCCCGTTTACTTCCCTATACATACACTACCATTACC	3	+						+			+				
PCI -3'Flank-15	5'-ATAGAACATTCTCAAAGTTGGACAAACTCTCAGAAATTTGTTTGCATAAATATTCATCCG	5	+	+					+		+	+				
PCI -3'Flank-16	5'-GTACGATGAATCCAAAGCTCTTTGCTTTTTTCATCTAATCTTGTCTATCCTCCTGTTGATG	4	+	+	+							+				
PCI -3'Flank-17	5'-CTATCTTGTTTTATTTTAAAAGTCTGAAAAATACAGCAGCGTTGGAGACATAGATATGCA	5	+		+				+		+		+			

+	Light grey represents insertion sites found in two koalas
+	Middle grey represents insertion sites found in 3 kolas
+	Dark grey represents insertion sites found in four or more koalas

Chapter III

Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without a reference genome

3.1 Summary

Retroviral integration into the host germline results in permanent viral colonization of vertebrate genomes. The koala retrovirus (KoRV) is currently invading the germline of the koala (*Phascolarctos cinereus*) and provides a unique opportunity for studying retroviral endogenization. Previous analysis of KoRV integration patterns in modern koalas demonstrate that they share integration sites primarily if they are related, indicating that the process is currently driven by vertical transmission rather than infection. However, due to methodological challenges, KoRV integrations have not been comprehensively characterized, nor have historical trends been examined. To overcome these challenges, we applied and compared three target enrichment techniques coupled with next generation sequencing (NGS) and a sequence-clustering based computational pipeline to determine the integration sites for 10 museum Northern Australian koala samples collected between the 1870s and late 1980s. Although three enrichment methods each exhibited bias in integration site retrieval, hybridization capture based methods performed best. The results suggest that the proportion of KoRVs shared among unrelated koalas has increased over the last 140 years.

3.2 Introduction

Vertebrate endogenous retroviruses (ERVs) descend from exogenous retroviruses that infected the ancestral germ line and have been subsequently transmitted vertically from parent to offspring through Mendelian inheritance (1). ERVs comprise up to 8-11% of vertebrate genomes (2, 3). Most ERVs colonized their host genomes millions of years ago (4, 5) making it difficult to study the process of retroviral invasion. The koala retrovirus (KoRV) spreads both horizontally and vertically among koalas (*Phascolarctos cinereus*) (6, 7, 8), and it is still in the process of endogenizing, unlike most other described ERVs (9). Therefore, KoRV provides a unique opportunity to study the processes underlying retroviral endogenization in real time. Historical DNA analysis from museum koala samples collected during the 19th and 20th centuries demonstrated that KoRV was already ubiquitous in northern Australia by the 19th century (10), and that its genome has remained strongly conserved (11). In contrast, KoRV integration sites among individuals are highly variable (11, 12).

The integrated provirus has identical sequences at the 5’ and 3’ ends of the proviral genome, which are termed the long terminal repeats (LTR). Distribution of retroviral integration sites in the host genome is generally regarded as non-random (13), for example, fewer integration events are observed near transcriptional hotspots (14). Integration site preference is associated with the viral integrase (15) and host chromosomal features (16). Retroviruses belonging to the same group tend to exhibit similar integration site preference (17). Despite these tendencies, integration of a specific retrovirus at a specific site is still a random event. Most individuals in a host population will share older ERV integration sites as they become fixed in the population over time through drift, as is now true for most human endogenous retroviruses (18). In contrast, if a retrovirus endogenized very recently, the integration site will be rare among all but the most closely related individuals such as offspring. This is the case for KoRV integrations, which appear to be largely unique to related koalas (11, 12). However, previous studies have not attempted a comprehensive survey of integration sites. The focus of the current study was to evaluate methods that may comprehensively characterize retroviral integrations and which could be applied to museum samples to examine historical trends in the frequency of shared or unique KoRV integration sites.

Inverse PCR has conventionally been used for retrieving retroviral integration sites (19). Methods such as rapid amplification of cDNA ends (RACE), ligation-mediated PCR, Linker-selection-mediated PCR, linear amplification–mediated PCR and genome walking (20; 21; 22; 23; 24; 25) have also been used. However, it is unclear if they can comprehensively detect integration sites particularly due to potential primer-target mismatch, and they have never been applied to ancient DNA (aDNA). DNA extracted from museum samples has the characteristics of aDNA, e.g., it is heavily fragmented and damaged, and in low concentration (26). The DNA degradation, fragmentation and contamination that occurs post mortem makes aDNA research difficult (27; 28), often preventing the use of conventional molecular biological methods such as PCR.

To overcome the limitations of working with historical DNA, we applied three target enrichment techniques followed by high-throughput Illumina sequencing. The three techniques, Primer Extension Capture (PEC) (29), Single Primer Extension (SPEX) (30), and hybridization capture (31) have been applied successfully to aDNA and could potentially be employed to determine sequences flanking targeted ERVs. Ten koala museum samples collected between the 1870s and the 1980s were successfully examined. Because no assembled koala genome is currently available, a reference-independent computational pipeline was established. The results are discussed in terms of performance of the three methods in retrieving KoRV integration sites per koala.

3.3 Materials and methods

3.3.1 Samples and ancient DNA extraction

A total of thirteen museum samples were examined as described in Table 1. DNA extractions were performed in the aDNA laboratory of the Department of Wildlife Diseases of the Leibniz Institute for Zoo and Wildlife Research in Berlin, Germany. The laboratory is dedicated to aDNA work and has never been used for molecular work on modern samples. The room is UV irradiated 4 hours every night by ceiling-mounted UV lights. All work performed in the facility follows procedures designed to minimize the possibility of contamination, such as use of laminar flow hoods and use of protective clothing to avoid sample contamination.

Approximately 250 mg of skin tissue per specimen were extracted using a silica-based extraction kit for aDNA (GENECLEAN Ancient DNA Extraction Kit, MP Biomedicals, USA). The protocol followed the manufacturer’s instructions and has been successfully applied to a variety of ancient sample types (32; 33). Mock extractions were performed with each set of koala museum specimens as negative controls during extraction. Subsequent to each extraction, the isolated DNA was further purified using a MinElute spin column (Qiagen, Hilden, Germany) as described in (34) to remove potential inhibitors for the subsequent enzymatic reactions.

3.3.2 NGS Library preparation

Illumina sequencing libraries were prepared from the extracts using a previously described protocol (35) with the following modifications: (A) All SPRI purification steps were substituted with spin column purification (MinElute PCR purification kit, Qiagen). (B) Adapter concentration in the ligation reaction was reduced to 0.2 mM per adapter. (C) The purification after adapter fill-in was substituted by heat inactivation at 80°C for 20 min. The libraries were then used directly as template for subsequent amplification following a two-step strategy (36). A quality control strategy (37) was also applied, which consisted of a qPCR to quantify the product after each step of library amplification. The qPCR results excluded three samples from further processing for which DNA quality was too poor for analysis.

In the first round of amplification, AmpliTaq Gold, a non-proof reading enzyme, and indexing primers (Table S1) were applied, adding a distinct P7 index to each library as described in (35), 10 indices for the 10 working samples and 3 and 4 negative control indices for PEC and SPEX respectively. Adding distinct indices to each library allows for multiple samples to be sequenced in a single sequencing run. Non-proof reading enzyme allows for amplification to be performed on templates containing deoxyuracils, which are common with aDNA (38). Except for removal of 1 µL for qPCR as a library quality control, the remaining libraries were used as template in 100 µL PCR reactions containing 1x Taq buffer II (Applied Biosystems), 5U AmpliTaq Gold (Applied Biosystems), 250 mM each dNTP and 100 nM each indexing primer. Cycling conditions followed manufacturer’s instructions: The pre-denaturation step lasted 12 min at 95°C, followed by 12 cycles of denaturation at 95°C for 20 s, annealing at 60°C for 30 s and elongation at 72°C for 40 s, with a final extension step of 72°C for 5 min. PCR products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany).

In the second round of amplification, 5 µl of the purified PCR product from the first round PCR was used as template for a second PCR. This involved 50 µL reactions containing Herculase II Fusion DNA Polymerase (Agilent Technologies Catalog 600677), which has proof reading activity, and primers IS5 and IS6 (35) at a final concentration of 400 nM each. Cycling conditions included an activation step of 3 min at 95°C, followed by 15-20 cycles of denaturation at 95°C for 20 s, annealing at 60°C for 25 s and elongation at 72°C for 30 s, with a final extension step at 72°C for 3 min. The number of cycles used in the PCR for every sample was dependent on the concentration of each of the libraries as determined by the qPCR assay. The PCR amplified libraries were then purified using the QIAquick PCR purification kit. Each library was separately used in subsequent PEC and hybridization capture experiments.

3.3.3 Bait preparation and integration site enrichment

Three methods were compared for retrieving integration sites: primer extension capture (PEC), single primer extension (SPEX) and hybridization capture. All three have been successfully applied to ancient and historical DNA samples and all are applicable to samples that would not be expected to yield results with conventional methods for integration site analysis. The same set of primers was used in PEC and SPEX experiments (Figure 1, Table S2). Because the two LTRs of a provirus are identical, the primers designed for targeting the 5’ integrations will also extend targeting the retroviral *env* gene and the primers designed for targeting the 3’ integrations will also extend targeting the retroviral gag leader sequence (Figure 1A). For both the 5’ and 3’ KoRV LTR, two 20 bp primers were developed which overlap such that the 3’ end of the first primer overlapped 8 bp with the 5’ end of the second primer (Figure 1 B primers 5.1 and 5.2 and 3.1 and 3.2 respectively). To avoid known LTR polymorphisms among KoRV proviruses, the two primers on each side of the LTR were located 17 bp from the 5’ end and 50 bp from the 3’ end of the LTRs respectively in conserved regions (Figure 1B). The baits used for hybridization capture were synthesized to generate 32 bp oligonucleotides that spanned the full length of sequence covered by primers 5.1 and 5.2 (32 bp) on the 5’ LTR and primers 3.1 and 3.2 (32 bp) on the 3’ end.

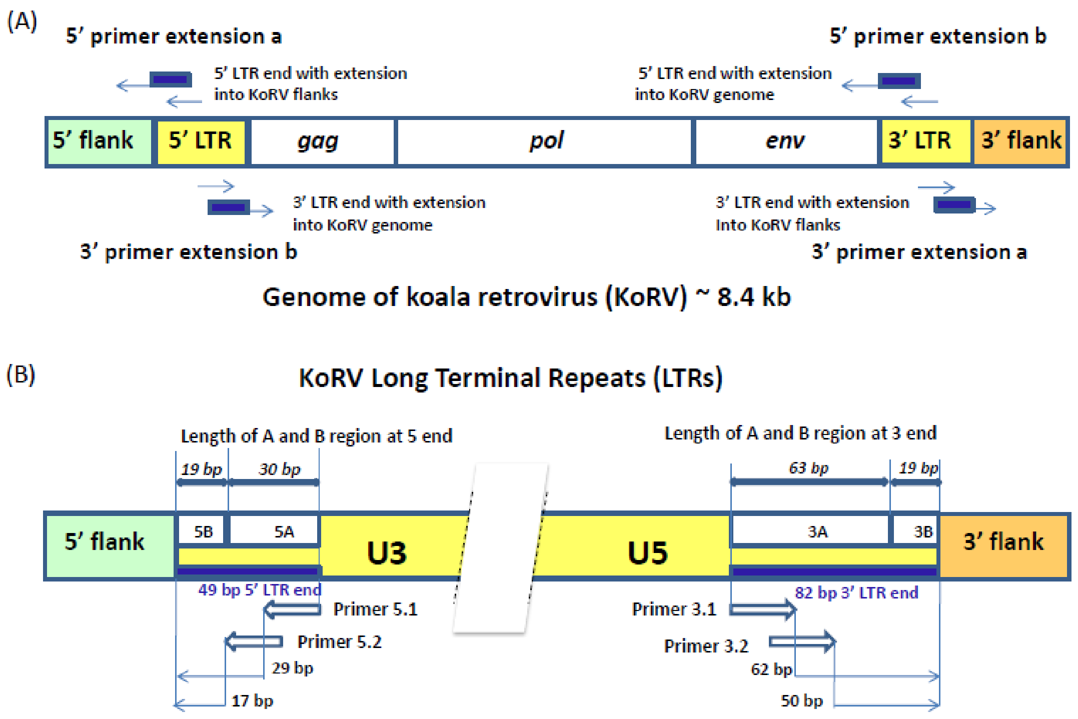

(A)
5′ primer extension a
5′ primer extension b
5′ LTR end with extension into KoRV flanks
5′ LTR end with extension into KoRV genome
5′ flank
5′ LTR
gag
pol
env
3′ LTR
3′ flank
3′ LTR end with extension into KoRV genome
3′ LTR end with extension Into KoRV flanks
3′ primer extension b
3′ primer extension a
Genome of koala retrovirus (KoRV) ~ 8.4 kb
(B)
KoRV Long Terminal Repeats (LTRs)
Length of A and B region at 5 end
Length of A and B region at 3 end
19 bp
30 bp
63 bp
19 bp
5′ flank
5B
5A
U3
U5
3A
3B
3′ flank
49 bp 5′ LTR end
82 bp 3′ LTR end
Primer 5.1
Primer 5.2
Primer 3.1
Primer 3.2
29 bp
17 bp
62 bp
50 bp

Figure 1. Experiment design for the identification of KoRV integration sites.

Panel A illustrates that the genome of the koala retrovirus (KoRV) has two identical long terminal repeats (LTRs) on both ends. The primers or baits can bind to both LTRs, so there should be two categories of products: A) products extending into the flanks from primer extension a; B) products extending into the middle of KoRV genome from primer extension b. In principle, there should be equal number of sequences for the two categories. Panel B indicates that the KoRV LTRs contain three components, U3, R and U5. For SPEX, primers were partially nested. All primers are 20 bp long and there is a 8 bp-overlap between the inner primers (3.1 and 5.1) and outer primers (3.2 and 5.2) respectively. To avoid known polymorphisms in the LTR, the 3' end of outer primers are 17 bp from the 5' end of LTR and 50 bp from the 3' end of LTR. Since the 5' LTR and 3' LTR of the same KoRV are identical products can also extend into the KoRV genome. The 5' and 3' flanks can be distinguished by their linked LTR end, with the 5' flank linked to 5' LTR and 3' flank linked to 3' LTR. Considering the longest deletion found at the end of LTR is 19 bp, the LTR end was divided into two segments for subsequent computational identification: the B region representing the last 19 bp of the LTR, and the A region representing the rest of LTR end.

3.3.4 Primer Extension Capture (PEC)

Indexed libraries were pooled in equi-molar ratios for primer extension following a published protocol (29). After each step, 1 µL of the product was quantified by qPCR. To minimize the amplification bias, each of the captured products was amplified in triplicate, using 5 µL of the captured product as template for each reaction, using the same kit and cycling conditions as described previously under NGS library preparation for second round amplification of Illumina indexed libraries, except that we ran 20 cycles of amplification for all samples. Amplified captured libraries were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany) and eluted in 50 µl of elution buffer (EB) and used as template for a second round of PEC.

3.3.5 Single Primer Extension (SPEX)

The SPEX experiments generally followed a published protocol (30) using DNA extracts prior to Illumina library construction with three modifications: (1) Illumina sequencing adaptors were attached to the 5' end of the primers used in in the first round of partially nested PCR; (2) MyTaq HS Mix (Bioline, BIO-25045) was used instead of Platinum Taq DNA Polymerase High Fidelity in the first round of a partially nested PCR; (3) only one round of a partially nested PCR amplification was performed. The nested PCR products were then quantified by qPCR and indexed using Illumina indexing primers (Table S1). The indexed PCR products were purified using a QIAquick PCR Purification Kit (Qiagen). The amplicons were quantified by qPCR and subjected to a second round of amplification using the same conditions as the first round. The products were purified again using the QIAquick PCR Purification Kit (Qiagen), quantified by qPCR and pooled at equi-molar ratios. All PEC and SPEX products were pooled and measured using High Sensitivity DNA chips on an Agilent 2100 Bioanalyzer, then sequenced at the National High-throughput DNA Sequencing Centre, Copenhagen, Denmark using Illumina MiSeq Reagent Kit v2 (300 cycle).

3.3.6 Hybridization capture

The amplified libraries were pooled in equi-molar ratios at final concentration of 2 µg. An established protocol was followed (31) except that synthesized oligonuceotide baits were used instead of PCR products and the EB volume for final elution using Qiagen MiniElute column was 20 µL instead of 15 µL. After 2 days of hybridization and subsequent elution steps, 1 µL of the final eluate was quantified by qPCR and 5 µL (in total 15 µL) was amplified in triplicate using the same kit and cycling conditions as described in the NGS library preparation for second round amplification of Illumina indexed libraries. The pooled PCR products were purified using the QIAquick PCR Purification Kit and was measured using the Tapestation 2200 (Agilent Technologies Catalog G2964AA). Hybridization capture libraries were sequenced at the National High-throughput DNA Sequencing Centre, Copenhagen, Denmark using Illumina MiSeq Reagent Kit v2 (300 cycle).

3.3.7 Preprocessing of sequence data

Adaptor sequences occur at the ends of each sequence read. Adaptor sequences were removed from sequence reads using cutadapt-1.2.1 (39), and quality trimming was performed using Trimmomatic-0.22 using default settings (40). The paired forward and reverse sequence reads were merged using Flash-1.2.5 where possible (41), and both the merged and unmerged reads were used for further analyses. PCR duplicates (clonality in the sequencing data) with 100% sequence identity were removed using cd-hit-v4.6.1 (42).

3.3.8 Identification of KoRV integration sites

Figure 2 and Table S3 summarize the computational pipeline used for the identification of KoRV integration sites. For its implementation, both existing software and customized perl scripts were used that made use of BioPerl (43). Because the nested primers or bait were designed near the ends of LTR, the primer extension products would include either the first 49 bp of the 5’ LTR or the last 82 bp of the 3’ LTR, which are designated “LTR ends” in Figure 1A. All sequences with a KoRV flank should contain an LTR end, as a result of the primer extension (Figure 1B). Therefore, KoRV integration sites could be identified as the sequence beyond the KoRV LTR end, since all integration sites would have this sequence. However, due to DNA degradation in museum samples, some primer extension product may not have a complete LTR end. Furthermore, minor deletions at the end of the integrated LTRs may be present (44); for example, a 19 bp deletion was found in a KoRV provirus (12). To get around these potential issues, identification of the LTR ends relied on sequentially selecting sample sequences that contain defined LTR segments; this was done in separate steps for the 5’ and 3’ flank-containing sequences. The LTR end was divided into two segments, designated A and B (Figure 1b): the B segment corresponds to the last 19 bp of the LTR and is referred to as 5B or 3B in the 5’ and 3’ LTR ends, respectively. The A segment is the remaining section of the LTR end, which has a length of 30 bp in the 5’ end (5A) and 63 bp in the 3’ end (3A).

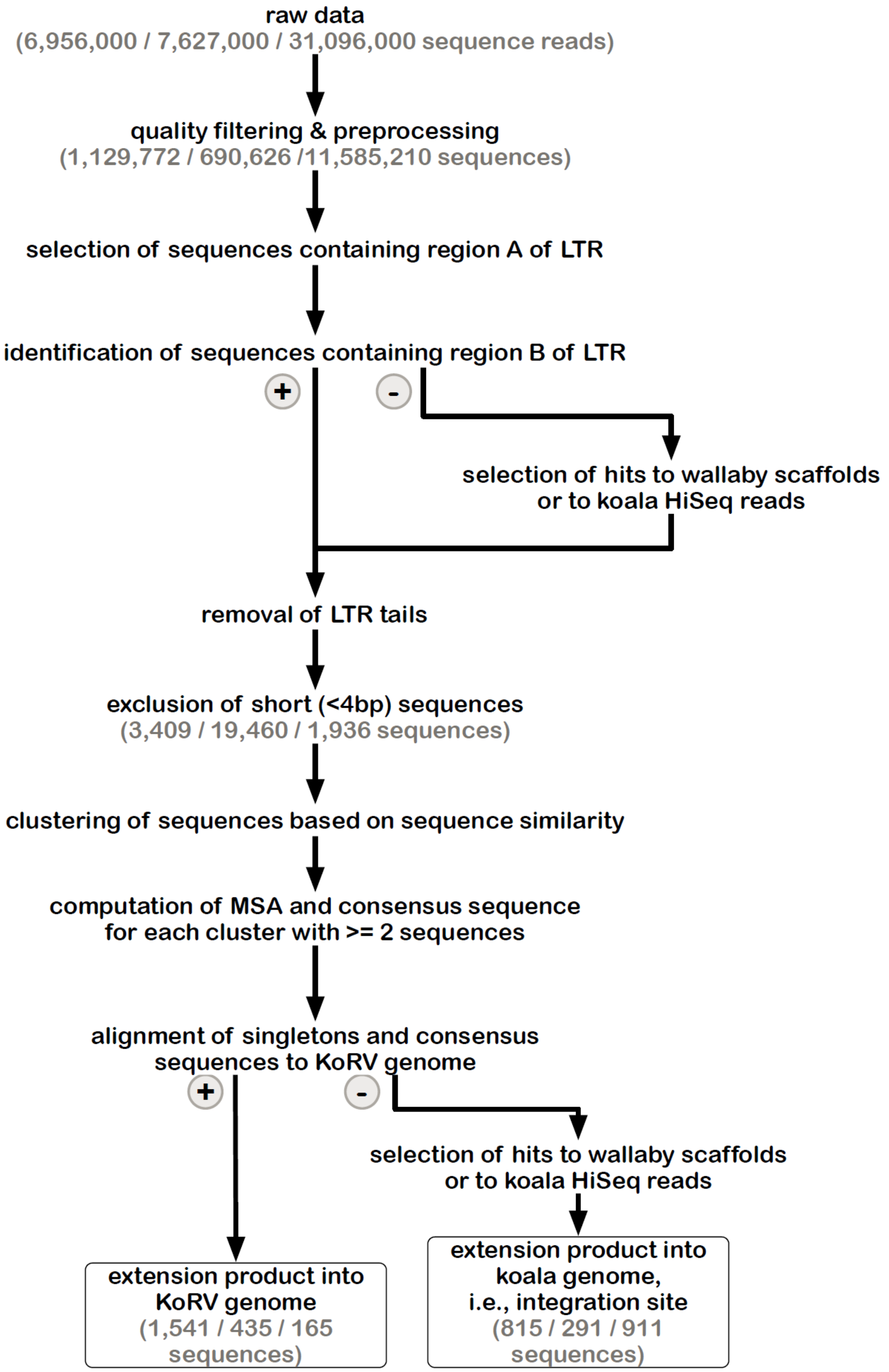

raw data
(6,956,000 / 7,627,000 / 31,096,000 sequence reads)
quality filtering & preprocessing
(1,129,772 / 690,626 /11,585,210 sequences)
selection of sequences containing region A of LTR
identification of sequences containing region B of LTR
+
-
selection of hits to wallaby scaffolds or to koala HiSeq reads
removal of LTR tails
exclusion of short (<4bp) sequences
(3,409 / 19,460 / 1,936 sequences)
clustering of sequences based on sequence similarity
computation of MSA and consensus sequence for each cluster with >= 2 sequences
alignment of singletons and consensus sequences to KoRV genome
+
-
selection of hits to wallaby scaffolds or to koala HiSeq reads
extension product into KoRV genome (1,541 / 435 / 165 sequences)
extension product into koala genome, i.e., integration site (815 / 291 / 911 sequences)

Figure 2. Bioinformatic pipeline for identification of KoRV integration sites. The pipeline was run separately for each data set obtained by three different techniques. For the key steps, the number of

sequences retained is indicated in parentheses for each technique in this order from left to right: PEC, SPEX and hybridization capture. After processing NGS reads, KoRV integration sites were identified in a two-step analysis of KoRV LTR ends, next to the host DNA flanking KoRV. The first round selection targeted the A region of the LTR end and its output was used for subsequent identification of the B region. The LTR ends of all sequences were trimmed off and only sequences longer than 4 bp were considered. Using a sequence clustering approach, unique vs shared integration sites were sorted into clusters. Sorting also included singleton clusters and non-singleton clusters. The consensus of each non-singleton cluster was computed using multiple sequence alignment. These consensus sequences and singleton sequences were queried against wallaby genomic scaffolds and koala Illumina Hiseq reads to determine whether they represented KoRV flanking sequences. At the same time extension products into the KoRV genome were identified.

Initially, sequences containing either of the two A regions in the KoRV LTR end (5A or 3A in Figure 1B) were identified. For this step, optimal local pairwise sequence alignments (Smith-Waterman, EMBOSS) were computed between each sample sequence and the A region in either 5’ or 3’ LTR end. Sequences were kept for further analysis if they could be aligned to at least 20 bp of the 30 bp 5A segment with at least 90% identity, or if they could be aligned to at least 43 of the 63 bp 3A segment with at least 90% identity (Table 2). Sequences not passing these criteria were discarded as artifacts. The LTR ends of all sequences meeting these criteria were trimmed to the last 19 bp and then used for further analyses.

From these sequences, B segments of either 3’ or 5’ LTR ends were identified (3B or 5B in Figure. 1B). For this step, optimal local sequence alignments were computed between each of the trimmed sequence and the B segment in either the 3’ or the 5’ LTR end. Only sequences that could be aligned to at least 12 bp of the 19 bp long B segment (3B or 5B) with at least 80% (Table 2) identity were selected. The last 19 bp of LTR ends were trimmed from all sequences meeting the selection criteria, leaving LTR free KoRV flanks or KoRV genomic DNA adjacent to the LTR.

All sequences that contained the A region, but for which the B region was not detected using the pairwise alignment strategy, were then subjected to another test. Specifically, these sequences were used as queries for two separate local database searches using BLAST (48). Such sequences represent LTRs that have suffered deletions at the end, a common occurrence in proviruses. One search was against HiSeq sequencing data of a koala from Queensland, Australia with 100X coverage. The data represent raw Illumina sequences and are not annotated or assembled. After adaptor and quality trimming, 6.469 billion reads from this koala, with a mean length of 78 bp, were used for this step. Sequences were considered KoRV integration sites when their non-LTR portion could be aligned with greater than 90% identity to the koala reads over 60% length of the sample sequence. A second search was against the Tammar wallaby (*Macropus eugenii*) genome (GenBank: ABQO000000000.2), which represents the closest related species to koala for which a genome has been assembled (46). Although the wallaby and koala lineages diverged more than 50 Mya (47), we expected that some of the koala genomic DNA (flanking KoRV) could be aligned to the homologous wallaby regions. Sequences with at least 70% identity over 50% length of the sample sequence to the wallaby genome were therefore considered to be KoRV integration sites. For the sequences with a match to the wallaby scaffolds or the koala data, the LTR sequences were trimmed and were then concatenated with the KoRV flanks (obtained in previous steps) for further analysis.

3.3.9 Sorting of sequences representing different integration sites

All sequences with matches to the different segments of the 3' and 5'LTR ends and/or to wallaby scaffolds or koala HiSeq data from each of the enrichment techniques were collected. The sequences matching 3' and 5' LTR ends were kept separate, resulting in a total of six different data sets for further analysis (two data sets each for the PEC, SPEX and capture). LTR ends had been removed from all sequences in these data sets. Before using these sequences to identify shared and unique integration sites, KoRV flank sequences shorter than 4 bp (the typical length of a KoRV target site duplication) were defined as “short insertion sites” (Table 3) and were excluded from further analysis; only KoRV flanks of 4 bp or longer (representing the length of target site duplication as identified in Ishida, Y., et al 2015) were used for further analysis. At this stage, the PEC data had 392 5' flank sequences and 2,347 3' flank sequences; the SPEX data 6521 5' flank sequences and 9,200 3' flank sequences; and hybridization capture 1,158 5' flank sequences and 28 3' flank sequences.

A clustering approach was used to sort all sequences in each of the six data sets into groups of similar sequences; each cluster representing a unique integration site. Sequences that did not share significant similarity with any other sequences in the input file were called singletons. For each of the six data sets, all-against-all BLAST comparisons were run, and the BLAST output was used as input for clustering using TRIBE-MCL (48), separately for each data set. Different combinations of E-values (all against all BLAST) and inflation values (TRIBE-MCL) were used for this step and the optimal parameter combination for each data set was evaluated. For all combinations of E-values and inflation values, multiple sequence alignments were computed for all clusters using MAFFT v7.127b (49). To assess the quality of the clustering, alignments of the 30 largest clusters of each clustering result were visualized in jalview (50) and were checked by eye. An alignment was considered high quality if the total number of mismatches and gaps in every sequence of the alignment was no more than

10% of the sequence length. If all 30 clusters were evaluated to be of high quality, the sequence was further analyzed. The parameter combinations for optimal clustering and related all against all BLAST are listed in Table 4.

Singletons and non-singleton clusters containing sequences derived from a single individual koala were considered to represent unique integration sites. Clusters containing sequences shared by more than one koala were considered to represent shared integration sites. A consensus sequence was computed from the alignment of each non-singleton cluster. Singletons and consensus sequences were then further evaluated first by computing pairwise alignments between these sequences and the *gag* or *env* part of KoRV genome (Figure 1A) (GenBank: AF151794.2). The sequences that could be aligned to the KoRV genes with at least 90% identity and of any length were categorized as primer extension or flank capture within the KoRV genome. The LTR sequences at the 5' and 3' ends of the KoRV genome are identical or nearly so and therefore 50% of the PCR products should extend into the KoRV genome (Figure 1A). Sequences that could not be mapped to KoRV genome were potential KoRV integration sites and were evaluated further. For such sequences, a length filtering was performed with threshold of 15 bp, since this is the minimum length that can be effectively identified in BLAST. The sequences longer than 15 bp were first used as query in BLAST to search against the koala shotgun Hiseq data; they were also mapped to wallaby genome (GenBank: ABQO000000000.2) in Geneious version 6.18 (http://www.geneious.com, 51). Identified sequences for either one of the two computations were considered to be KoRV integration sites. Sequences shorter than 15 bp are too short for efficient mapping or BLAST; however, because they contained an LTR end, were included in the KoRV specific enrichment statistics (Table 3), although they were not further analyzed.

3.3.10 Pairing of 5' and 3' integration site to one KoRV provirus

Ishida, et al 2015 identified the length of the retroviral target site duplication (a stretch of host DNA directly adjacent to retrovirus which is duplicated during retroviral integration) for KoRV to be 4 bp. Based on this target site duplication length (Figure 3), all 5' and 3' integration sites were examined for shared 4 bp target site identity. Only flanks longer than 16 bp were used for matching 5' and 3' integration sites. The minimum 28 bp (32 bp minus the 4 bp target site duplication) combined length discriminated true wallaby matches from non-significant blastn results.

The paired 5'-3' integration sites were 1) mapped against the wallaby genome using the mapping tool in Geneious using default settings, where only the paired 5'-3' integration sites that could be mapped to the wallaby genome with over 70% of their total length were scored as positively identified; 2) used as query to search in the Hiseq data (a Queensland wild koala) using BLAST. Here, only the paired 5’-3’ integration sites that could be aligned with over 90% identity with the koala Hiseq reads were considered positive.

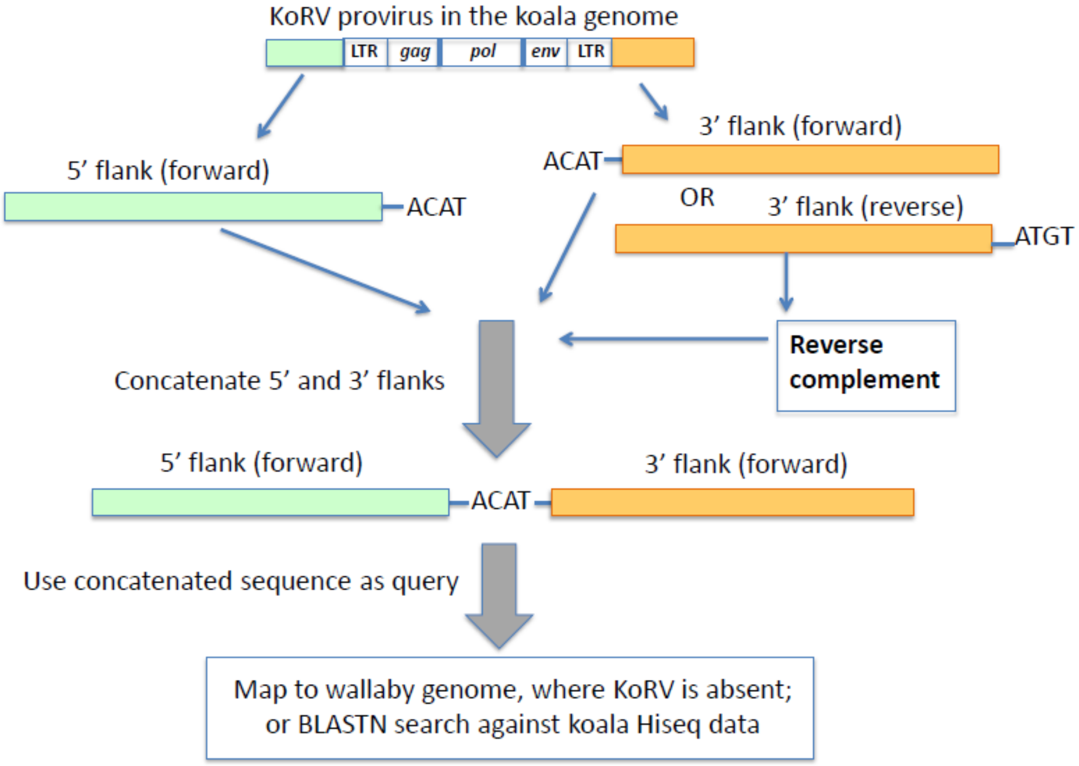

KoRV provirus in the koala genome
LTR
gag
pol
env
LTR
5’ flank (forward)
ACAT
ACAT
3’ flank (forward)
OR
3’ flank (reverse)
ATGT
Reverse complement
Concatenate 5’ and 3’ flanks
5’ flank (forward)
ACAT
3’ flank (forward)
Use concatenated sequence as query
Map to wallaby genome, where KoRV is absent; or BLASTN search against koala Hiseq data

Figure 3. Pairing of 5’ and 3’ integration sites. The first 4 bp beyond the KoRV 5’ LTR is the target site duplication (eg. ACAT in this figure), and the same 4 bp is found at the beginning of a 3’ flank (Ishida et al. 2015). One copy of the target site duplication was trimmed off and the 2 flanks were concatenated. The paired 5’-3’ integration sites were then screened against the wallaby draft genome and koala Hiseq genomic sequences.

3.3.11 Statistical analysis of shared integration sites

Statistical tests were performed to check if the occurrences of KoRV at sampled integration sites increased as the samples became younger among the 10 museum koala samples. Two logistic regression models were employed: one for 5’ integration sites and one for 3’ integration sites. Both models had the same structure. The occurrence was considered (binary: 1=presence, 0=absence) as the response variable and time as a continuous fixed effect. Because results were qualitatively similar irrespective of expressing "time" as rank or directly as years, for the sake of simplicity, only the latter was reported. The identity of koalas and of insertion sites were considered as two Gaussian random effects, making this logistic regression a Generalised Mixed effect Model (GLMM). The GLMM was fitted using the function HLfit from the R package spaMM 1.4.1 (52), considering a Binomial error structure.

The effect of time was tested by performing an asymptotic Likelihood Ratio Test (LRT) using the function anova.HLfit from the same package.

4 Results

NGS sequencing post enrichment by all three tested methods generated hundreds of thousands to millions of reads. After the pre-processing steps, 1,129,772 sequences from the PEC approach were available for further analysis, 690,626 from SPEX, and 11,585,210 from hybridization capture.

4.1 Single primer extension

Using SPEX to target the 5’ LTR flanks, 66 integration sites unique to a single koala, and 15 integration sites shared by more than one koala were identified across the 10 koala samples. Integration sites derived from consensus sequences generated from sequence clusters with at least 4 bp of sequence flanking the KoRV LTR. An additional 15,822 sequences were too short (less than 4 bp) for further biological interpretation. A total of 212 sequences contained only the KoRV genome, *env* to 5’ LTR. This is a consequence of the identical primer binding sites in the 5’ and 3’ LTRs (Figure 1A), since KoRV 5’ and 3’ LTRs are identical or nearly so (12). Thus, approximately 50% of the sequences are expected to extend from the LTR into the virus rather into the host flanking region. Sequences that extended into KoRV were categorized separately but included in the total enrichment efficiency evaluation. SPEX also identified 182 unique and 28 shared 3’ LTR flanks; an additional 1,527 sequences were too short to further analyze and 223 were found extending into the KoRV genome (Table 3).

4.2 Primer extension capture

PEC designed to identify flanking regions 5’ of integration sites detected 126 unique and 17 shared integration sites; an additional 496 sequences were too short to further characterize and 135 sequences extended into the KoRV genome. PEC targeting regions downstream of 3’ LTR integration sites identified 538 unique and 134 shared integration sites; an additional 1,806 sequences were too short to characterize further and 1,406 sequences extended into the KoRV genome (Table 3).

4.3 Hybridization capture

Using the 5’ LTR region as bait, 862 unique and 25 shared 5’ flanking regions were identified. An additional 191 sequences were too short to further characterize, while 151 sequences extended into the KoRV genome. Additionally, 24 unique and no shared integration sites were identified by hybridization using the 3’ LTR as bait. The strong bias towards the 5’ integration sites has been observed previously (11) although it is unclear why the preferential LTR enrichment occurs. Additionally, 41 sequences were too short to further characterize and 14 sequences extended into the KoRV genome (Table 3).

4.4 Summary of computational data processing

At each step of our bioinformatics pipeline, we recorded for each experiment the number of sequences that met our screening criteria. Additional information like mean length, minimum length and maximum length of sequences was also computed at each step (Table S3). Before any screening criteria were applied, PEC produced 6,956 million reads, SPEX produced 7,627 million, and hybridization capture produced 31,096 million. After pre-processing (including PCR duplicate removal) of this sequencing data, 16.24% of the initial sequencing reads were kept for PEC, 9.05% for SPEX and 37.25% for hybridization capture. Clonality was more prevalent for SPEX than for either PEC or hybridization capture.

After the first round of LTR end identification, 31,787 (2.67%) LTR positive sequences were identified for PEC, 142577 (19.94%) for SPEX and 5,648 (0.0483%) for hybridization capture. Sequences passing the second round of LTR end selection were 5,692 for PEC, 31,941 for SPEX, and 1,503 for hybridization capture. No KoRV flanks were detected in negative controls, extraction or PCR controls lacking template, for any experiment.

4.5 Cross-technique comparisons

Efficiency of target enrichment for each technique was calculated as the total number of identified integration sites divided by the total number of sequences after removal of clonality. The total number of identified integration sites included KoRV flanking sequences (including sequences shorter or longer than 4 bp) and reads extending into the KoRV genome. Sequences extending into the KoRV genome are not the desired target but because of the identical or nearly identical sequences of the 5’ and 3’ LTRs all such sequences represent correctly targeted enriched sequences.

As shown in Table 2, PEC enriched the highest total number of 3’ integration sites, 531, whereas hybridization capture enriched the most 5’ integration sites, 762. As a percentage of the total sequences retrieved, SPEX achieved the highest target enrichment efficiency (4.684%). Both PEC and hybridization capture exhibited lower enrichment percentages (0.554% and 0.0135% respectively).

Due to a phenomenon known as CapFlank (53), koala genome sequences near the integration sites may be enriched together with KoRV flanks by concatenation of library molecules on the baits. To estimate the number of such target flanks, after PCR duplicate removal all sequences were screened against the wallaby genome using BLAST. Hybridization capture exhibited the lowest efficiency of on-target enrichment (0.0135%, Table 3) and highest ratio of CapFlank enrichment (16.409%), while SPEX achieved the highest efficiency of on-target enrichment (4.684%) and lowest ratio of CapFlank enrichment (0.226%).

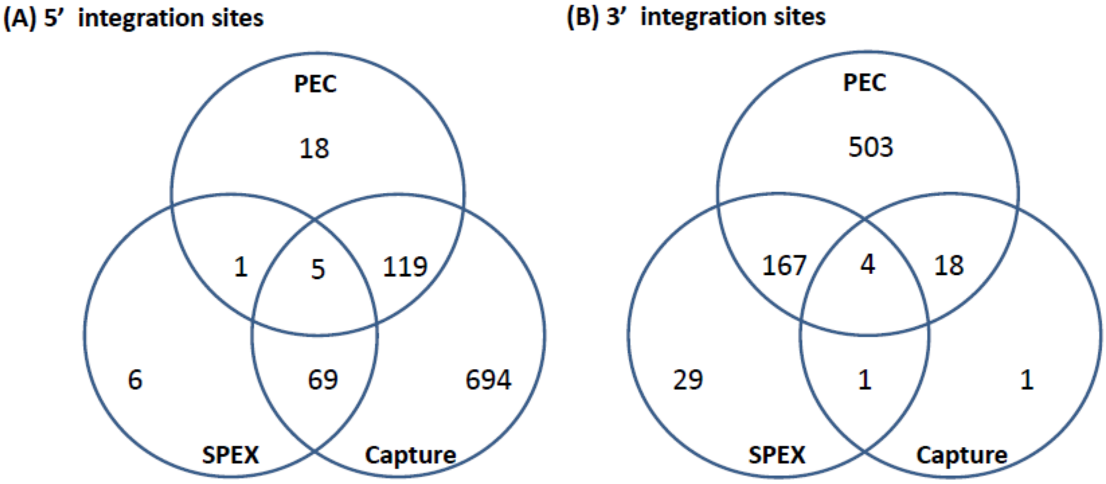


Figure 4. Venn diagrams of KoRV integration sites found by different methods. (A) For 5' integration sites, HC (hybridization capture) yielded the highest total number of integration sites (887), and covered 91.3% of the integration sites found by SPEX and 86.7% of the integration sites found by PEC. (B) For 3' integration sites , PEC yielded the highest total number of integration sites (672), and covered 81.4% of the integration sites found by SPEX and 91.7% of the integration sites found by HC (capture hybridization).

As illustrated in Figure 4, for the 5' LTR integration sites, hybridization capture yielded the highest total number of integration sites, 887, and contained 91.3% of the integration sites identified in the SPEX data set and 86.7% of the integration sites identified in PEC data set. The 3' LTR integration data followed a different profile with PEC generating the highest total number of integration sites, 672, containing 81.4% of the integration sites in the SPEX data set and 91.7% of the integration sites in the hybridization capture data set.

4.6 Shared and unique integration sites

After identical integration sites across the data sets generated by the 3 techniques were combined, 52 shared and 865 unique 5' KoRV host flanks could be identified. Shared integration sites accounted for 5.7% of the total identified using 5’ flanking host sequences, a similar percentage as estimated in previous studies (11). Among the 3’ flanking regions, 146 shared and 570 unique integration sites were identified, with shared sites accounting for 20.4% of total integration sites identified using 3’ host genomic sequences.

4.7 Pairing of 5’ and 3’ flanking regions to identify individual proviral integration sites

KoRV typically produces a 4 bp target site duplication upstream and downstream of its integration site (12). By comparing the 4 bp target site duplication, 1,690 5’ and 3’ host flanking regions were screened in the koala genome to identify potential paired flanking regions. Sixty three pairs of 5’ and 3’ KoRV integration sites were identified as originating from a same proviral loci. Of these 63 pairs, 40 were derived from a single koala (Supplement List 1), whereas 23 matches were identified by pairing 5’ and 3’ flanks identified in different koalas (Supplement list 2).

4.8 Statistical modeling of shared KoRV integration sites among 10 koalas

The proportion of 5’ integration sites that were shared with other koalas was significantly higher in more recently collected koala specimens than in specimens collected further in the past. This was true both when the influence of the identity of the koala and of the insertion site were accounted for in a statistical model (time effect in the GLMM: LRT=5.06, df=1, pv=0.0024), and on raw mean occurrence frequencies pooled across insertion sites (Spearman correlation test, rho=0.75, pv=0.033; Fig. 5A). For the 3’ data set, the increase with time in raw mean occurrence frequencies pooled across integration sites did not reach significance (Spearman correlation test, rho=0.57, pv=0.15) due to the low prevalence of integration sites (7.53 %) observed in the koala sampled in 1960 (Figure 5B). However, similar to the 5’ data set, the prevalence of shared integration sites did increase with time when controlling for the effect of koalas and insertion site (time effect in the GLMM: LRT=5.53, df=1, pv=0.019).

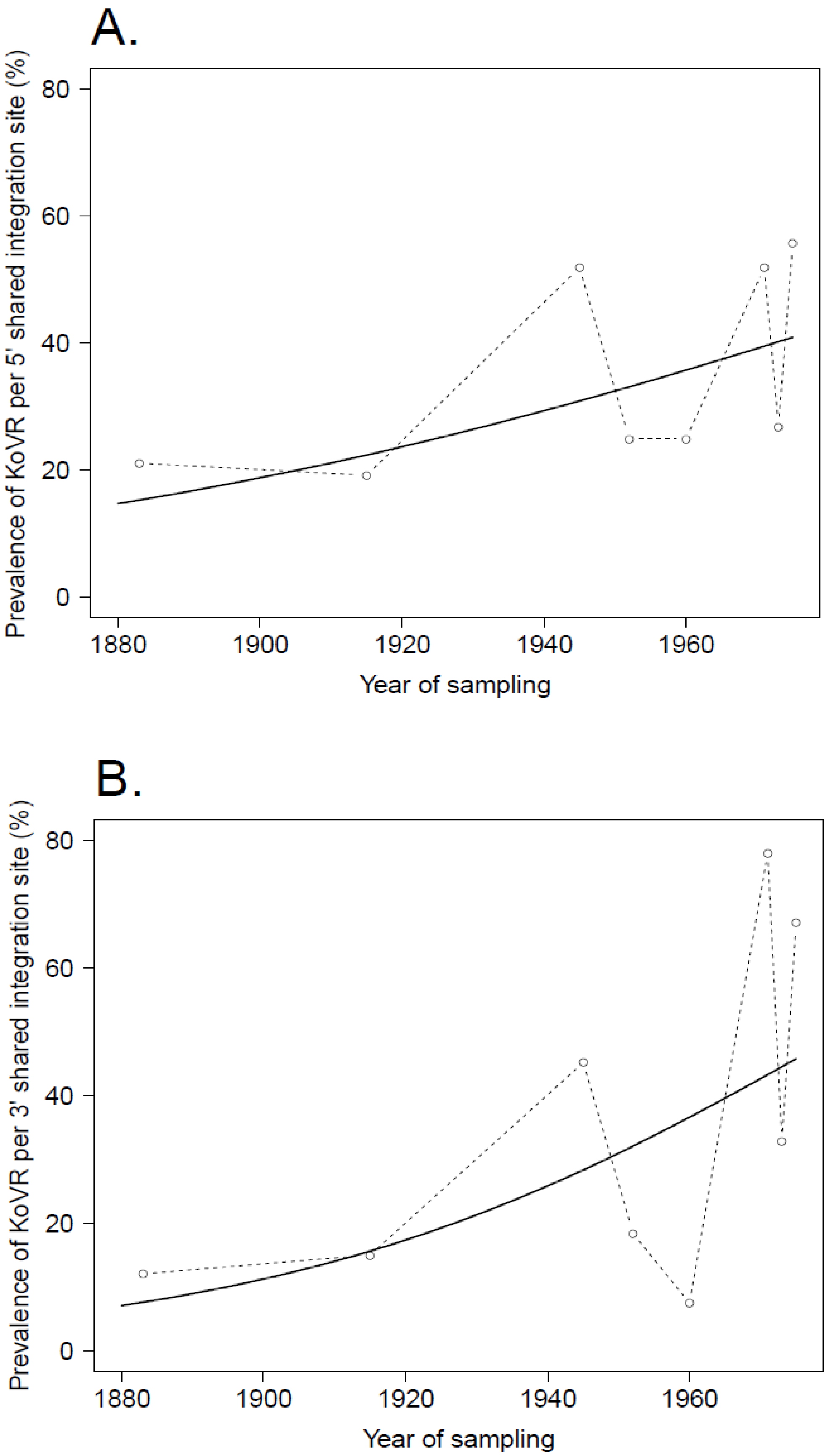


Figure 5. The proportion of KoRV integration sites that are shared among koalas may be increasing over time. The horizontal axis shows the year of collection of museum koala samples

screened for KoRV. The vertical axis shows the proportion of KoRV integration sites within a koala sample that were also detected in other koalas. Dots connected by dashed lines represent the mean prevalence for each year of sampling. The full line represents the prediction from the statistical analysis (Generalised Mixed effect Model): it shows that every 20 years, the odd for shared KoRV integration sites increased by 1.26 times for the 5' data set (LRT=5.06, df=1, pv=0.0024) and by 1.87 times for the 3' data set (LRT=5.53, df=1, pv=0.019), among the ten koala specimens examined.

5 Discussion

The currently available software for identifying viral integration sites using NGS data require an assembled host genome as a reference, e.g., SLOPE (54), VirusFinder (55) and VirusSeq (56). For the koala, however, no assembled genome but only raw sequence reads averaging 98 bp in length are available. We therefore established a customized computational pipeline that was largely reference-independent but made use of the Illumina Hiseq reads of koala and of assembled scaffolds of wallaby (the closest relative to the koala with a genome assembled).

Given the degraded state of DNA in the museum specimens, many of the captured or extended molecules either did not extend beyond the LTR or extended only a few bases into the flank. However, such sequences still represent successful targeted enrichment even if they did not provide extensive integration site information. Primers closer to the ends of the LTRs may have retrieved more and longer integration site data. However, polymorphisms within the ends of the LTRs would likely have caused the loss in the ability of all three methods to identify integration sites, due to the reduced ability of the mismatching oligonucleotides to bind. The distance from the 5' LTR, 37 bp, compared with the 3' LTR, 70 bp, may explain why capture yielded an overabundance of 5' flanking regions as compared to 3' flanking regions. But distance alone is not the explanation as both PEC and SPEX yielded more 3' integration sites overall even though the oligonucleotides were identically positioned. Of note, both techniques that involve extension from a primer (SPEX and PEC) were biased toward the 3' integration sites whereas techniques that did not extend from a primer (hybridization capture or genome-walking) were not. Further analysis will be required to determine the underlying mechanisms generating this bias. Of note, several koala samples in the current study overlap with those examined by PCR (around 100 bp amplifications) in Avila-Arcos et al. 2012 (Table 1). Several samples in that study failed to yield PCR products but were successful here likely because shorter sequences, less than 100 bp, are easily retrieved by the methods applied in the current study.

Hybridization capture found the greatest number of 5' integration sites which included all integration sites identified by SPEX and 87.68% of the integration sites identified by PEC (Figure 4). In contrast, for the 3' LTRs, PEC yielded the most integration sites including 91.28% and 94.12% of the integration sites identified by SPEX and hybridization capture respectively. Considering the output of the methods, the most reliable and comprehensive screening of museum DNA for sequences flanking a target would be achieved by performing PEC and hybridization capture in combination. Both methods covered the full diversity of integration sites identified by SPEX. However, PEC and hybridization capture each retrieved integration sites unique to the method and had reciprocal biases in retrieving 5' and 3' integration sites. It should also be considered that because not all integration sites could be paired for 5' and 3' LTRs, it is clear that not all integration sites present in the samples were retrieved, even when combining all methods. The strong biases towards the 5' or 3' integration sites may prevent such comprehensive analysis from historical samples without very high sequence coverage depth for example, Illumina HiSeq sequencing.

Querying of concatenated 5' and 3' flanks on either side of an integration site yielded 63 matches using the wallaby genome as a reference. The success rate would likely improve upon the availability of an assembled koala reference genome (genome data available to this project was represented by unassembled raw reads of 98 bp average length). Among 63 paired flanking sequences, 40 matches were between 5' and 3' flanks derived from the same individual koala. Twenty three pairs were identified by matching the 5' and 3' flanking sites from different koala individuals. This result demonstrates that although many integration sites were identified per koala, saturation was not achieved and some integration sites were missed. Considering that there are an estimated 165 KoRV copies per haploid genome in Queensland koalas (7), saturation would have required identification of 1,650 5' and 3' integration sites for the 10 koalas for which sequences could be obtained. The average may be an overestimate or underestimate as it was determined by qPCR. However, for aDNA, reaching saturation would be challenging for most samples due to the poor and variable condition of the samples regardless of the actual copy number of KoRV.

KoRV integrations demonstrate significant increased sharing of integration sites among museum koalas in the more recently collected samples. While DNA degradation may alter the detection of both shared and non-shared integrations, modern and historical koalas demonstrated a strong bias against shared integration sites. Moreover, the ancient DNA samples from the current data set did not demonstrate a linear pattern of poorer sample performance based on age. Therefore, data suggests that the proportion of KoRVs shared across koalas has increased in over a period of 110 years. As some of the samples, particularly the oldest were from New South Wales and the younger samples from Queensland, the results could also be explained by geographical differences in specific KoRV integrations. The more commonly shared integrations may represent older KoRV integrations that endogenized earlier and that have had more time for drift to increase their frequency in the population and their geographic extent within the koala population. The methods described here should facilitate the characterization of target flanking sequences of any kind from modern and historical samples.

4.6 REFERENCES

1. Boeke,J.D. & Stoye,J.P. (1997) in *Retroviruses*, eds. Coffin,J.M., Hughes,S.H. & Varmus,H.E. *Cold Spring Harbor Lab. Press*, Plainview, NY, pp. 343–435.

2. Bromham,L. (2002) The human zoo: retroviruses in the human genome. *Trends Ecol. Evol.,* 17, 91-97.

3. Pontius,J.U., Mullikin,J.C., et al. (2007). Initial sequence and comparative analysis of the cat genome. *Genome Res.,* 17(11): 1675-1689.

4. Khodosevich,K., Lebedev,L., Sverdolv,E. (Oct 2002). Endogenous retroviruses and human evolution. *Comp. Funct. Genomics,* (3): 494–98. Doi:10.1002/cfg216.

5. Gifford,R., Tristem,M. (May 2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, 26 (3): 291–315.

6. Tarlinton,R.E., Meers,J., Hanger,J., and Young,P.R. (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(3): 783–787.

7. Tarlinton,R.E., Meers,J., and Young,P.R. (2006) Retroviral invasion of the koala genome. *Nature,* 442(7098): 79–81.

8. Simmons,G.S., Young, P.R., Hanger,J.J., Jones, K., Clarke,D.T.W., McKee,J.J., and Meers,J. (2012) Prevalence of koala retrovirus in geographically diverse populations in Australia. *Aust. Vet. J.,* 90(10): 404–409.

9. Tarlinton,R.E., Meers,J., and Young,P.R. (2008) Biology and evolution of the endogenous koala retrovirus. *Cell. Mol. Life Sci.,* 65: 3413–3421.

10. Ávila-Arcos,M.C., Ho, S.Y.W., et al. (2013) One Hundred Twenty Years of Koala Retrovirus Evolution Determined from Museum Skins. *Mol. Biol. Evol.*, 30(2): 299-304.

11. Tsangaras,K, Siracusa,MC, Nikolaidis,N, Ishida,Y, Cui,P, et al. (2014) Hybridization Capture Reveals Evolution and Conservation across the Entire Koala Retrovirus Genome. *PLoS ONE*, 9(4): e95633.

12. Ishida,Y., Zhao,K., Greenwood,A.D. and Roca,A.L. (2015). Proliferation of Endogenous Retroviruses in the Early Stages of a Host Germ Line Invasion. *Mol. Biol. Evol.,* 32(1): 109-120.

13. Taruscio,D., Manuelidis,L. (1991) Integration site preferences of endogenous retroviruses. *Chromosoma,* 101:141-156.

14. Maxfield,L. F., Fraize, C. D., et al. (2005) Relationship between retroviral DNA-integration-site selection and host cell transcription. *Proc. Natl Acad. Sci. USA,* 102(5): 1436-1441.

15. Lewinski,M.K., Yamashita,M., Emerman,M., Ciuffi,A., Marshall,H., et al. (2006) Retroviral DNA Integration: Viral and Cellular Determinants of Target-Site Selection. *PLoS Pathog,* 2(6): e60. doi:10.1371/journal.ppat.0020060

16. Santoni,F.A., Hartley,O., Luban, J. (2010) Deciphering the Code for Retroviral Integration Target Site Selection. *PLoS Comput Biol,* 6(11): e1001008. doi:10.1371/journal.pcbi.1001008

17. Mitchell,R.S., Beitzel,B.F., Schroder,A.R.W., Shinn,P., Chen,H., et al. (2004) Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site Preferences. *PLoS Biol* 2(8): e234. doi: 10.1371/journal.pbio.0020234

18. Blikstad,V., Benachenhou,F., et al. (2008). Evolution of human endogenous retroviral sequences: a conceptual account. . *Cell. Mol. Life Sci.,* 65(21): 3348-3365.

19. Nowrouzi,A., Dittrich,M., et al. (2006). Genome-wide mapping of foamy virus vector integrations into a human cell line. *J. Gen.Virol.,* 87(5): 1339-1347.

20. Bushman,F., Lewinski,M., et al. (2005). Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Micro.,* 3(11): 848-858.

21. Moalic,Y., Blanchard,Y., et al. (2006). Porcine Endogenous Retrovirus Integration Sites in the Human Genome: Features in Common with Those of Murine Leukemia Virus. *J. Virol.,*80(22): 10980-10988.

22. Schmidt,M., Schwarzwaelder,K.,et al. (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nat. Meth.,* 4(12): 1051-1057.

23. Ciuffi,A. and Barr,S.D. (2011). Identification of HIV integration sites in infected host genomic DNA. *Methods,* 53(1): 39-46.

24. Kustikova,O., Modlich,U., et al. (2009). Retroviral Insertion Site Analysis in Dominant Haematopoietic Clones. *Methods Mol. Biol.,* 506: 373-390.

25. Hüser,D., Gogol-Döring,A., et al. (2010). Integration Preferences of Wildtype AAV-2 for Consensus Rep-Binding Sites at Numerous Loci in the Human Genome. *PLoS Pathog,* 6(7): e1000985.

26. Willerslev, E., & Cooper, A. (2005). Ancient DNA. *Pro. R. Soc. Lond. [Biol],* 272(1558), 3–16. doi:10.1098/rspb.2004.2813

27. Pääbo,S., Poinar,H., Serre, D., Jaenicke-Després,V., Hebler,J., et al. (Dec 2004) Genetic analyses from ancient DNA. *Annu. Rev. Genet.,* 38,645 -679

28. Allentoft,M.E., Collins,M., et al. (Dec 2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci***.,** 7;279(1748):4724-33.

29. Briggs,A.W., Good,J.M., Green,R.E., Krause,J., Maricic,T., Stenzel,U., *et al.* (2009). Primer Extension Capture: Targeted Sequence Retrieval from Heavily Degraded DNA Sources. *J. Vis. Exp.* (31), e1573.

30. Brotherton,P., Endicott,P. et al. (2007). Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res.,* 35(17): 5717-5728.

31. Maricic,T., Whitten,M., et al. (2010). Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE,* 5(11): e14004.

32. Roca,A.L., Ishida,Y., Nikolaidis,N., Kolokotronis,S.O., Fratpietro,S., et al. (Sep 2009) Genetic variation at hair length candidate genes in elephants and the extinct woolly mammoth. *BMC Evol Biol.,* 11;9:232.

33. Wyatt,K.B., Campos,P.F., et al. (2008) Historical Mammal Extinction on Christmas Island (Indian Ocean) Correlates with Introduced Infectious Disease. *PLoS ONE*, 3(11): e3602. doi:10.1371/journal.pone.0003602.

34. Gilbert,M.T.P., Tomsho,L.P., et al. (2007). Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts. *Science*, 317(5846): 1927-1930.

35. Meyer,M. and Kircher,M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc.,* 2010(6): pdb.prot5448.

36. Kircher,M., Sawyer,S., et al. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.,* 40(1): e3.

37. Meyer, M., A. W. Briggs, et al. (2008). From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Res.,* 36(1): e5.

38. Der Sarkissian,C., et al. (2015) Ancient genomics. *Phil. Trans. R. Soc. B.,* 370: 20130387.

39. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, [S.l.], v. 17, n. 1, p. pp. 10-12, may. 2011. ISSN 2226-6089.

40. Bolger,A.M., Lohse,M., and Usadel,B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

41. Magoc,T., and Salzberg,S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics,* 27: 2957-2963.

42. Li,W., Jaroszewski,L., and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics*, 17:282-283.

43. Stajich,J.E., et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12, 1611-8.

44. Fields, B.N., David M. Knipe, D.M., and Howley,P.M. (1996) Fields Virology, Volume 1(3rd Edition) *Lippincott-Raven*, Philedelphia, PA. ASIN: B000TGXAAM.

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

46. Renfree,M.B., Papenfuss,A.T., Deakin,J.E., Lindsay,J., Heider,T., et al. (2011) Genome sequence of an Australian kangaroo,*Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. *Genome Biol.*, *12*(8), R81. doi:10.1186/gb-2011-12-8-r81

47. Meredith,R.W., Westerman,M., et al. (2009). A phylogeny of Diprotodontia (Marsupialia) based on sequences for five nuclear genes. *Mol. Phylogenet. Evol.* 51(3): 554-571.

48. Enright,A.J., Van Dongen,S., and Ouzounis,A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* 30 (7): 1575-1584.

49 Katoh,K., Misawa,K., Kuma,K., and Miyata,T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

50. Waterhouse,A.M., Procter, J.B., Martin,D.M.A, Clamp, M. and Barton, G. J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25 (9) 1189-1191.

51. Kearse,M., Moir,R., Wilson,A., Stones-Havas,S., Cheung,M., et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* 28(12), 1647-1649.

52. Rousset,F., and Ferdy,J.-B. (2014). Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography,* 37(8): 781-790.

53. Tsangaras,K., Wales,N., Sicheritz-Pontén,T., Rasmussen,S., Michaux,J., et al. (2014) Hybridization Capture Using Short PCR Products Enriches Small Genomes by *Cap*turing *Flank*ing Sequences (CapFlank). *PLoS ONE,* 9(10): e109101.

54. Duncavage,E.J., Magrini,V., Becker,N., Armstrong,J.R., Demeter,R.T., et al. (2011) Hybrid Capture and Next-Generation Sequencing Identify Viral Integration Sites from Formalin-Fixed, Paraffin-Embedded Tissue. *J. Mol. Diagn., s : JMD*, *13*(3), 325–333. doi:10.1016/j.jmoldx.2011.01.006

55. Wang,Q., Jia,P., Zhao,Z. (2013) VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. *PLoS ONE,* 8(5): e64465.

56. Chen,Y., Yao,H., et al. (2013). VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics,* 29(2): 266-267.

Figure legends

Figure 1. Experiment design for the identification of KoRV integration sites.

Panel A illustrates that the genome of the koala retrovirus (KoRV) has two identical long terminal repeats (LTRs) on both ends. The primers or baits can bind to both LTRs, so there should be two categories of products: A) products extending into the flanks from primer extension a; B) products extending into the middle of KoRV genome from primer extension b. In principle, there should be equal number of sequences for the two categories. Panel B indicates that the KoRV LTRs contain three components, U3, R and U5. For SPEX, primers were partially nested. All primers are 20 bp long and there is a 8 bp-overlap between the inner primers (3.1 and 5.1) and outer primers (3.2 and 5.2) respectively. To avoid known polymorphisms in the LTR, the 3' end of outer primers are 17 bp from the 5' end of LTR and 50 bp from the 3' end of LTR. Since the 5' LTR and 3' LTR of the same KoRV are identical products can also extend into the KoRV genome. The 5' and 3' flanks can be distinguished by their linked LTR end, with the 5' flank linked to 5' LTR and 3' flank linked to 3' LTR. Considering the longest deletion found at the end of LTR is 19 bp, the LTR end was divided into two segments for subsequent computational identification: the B region representing the last 19 bp of the LTR, and the A region representing the rest of LTR end.

Figure 2. Bioinformatic pipeline for identification of KoRV integration sites. The pipeline was run separately for each data set obtained by three different techniques. For the key steps, the number of sequences retained is indicated in parentheses for each technique in this order from left to right: PEC, SPEX and hybridization capture. After processing NGS reads, KoRV integration sites were identified in a two-step analysis of KoRV LTR ends, next to the host DNA flanking KoRV. The first round selection targeted the A region of the LTR end and its output was used for subsequent identification of the B region. The LTR ends of all sequences were trimmed off and only sequences longer than 4 bp were considered. Using a sequence clustering approach, unique vs shared integration sites were sorted into clusters. Sorting also included singleton clusters and non-singleton clusters. The consensus of each non-singleton cluster was computed using multiple sequence alignment. These consensus sequences and singleton sequences were queried against wallaby genomic scaffolds and koala Illumina Hiseq reads to determine whether they represented KoRV flanking sequences. At the same time extension products into the KoRV genome were identified.

Figure 3. Pairing of 5' and 3' integration sites. The first 4 bp beyond the KoRV 5' LTR is the target site duplication (eg. ACAT in this figure), and the same 4 bp is found at the beginning of a 3' flank (Ishida et al. 2015). One copy of the target site duplication was trimmed off and the 2 flanks were concatenated. The paired 5'-3' integration sites were then screened against the wallaby draft genome and koala Hiseq genomic sequences.

Figure 4. Venn diagrams of KoRV integration sites found by different methods. (A) For 5’ integration sites, HC (hybridization capture) yielded the highest total number of integration sites (887), and covered 91.3% of the integration sites found by SPEX and 86.7% of the integration sites found by PEC. (B) For 3’ integration sites , PEC yielded the highest total number of integration sites (672), and covered 81.4% of the integration sites found by SPEX and 91.7% of the integration sites found by HC (capture hybridization).

Figure 5. The proportion of KoRV integration sites that are shared among koalas may be increasing over time. The horizontal axis shows the year of collection of museum koala samples screened for KoRV. The vertical axis shows the proportion of KoRV integration sites within a koala sample that were also detected in other koalas. Dots connected by dashed lines represent the mean prevalence for each year of sampling. The full line represents the prediction from the statistical analysis (Generalised Mixed effect Model): it shows that every 20 years, the odd for shared KoRV integration sites increased by 1.26 times for the 5’ data set (LRT=5.06, df=1, pv=0.0024) and by 1.87 times for the 3’ data set (LRT=5.53, df=1, pv=0.019), among the ten koala specimens examined.

3.7 Supplementary material

3.7.1 Paired 5' and 3' integration sites from the same koala

```
>P_3rU_6s8_cluster_446_cluster_937__C_5fU_S8.120509_S8.334494
gaattagaaatcgaggagatgcccatcaattggggaatggctgaacaagtcatggtatAAGGAATTATGATGTGATG
TTGTCTTCATGATCCCATTTAGAGTTTTCTTGGCAAAGA
>P_3rU_cluster_564_S_3fS_cluster_29_cluster_977_cluster_4_cluster_n1_cluster_28__C_5f
U_2S3_S_5rS_cluster_7_C_5rS_S3.746665_S4.29352_S_S9.36057_S1.30664_S5.12084_S1
0.22389
gaacccacaaaaagacagaatgaaacaaatatccagcccaatacagcctggatgATTGATCAGAAGGGTGTATC
GCGCCAGGCTGGGAGCACAG
>S_3rU_S17.112853_P_3rU_8S2__P_5fS_cluster_35
GTAAATTGGTCTgaggctggatttgaactcagatcctcctgact
>P_3rU_S8.18869_S8.19758_S8.24709__P_5fU_cluster_54_C_5fU_2S8
AGTGGCTTGCCCCGGGGCACACAGCTAGTATGTATCGGAGGCTGGATTTGAACTC
AGGT
>P_3rS_cluster_175.maffT1s4-6s8__P_5fU_cluster_54_C_5fU_2S8
aggaaactgaggcaaagttaagtgatttgccgggtcacacagctagTATGTATCGGAGGCTGGATTTGAACT
CAGGT
>P_3rU_S2.133819_S2.154135__P_5fU_S2.2639_C_5fU_S2.23072_S2.547057
GACAGCATTTTCCATCCTGCGGCCTCTGGAGATGTCTTAGATCCTTGTATTGCTGA
>P_3rS_S10.119113_7S2_P_3fU_S2.18246_S_3fS_cluster_68_S_3rU_S15.6385__P_5fU_S
2.2639_C_5fU_S2.23072_S2.547057
tatgcttcaatctacattcaaactccgtagttctttctttggatgtggatagcatttttcatcatgaggcctttggagatgtcttagatccttgT
ATTGCTGA
>P_3rS_cluster_31_cluster_252_P_3fS_S2.18623_S2.14939_S_3rU_2S17__P_5fU_S2.2639
```

```
_C_5fU_S2.23072_S2.547057
cttttccgatttgacagcattttccatcctgaggcctctggagatgtcttagatccttgTATTGCTGA
>P_3rU_2S8_P_3fU_7S8__C_5fU_3S8
ccgatctttgcctgaggttgcacacttggtcatatttgagctcaggcaactggggttaagtgacttgcccagagtcacaagtctgaggtcg
gatttgaa
>P_3rU_S8.12674_S8.39898__C_5fU_S8.950860
gacacctcggtgtgtctcagtttcctcatctataaTATGAGCTGGAAAAGGAAATGGCACACTACTCTA
GTATCTTTGCCAAGAAAACCC
>P_3rU_7S3__C_5fU_S3.480683
TATTGTGTCATGTAACCATATAGTATCTCATTGTAAACAAATCATTACATGCACAC
ATTCCCACCAATGCATGCTGGACTTCCTGACAAAGTACAACATGCTCACCTGCCA
ACACTTGCTTGTTAAGACCTGTCAGTGGACTTGACCACTGATGGGTTATAGCTTG
CATTT
>P_3rU_cluster_231__C_5fU_2S8
attcagcctcagacactttccagctgtgtgaccctgggcaagtcagttAACCCCGTCTGCCTCAGTTTCTCCATC
CATAAAATGAGCTGGAGAAAGAAATGGCAAACCACTCCAGGATCTTTGCCAAGA
AATCCCCAAATGGGG
>P_3rU_S8.36760__C_5fU_S8.958388
CATATCAAATGACTTGTCCACAGTCACACAGC
>P_3rU_S8.36760__C_5rU_4S8
CATATCAAATGACTTGTCCACAGTCACACAGCTAGTTTCAGAGGTGAGATTTGAA
>P_3rU_cluster_446_cluster_937__P_5rU_3S8_C_5rU_S8.15867
gaattagaaatcgaggagatgcccatcaattggggaatggctgaacaagtcatggtatAAGGAATTATGATGTGA
>P_3rU_cluster_446_cluster_937__C_5rU_S8.170262
gaattagaaatcgaggagatgcccatcaattggggaatggctgaacaagtcatggtatATGAATGTAATGGAATACT
ATCGTGCTATAAGAAAGA
>P_3rS_8S3_S8.47689__S_5rU_cluster_541_C_5rU_S8.50529
ATACACATAaattagagataagaggcagagttgcacagtcatcagcctcactttctc
>P_3rU_4S8_P_3fU_S8.70215__C_5rU_S8.1022487
ttaaagacagcaatcttctgtctattcttcaagattggtttCAGGAAACTTTAACTGGGGGCTGGAAAAGCA
TCCCATCATTTCTGACTTCCATCCTCCTTCTACTG
>P_3rU_6s3_P_3fS_3S2_2S9_S8.5461_S5.1115_2S4_3S3_S_3fS_cluster_29_cluster_977_c
luster_4_cluster_n1_cluster_28__S_5rS_clsuter_cp1_C_5fU_2S7_P_5fU_cluster_38_C_5rS_
S7.10526_2S10_2S3_S4.17612_S5.4886_S9.1880_S1.11056_2S8_S2.123759
gaacccacaaaaagacagaatgaaacaaatatccagcccaatacagcctggatgATTGATCAGAAGGGTGTATC
GCGCCTTGCTAGGAGCGCAGTGCAGCGCGGTGTGGGCGCACAGGCTGCAGCAAA
CCTGGAGCAGGCCTCAGACTGAATCATGGGCAGCTG
>P_3rU_6s3_P_3fS_3S2_2S9_S8.5461_S5.1115_2S4_3S3_S_3fS_cluster_29_cluster_977_c
luster_4_cluster_n1_cluster_28__S_5rU_S20.31638_C_5rU_S3.50689
gaacccacaaaaagacagaatgaaacaaatatccagcccaatacagcctggatgCTTGATCGGA
>P_3rS_cluster_31_cluster_252_P_3fU_3S2_S_3rS_2S17_S15.6385_S_3fS_cluster_68__S_
5rU_S18.27629_C_5fU_S2.1235481_C_5rU_S2.466916
GACAGCATTTTCCATCCTGAGGCCTCTGGAGATGTCTTAGATCCTTGTATTGCTGA
GAAGGGTTAAGTCTATTAATATTAGTACCTAACTGATTATGTTATTCTCTTCTTGA
GCCAAATCTGATGAGAGTAAGGTTCAAACAATGCTAATATCCGTC
>P_3rS_S5.47458_6S8__S_5rS_cluster_43_C_5fU_S8.854973
ATAAATGAGGAACCTGATATCCaaagaactgaaatgacttaccaaggtcacacagctgatgagtagcagaagca
agaagagaaacaaaatcttctgattcccaggttcctgccac
```

```
>P_3rU_cluster_231__C_5rU_S8.503819
attcagcctcagacactttccagctgtgtgaccctgggcaagtcagttAACCCCGTCTGCCTCAGTTTCC
>P_3rU_S8.12674_S8.39898__C_5rU_S8.911698
gacacctcggtgtgtctcagtttcctcatctataaAATGAGCTGGAGAAGGAAATGACAAACCACTCTA
GTATCTTTGCCAAGAAAACCCCAAATGAGATCA
>P_5fU_cluster_54.maffT3s8_C_5fU_S8.30495_S8.561570__P_3fU_S8.88040_S8.3563
ACCTGAGTTCAAATCCAGCCTCCGATACATACTAGCTGTGCGACCCGGGGCAAGC
CACT
>P_5fU_cluster_54.maffT3s8_C_5fU_S8.30495_S8.561570__P_3fU_6S8_3fU_S8.105580
ACCTGAGTTCAAATCCAGCCTCCGATACATACTAGctgtgtgacccggggcaagccact
>P_5fU_S2.26104_C_5fU_S2.82531__P_3fU_S2.2953
CTGGGAGTTAGGGAGGACCTGAGTTCAAATCCAGCCTCAGACACATAACACTTA
GCATATGTGATG
>C_5fU_S2.410158__S_3fU_S18.5221_P_3fU_S2.310
CATTTTTATTTATTCATACATACTTCCAATCATCAATATGAGAACCATTTTATGTG
CAATACATTGTGCTTTCCAGAACAGTGGAGCCAATTCCCAGCTCCACCAACAATG
CATCAGTG
>C_5fU_S8.589094__C_3fU_S8.917292_P_3fU_S8.49463
CCCAAGAAGTGGTATTGCTGGATCAAAGGGTATGCAGTTTTATAGCCCTTTGGGC
ATAGTTCCAAAT
>C_5fU_S3.480683__P_3fU_S3.28583_S8.82168_S3.90397_S3.64168_S_3fU_S20.10443
AAATGCAAGCTATAACCCATCAGTGGTCAAGTCCACTGACAGGTCTTAACAAGCA
AGTGTTGGCAGGTGAGCATGTTGTACTTTGTCAGGAAGTCCAGCATGCATTGGTG
GGAATGTGTGCATGTAATGATTTGTTTACAATGAGATACTATATGGTTAcatgacacaat
attgtgtcatcaaaattttgatgcaggggaaaaactcacaaatttacaataaatt
>C_5fU_S8.92543_S8.302794__P_3fU_S8.42372
CCCCATTTGGGGATTTCTTGGCAAAGATCCTGGAGTGGTTTGCCATTTCTTTCTCC
AGCTCATTTTATGGATGGAGAAACTGAGGCAGACGGGGTTAACTGACTTGCCCAG
GGTCATACAACTAGGAAGTGTCTGAGGCCAGATTTGAATCCAAGAAGATAAGTC
CTCCTGACTCCGGGTTTGGCAGTCTGTCCACTATGAC
>C_5fU_S3.102282_S3.954929__S_3fU_cluster_195_P_3fU_S3.39389
ATATTATATTCCATGCCCAGCGGTCCTTTAATGTAGaagctgctaaaacttgtgttatcctgattgtgttt
ccactatacttgaattgtttctttcttgcagcttgtaatatat
>C_5fU_S8.242952_S8.249062__S_3fU_S17.14523_C_3fU_S8.70423_C_3rU_cluster_339
TGCTTGGAACCATCGGTTATAGcaaatggagaagttttgaactctgtggataagttcacttacctcggtagtgtacta
>P_5rU_cluster_19_C_5rU_S8.10722__S_3fU_S17.14523_C_3fU_S8.70423_C_3rU_cluster
_339
TTTCTCCACCAGCTGGCACCACATCATCTATGCTTGGAACCATCGGTTatagcaaatgga
gaagttttgaactctgtggataagttcacttacctcggtagtgtacta
>C_5rU_2S3__S_3fU_4S20_P_3fU_S3.74019
GGCCGGTGCTCTATTCACTGTGCCACCTAGATGCCCCTGAAGAATATATTTTAGG
CATATAAATGTGTAT
>C_5rU_S2.906334__P_3fS_5S3_S2.110884_S8.68922
GGTCACCCAGCTAGTAAATATCTGAGGCCAGATTTGAGTCTTTCTGACTTCAGGC
CCTGCACTTTATTCACTGTGCCACCTAGATGAGATCG
>S_5rU_cluster_645.maffT2s20_C_5rU_S3.16848__S_3fU_S20.29912
gaccagactgattagaagcataactacaaaattctgattattgCAATAACCCTTAAGTATAATATTCCAATTA
AGACATCCAAGAGTCACATTTAAATATTGCACTCTTC
```

```
>C_5rU_S8.108028__P_3fU_S8.380
GAATGGGGCACAGTAGTAATAACATTAAAAAGACACACAACTTTGAGAGAATTA
AGGACTTTGATCAACCTAATGACTAACCACAGTTCCAG
>C_5rU_S8.490159__P_3fU_S8.380
GAAGGAGCAGAAATAACATTAAAAAGACACACAACTTTGAGAGAATTAAGGACT
TTGATCAACCTAATGACTAACCACAGTTCCAG
>C_5rU_S8.503819__P_3fU_S8.42372
GGAAACTGAGGCAGACGGGGTTAACTGACTTGCCCAGGGTCATACAACTAGGAA
GTGTCTGAGGCCAGATTTGAATCCAAGAAGATAAGTCCTCCTGACTCCGGGTTTG
GCAGTCTGTCCACTATGACA
```

3.7.2 Paired 5' and 3' integration sites from different koalas

```
>P_3rS_S2.74718_S3.57590_11S8__C_5fU_3S8
TATTTGAGCTCAGGCaactggggttaagtgacttgcccagagtcacaagtctgaggtcggatttgaa
>P_3rU_4S8_P_3fU_S8.70215__C_5fU_S4.684182_5fU_S4.1661030
ttaaagacagcaatcttctgtctattcttcaagattggtttGCCATTTCCTTCTCCAGCTCATTTTATGGAT
>P_3rU_4S8_P_3fU_S8.70215__C_5fU_S4.1683056
ttaaagacagcaatcttctgtctattcttcaagattggtttCCTCATCTGTAAAATGGGGATAATAACA
>P_3rU_S3.26225__C_5fU_S8.589094
AATTGAGGAACTATGCCCAAAGGGCTATAAAACTGCATACCCTTTGATCCAGCAA
TACCACTTCTTGGG
>P_3rU_S8.97760__C_5fU_S3.480683
ACTGAGCCATATAACCATATAGTATCTCATTGTAAACAAATCATTACATGCACAC
ATTCCCACCAATGCATGCTGGACTTCCTGACAAAGTACAACATGCTCACCTGCCA
ACACTTGCTTGTTAAGACCTGTCAGTGGACTTGACCACTGATGGGTTATAGCTTG
CATTT
>P_3rS_S5.105847_5S8__C_5fU_S4.1481035
TGAATGAACATTTTCTCTACTCCGCCATCTTGGCTCCACCCCCC
>C_3rU_S8.1018030_P_3rS_S4.143028_6S3_S2.99154_17S8_P_3fU_2S8_S_3fU_S20.580
49__P_5rU_S9.30145
ggctgattcggactcaggtgagtcttccaactccagggctggcactctatccattcccccacctacctgccctccccacATTCCTT
CAAACCCTCTGTC
>P_3rS_S3.41285_S4.156475_S5.66477_S7.113364_S8.73375_S9.5274_36S2__S_5rU_S18
.20949_C_5rU_S2.111871
AGGAGGTTGAACCAGATGACCTCTGGGGTCTCTTTTAGCC
>S_3rU_S17.112853_P_3rU_8S2__S_5rU_S23.5584_2S20_C_5rS_3S3_S7.4344
GTAAATTGGTCTGAGGCTGGATTTGAACTCAGGTC
>P_3rU_4S8_P_3fU_S8.70215__C_5rU_S4.1173169
ttaaagacagcaatcttctgtctattcttcaagattggtttGTCATTTCCTTCTCCAGCTCATG
>P_3rS_S2.74718_S3.57590_11S8__C_5rU_S9.964451
TATTTGAGCTCAGGCAACTGGGGTTAAGTGACTTGCCAGATCGGAAGAGCGT
>C_5fU_S2.547057__P_3fU_S4.4833
TCAGCAATACAAGGATCTAAGACATCTCCAA
>C_5fU_S8.1066112_S8.1022169_S8.1055816__S_3fU_S20.2502
ttcaaatccgacctcagacttgtgactctgggcaagtcacttaaccccagttgcctCAGATCCAATTCACAT
>C_5fU_S4.684182_5fU_S4.1661030__S_3fU_S17.64559_P_3fU_S8.56518
```

```
ATCCATAAAATGAGCTGGAGAAGGAAATGGCAAACTAGTTTCCCAGTCTATTTCA
TCCGGAGTTGT
>C_5fU_S4.1683056__S_3fU_S17.64559_P_3fU_S8.56518
TGTTATTATCCCCATTTTACAGATGAGGAAACTAGTTTCCCAGTCTATTTCATCCG
GAGTTGT
>C_5fU_S2.729164__S_3fU_4S20_P_3fU_S3.74019
CTTTATTCACTGTGCCACCTAGATGCCCCTGAAGAATATATTTTAGGCATATAAAT
GTGTAT
>C_5fU_S4.1481035__P_3fU_S8.33956
GGGGGGTGGAGCCAAGATGGCGGAGTagaagatgaggaaaaggctcacagaagg
>S_5rU_cluster_620_C_5rU_2S8_P_5rU_S8.78052__P_3fU_S3.22152
agaggtccaggcaaagagCTATGATCCCCGGTTTCTGCTTTCCTTCTAGTTAAATCGGA
>C_5rU_S2.906334__S_3fU_4S20_P_3fU_S3.74019
GGTCACCCAGCTAGTAAATATCTGAGGCCAGATTTGAGTCTTTCTGACTTCAGGC
CCTGCACTTTATTCACTGTGCCACCTAGATGCCCCTGAAGAATATATTTTAGGCAT
ATAAATGTGTAT
>C_5rU_2S3__P_3fS_5S3_S2.110884_S8.68922
GGCCGGTGCTCTATTCACTGTGCCACCTAGATGAGATCG
>C_5rU_S4.1173169__S_3fU_S17.64559_P_3fU_S8.56518
CATGAGCTGGAGAAGGAAATGACAAACTAGTTTCCCAGTCTATTTCATCCGGAGT
TGT
>S_5rU_S23.5584_2S20_C_5rS_3S3_S7.4344__P_3fS_8s2_1s5
GACCTGAGTTCAAATCCAGCCTCAGACAAATTCCAAGAAAAAGTTAGCTCTTTCC
CCTTCCTCCCCCTCCTGTGCCAT
>P_5rU_S4.11115__S_3fU_S20.10939_P_3fU_S3.74062
TTCAAGCAGAAGACGGCATACGAGATAGAGGCGGTGACTGGAGTTCAGACGTGT
GCTTTGCCGATCTGAAGACACAGTAGTCAATGACTATAGTAGTCTTC
```

Table S1. Indexing primers for indexed Illumina library construction

index primer sequence*	**Experiment Method**
CAAGCAGAAGACGGCATACGAGATgccatctGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATaacctggGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATctaacggGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATagaggcgGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATccgcaagGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATctccgccGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATacgtccaGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATcatggttGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATcttcctgGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATaggtatgGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATggattggGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATacgccggGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATgcggcaaGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATtgatagGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATtatacgGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATcgatgaGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATatacacGTGACTGGAGTTCAGACGTGT	SPEX

CAAGCAGAAGACGGCATACGAGATatagcgGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATtgttcaGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATagatacGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATtagctgGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATgtatgtGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATggctcaGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATcatgctGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATtcatcgGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATcatctaGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATgtcacaGTGACTGGAGTTCAGACGTGT	SPEX

*** The Illumina indice (6-7bp long) are embeded in the primers in lower letters**

Table S2. Primers and baits used in the experiments

Primer name	**Sequence in 5' to 3'direction**
SPEX_Primer_3.1	Biotin- ATTTGCATCCGGAGTTGTGT
SPEX_Primer_5.1	Biotin- CGGAATGATTTCTGCCTCAT
SPEX_Primer_3.2	Biotin- AGTTGTGTTCGCGTTGATCC
SPEX_Primer_5.2	Biotin- TTCCATACTCCACGGAATGA
SPEX-2R_illumina P7	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATATATGGGIIGGGIIGGGIIGGG
SPEX-2F_5_illumina P5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGGAATGATTTCTGCCTCAT
SPEX-2F_3_illumina P5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTGTGTTCGCGTTGATCC
HybCap_KoRV3LTR_F	Biotin- TCAAGGACATCC*GATTTGCATCCGGAGTTGTGTTCGCGTTGATCC
HybCap_KoRV5LTR_R	Biotin- TCAAGGACATCC*GTTCCATACTCCACGGAATGATTTCTGCCTCAT
PEC_Primer_3.1	Biotin- CAAGGACATCC*GATTTGCATCCGGAGTTGTGT
PEC_Primer_3.2	Biotin- CAAGGACATCC*GAGTTGTGTTCGCGTTGATCC
PEC_Primer_5.1	Biotin- CAAGGACATCC*GCGGAATGATTTCTGCCTCAT
PEC_Primer_5.2	Biotin- CAAGGACATCC*GTTCCATACTCCACGGAATGA

I = deoxyinosine.

*** phosphorothioate bond to render the oligonucelotides resistant to nuclease degradation**

Chapter IV

Evolutionary relationships among extinct and extant sloths: the evidence of mitogenomes and retroviruses

Chapter III

Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without a reference genome

3.1 Summary

Retroviral integration into the host germline results in permanent viral colonization of vertebrate genomes. The koala retrovirus (KoRV) is currently invading the germline of the koala (*Phascolarctos cinereus*) and provides a unique opportunity for studying retroviral endogenization. Previous analysis of KoRV integration patterns in modern koalas demonstrate that they share integration sites primarily if they are related, indicating that the process is currently driven by vertical transmission rather than infection. However, due to methodological challenges, KoRV integrations have not been comprehensively characterized, nor have historical trends been examined. To overcome these challenges, we applied and compared three target enrichment techniques coupled with next generation sequencing (NGS) and a sequence-clustering based computational pipeline to determine the integration sites for 10 museum Northern Australian koala samples collected between the 1870s and late 1980s. Although three enrichment methods each exhibited bias in integration site retrieval, hybridization capture based methods performed best. The results suggest that the proportion of KoRVs shared among unrelated koalas has increased over the last 140 years.

3.2 Introduction

Vertebrate endogenous retroviruses (ERVs) descend from exogenous retroviruses that infected the ancestral germ line and have been subsequently transmitted vertically from parent to offspring through Mendelian inheritance (1). ERVs comprise up to 8-11% of vertebrate genomes (2, 3). Most ERVs colonized their host genomes millions of years ago (4, 5) making it difficult to study the process of retroviral invasion. The koala retrovirus (KoRV) spreads both horizontally and vertically among koalas (*Phascolarctos cinereus*) (6, 7, 8), and it is still in the process of endogenizing, unlike most other described ERVs (9). Therefore, KoRV provides a unique opportunity to study the processes underlying retroviral endogenization in real time. Historical DNA analysis from museum koala samples collected during the 19th and 20th centuries demonstrated that KoRV was already ubiquitous in northern Australia by the 19th century (10), and that its genome has remained strongly conserved (11). In contrast, KoRV integration sites among individuals are highly variable (11, 12).

The integrated provirus has identical sequences at the 5' and 3' ends of the proviral genome, which are termed the long terminal repeats (LTR). Distribution of retroviral integration sites in the host genome is generally regarded as non-random (13), for example, fewer integration events are observed near transcriptional hotspots (14). Integration site preference is associated with the viral integrase (15) and host chromosomal features (16). Retroviruses belonging to the same group tend to exhibit similar integration site preference (17). Despite these tendencies, integration of a specific retrovirus at a specific site is still a random event. Most individuals in a host population will share older ERV integration sites as they become fixed in the population over time through drift, as is now true for most human endogenous retroviruses (18). In contrast, if a retrovirus endogenized very recently, the integration site will be rare among all but the most closely related individuals such as offspring. This is the case for KoRV integrations, which appear to be largely unique to related koalas (11, 12). However, previous studies have not attempted a comprehensive survey of integration sites. The focus of the current study was to evaluate methods that may comprehensively characterize retroviral integrations and which could be applied to museum samples to examine historical trends in the frequency of shared or unique KoRV integration sites.

Inverse PCR has conventionally been used for retrieving retroviral integration sites (19). Methods such as rapid amplification of cDNA ends (RACE), ligation-mediated PCR, Linker-selection-mediated PCR, linear amplification–mediated PCR and genome walking (20; 21; 22; 23; 24; 25) have also been used. However, it is unclear if they can comprehensively detect integration sites particularly due to potential primer-target mismatch, and they have never been applied to ancient DNA (aDNA). DNA extracted from museum samples has the characteristics of aDNA, e.g., it is heavily fragmented and damaged, and in low concentration (26). The DNA degradation, fragmentation and contamination that occurs post mortem makes aDNA research difficult (27; 28), often preventing the use of conventional molecular biological methods such as PCR.

To overcome the limitations of working with historical DNA, we applied three target enrichment techniques followed by high-throughput Illumina sequencing. The three techniques, Primer Extension Capture (PEC) (29), Single Primer Extension (SPEX) (30), and hybridization capture (31) have been applied successfully to aDNA and could potentially be employed to determine sequences flanking targeted ERVs. Ten koala museum samples collected between the 1870s and the 1980s were successfully examined. Because no assembled koala genome is currently available, a reference-independent computational pipeline was established. The results are discussed in terms of performance of the three methods in retrieving KoRV integration sites per koala.

3.3 Materials and methods

3.3.1 Samples and ancient DNA extraction

A total of thirteen museum samples were examined as described in Table 1. DNA extractions were performed in the aDNA laboratory of the Department of Wildlife Diseases of the Leibniz Institute for Zoo and Wildlife Research in Berlin, Germany. The laboratory is dedicated to aDNA work and has never been used for molecular work on modern samples. The room is UV irradiated 4 hours every night by ceiling-mounted UV lights. All work performed in the facility follows procedures designed to minimize the possibility of contamination, such as use of laminar flow hoods and use of protective clothing to avoid sample contamination.

Approximately 250 mg of skin tissue per specimen were extracted using a silica-based extraction kit for aDNA (GENECLEAN Ancient DNA Extraction Kit, MP Biomedicals, USA). The protocol followed the manufacturer's instructions and has been successfully applied to a variety of ancient sample types (32; 33). Mock extractions were performed with each set of koala museum specimens as negative controls during extraction. Subsequent to each extraction, the isolated DNA was further purified using a MinElute spin column (Qiagen, Hilden, Germany) as described in (34) to remove potential inhibitors for the subsequent enzymatic reactions.

3.3.2 NGS Library preparation

Illumina sequencing libraries were prepared from the extracts using a previously described protocol (35) with the following modifications: (A) All SPRI purification steps were substituted with spin column purification (MinElute PCR purification kit, Qiagen). (B) Adapter concentration in the ligation reaction was reduced to 0.2 mM per adapter. (C) The purification after adapter fill-in was substituted by heat inactivation at 80°C for 20 min. The libraries were then used directly as template for subsequent amplification following a two-step strategy (36). A quality control strategy (37) was also applied, which consisted of a qPCR to quantify the product after each step of library amplification. The qPCR results excluded three samples from further processing for which DNA quality was too poor for analysis.

In the first round of amplification, AmpliTaq Gold, a non-proof reading enzyme, and indexing primers (Table S1) were applied, adding a distinct P7 index to each library as described in (35), 10 indices for the 10 working samples and 3 and 4 negative control indices for PEC and SPEX respectively. Adding distinct indices to each library allows for multiple samples to be sequenced in a single sequencing run. Non-proof reading enzyme allows for amplification to be performed on templates containing deoxyuracils, which are common with aDNA (38). Except for removal of 1 µL for qPCR as a library quality control, the remaining libraries were used as template in 100 µL PCR reactions containing 1x Taq buffer II (Applied Biosystems), 5U AmpliTaq Gold (Applied Biosystems), 250 mM each dNTP and 100 nM each indexing primer. Cycling conditions followed manufacturer's instructions: The pre-denaturation step lasted 12 min at 95°C, followed by 12 cycles of denaturation at 95°C for 20 s, annealing at 60°C for 30 s and elongation at 72°C for 40 s, with a final extension step of 72°C for 5 min. PCR products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany).

In the second round of amplification, 5 µl of the purified PCR product from the first round PCR was used as template for a second PCR. This involved 50 µL reactions containing Herculase II Fusion DNA Polymerase (Agilent Technologies Catalog 600677), which has proof reading activity, and primers IS5 and IS6 (35) at a final concentration of 400 nM each. Cycling conditions included an activation step of 3 min at 95°C, followed by 15-20 cycles of denaturation at 95°C for 20 s, annealing at 60°C for 25 s and elongation at 72°C for 30 s, with a final extension step at 72°C for 3 min. The number of cycles used in the PCR for every sample was dependent on the concentration of each of the libraries as determined by the qPCR assay. The PCR amplified libraries were then purified using the QIAquick PCR purification kit. Each library was separately used in subsequent PEC and hybridization capture experiments.

3.3.3 Bait preparation and integration site enrichment

Three methods were compared for retrieving integration sites: primer extension capture (PEC), single primer extension (SPEX) and hybridization capture. All three have been successfully applied to ancient and historical DNA samples and all are applicable to samples that would not be expected to yield results with conventional methods for integration site analysis. The same set of primers was used in PEC and SPEX experiments (Figure 1, Table S2). Because the two LTRs of a provirus are identical, the primers designed for targeting the 5' integrations will also extend targeting the retroviral *env* gene and the primers designed for targeting the 3' integrations will also extend targeting the retroviral gag leader sequence (Figure 1A). For both the 5' and 3' KoRV LTR, two 20 bp primers were developed which overlap such that the 3' end of the first primer overlapped 8 bp with the 5' end of the second primer (Figure 1 B primers 5.1 and 5.2 and 3.1 and 3.2 respectively). To avoid known LTR polymorphisms among KoRV proviruses, the two primers on each side of the LTR were located 17 bp from the 5' end and 50 bp from the 3' end of the LTRs respectively in conserved regions (Figure 1B). The baits used for hybridization capture were synthesized to generate 32 bp oligonucleotides that spanned the full length of sequence covered by primers 5.1 and 5.2 (32 bp) on the 5' LTR and primers 3.1 and 3.2 (32 bp) on the 3' end.

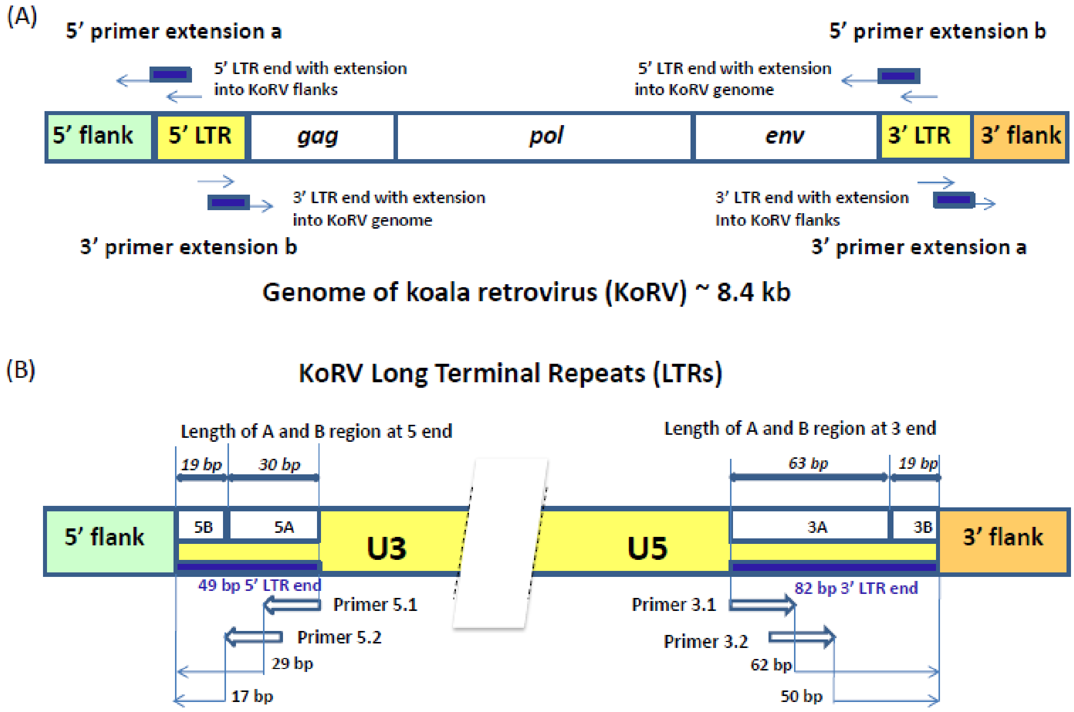

(A)
5' primer extension a
5' primer extension b
5' LTR end with extension into KoRV flanks
5' LTR end with extension into KoRV genome
5' flank
5' LTR
gag
pol
env
3' LTR
3' flank
3' LTR end with extension into KoRV genome
3' LTR end with extension Into KoRV flanks
3' primer extension b
3' primer extension a
Genome of koala retrovirus (KoRV) ~ 8.4 kb
(B)
KoRV Long Terminal Repeats (LTRs)
Length of A and B region at 5 end
Length of A and B region at 3 end
19 bp
30 bp
63 bp
19 bp
5' flank
5B
5A
U3
U5
3A
3B
3' flank
49 bp 5' LTR end
82 bp 3' LTR end
Primer 5.1
Primer 5.2
Primer 3.1
Primer 3.2
29 bp
17 bp
62 bp
50 bp

Figure 1. Experiment design for the identification of KoRV integration sites.

Panel A illustrates that the genome of the koala retrovirus (KoRV) has two identical long terminal repeats (LTRs) on both ends. The primers or baits can bind to both LTRs, so there should be two categories of products: A) products extending into the flanks from primer extension a; B) products extending into the middle of KoRV genome from primer extension b. In principle, there should be equal number of sequences for the two categories. Panel B indicates that the KoRV LTRs contain three components, U3, R and U5. For SPEX, primers were partially nested. All primers are 20 bp long and there is a 8 bp-overlap between the inner primers (3.1 and 5.1) and outer primers (3.2 and 5.2) respectively. To avoid known polymorphisms in the LTR, the 3' end of outer primers are 17 bp from the 5' end of LTR and 50 bp from the 3' end of LTR. Since the 5' LTR and 3' LTR of the same KoRV are identical products can also extend into the KoRV genome. The 5' and 3' flanks can be distinguished by their linked LTR end, with the 5' flank linked to 5' LTR and 3' flank linked to 3' LTR. Considering the longest deletion found at the end of LTR is 19 bp, the LTR end was divided into two segments for subsequent computational identification: the B region representing the last 19 bp of the LTR, and the A region representing the rest of LTR end.

3.3.4 Primer Extension Capture (PEC)

Indexed libraries were pooled in equi-molar ratios for primer extension following a published protocol (29). After each step, 1 µL of the product was quantified by qPCR. To minimize the amplification bias, each of the captured products was amplified in triplicate, using 5 µL of the captured product as template for each reaction, using the same kit and cycling conditions as described previously under NGS library preparation for second round amplification of Illumina indexed libraries, except that we ran 20 cycles of amplification for all samples. Amplified captured libraries were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany) and eluted in 50 µl of elution buffer (EB) and used as template for a second round of PEC.

3.3.5 Single Primer Extension (SPEX)

The SPEX experiments generally followed a published protocol (30) using DNA extracts prior to Illumina library construction with three modifications: (1) Illumina sequencing adaptors were attached to the 5’ end of the primers used in in the first round of partially nested PCR; (2) MyTaq HS Mix (Bioline, BIO-25045) was used instead of Platinum Taq DNA Polymerase High Fidelity in the first round of a partially nested PCR; (3) only one round of a partially nested PCR amplification was performed. The nested PCR products were then quantified by qPCR and indexed using Illumina indexing primers (Table S1). The indexed PCR products were purified using a QIAquick PCR Purification Kit (Qiagen). The amplicons were quantified by qPCR and subjected to a second round of amplification using the same conditions as the first round. The products were purified again using the QIAquick PCR Purification Kit (Qiagen), quantified by qPCR and pooled at equi-molar ratios. All PEC and SPEX products were pooled and measured using High Sensitivity DNA chips on an Agilent 2100 Bioanalyzer, then sequenced at the National High-throughput DNA Sequencing Centre, Copenhagen, Denmark using Illumina MiSeq Reagent Kit v2 (300 cycle).

3.3.6 Hybridization capture

The amplified libraries were pooled in equi-molar ratios at final concentration of 2 µg. An established protocol was followed (31) except that synthesized oligonuceotide baits were used instead of PCR products and the EB volume for final elution using Qiagen MiniElute column was 20 µL instead of 15 µL. After 2 days of hybridization and subsequent elution steps, 1 µL of the final eluate was quantified by qPCR and 5 µL (in total 15 µL) was amplified in triplicate using the same kit and cycling conditions as described in the NGS library preparation for second round amplification of Illumina indexed libraries. The pooled PCR products were purified using the QIAquick PCR Purification Kit and was measured using the Tapestation 2200 (Agilent Technologies Catalog G2964AA). Hybridization capture libraries were sequenced at the National High-throughput DNA Sequencing Centre, Copenhagen, Denmark using Illumina MiSeq Reagent Kit v2 (300 cycle).

3.3.7 Preprocessing of sequence data

Adaptor sequences occur at the ends of each sequence read. Adaptor sequences were removed from sequence reads using cutadapt-1.2.1 (39), and quality trimming was performed using Trimmomatic-0.22 using default settings (40). The paired forward and reverse sequence reads were merged using Flash-1.2.5 where possible (41), and both the merged and unmerged reads were used for further analyses. PCR duplicates (clonality in the sequencing data) with 100% sequence identity were removed using cd-hit-v4.6.1 (42).

3.3.8 Identification of KoRV integration sites

Figure 2 and Table S3 summarize the computational pipeline used for the identification of KoRV integration sites. For its implementation, both existing software and customized perl scripts were used that made use of BioPerl (43). Because the nested primers or bait were designed near the ends of LTR, the primer extension products would include either the first 49 bp of the 5' LTR or the last 82 bp of the 3' LTR, which are designated "LTR ends" in Figure 1A. All sequences with a KoRV flank should contain an LTR end, as a result of the primer extension (Figure 1B). Therefore, KoRV integration sites could be identified as the sequence beyond the KoRV LTR end, since all integration sites would have this sequence. However, due to DNA degradation in museum samples, some primer extension product may not have a complete LTR end. Furthermore, minor deletions at the end of the integrated LTRs may be present (44); for example, a 19 bp deletion was found in a KoRV provirus (12). To get around these potential issues, identification of the LTR ends relied on sequentially selecting sample sequences that contain defined LTR segments; this was done in separate steps for the 5' and 3' flank-containing sequences. The LTR end was divided into two segments, designated A and B (Figure 1b): the B segment corresponds to the last 19 bp of the LTR and is referred to as 5B or 3B in the 5' and 3' LTR ends, respectively. The A segment is the remaining section of the LTR end, which has a length of 30 bp in the 5' end (5A) and 63 bp in the 3' end (3A).

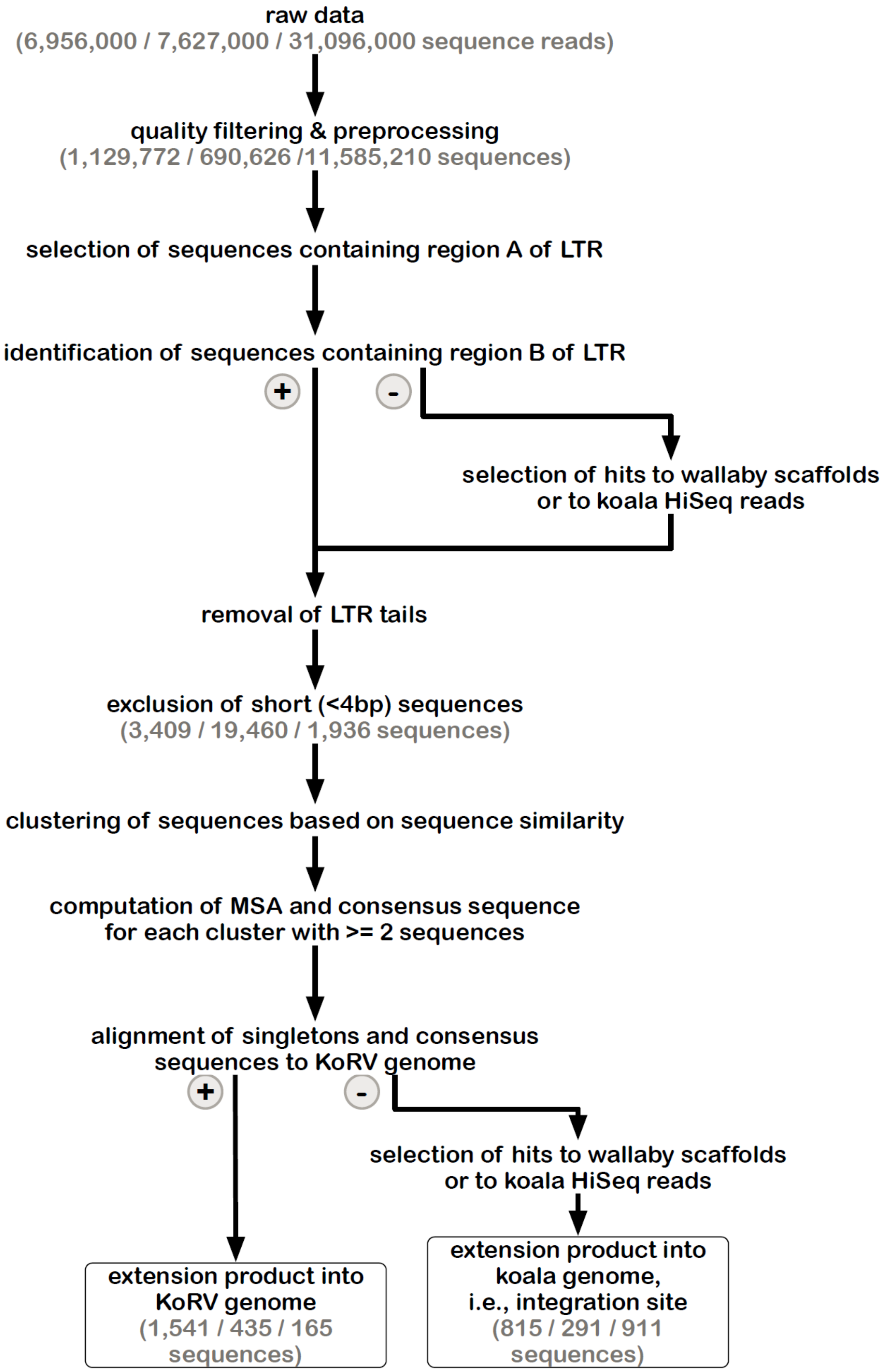

raw data
(6,956,000 / 7,627,000 / 31,096,000 sequence reads)
quality filtering & preprocessing
(1,129,772 / 690,626 /11,585,210 sequences)
selection of sequences containing region A of LTR
identification of sequences containing region B of LTR
+
-
selection of hits to wallaby scaffolds
or to koala HiSeq reads
removal of LTR tails
exclusion of short (<4bp) sequences
(3,409 / 19,460 / 1,936 sequences)
clustering of sequences based on sequence similarity
computation of MSA and consensus sequence
for each cluster with >= 2 sequences
alignment of singletons and consensus
sequences to KoRV genome
+
-
selection of hits to wallaby scaffolds
or to koala HiSeq reads
extension product into
KoRV genome
(1,541 / 435 / 165
sequences)
extension product into
koala genome,
i.e., integration site
(815 / 291 / 911
sequences)

Figure 2. Bioinformatic pipeline for identification of KoRV integration sites. The pipeline was run separately for each data set obtained by three different techniques. For the key steps, the number of

sequences retained is indicated in parentheses for each technique in this order from left to right: PEC, SPEX and hybridization capture. After processing NGS reads, KoRV integration sites were identified in a two-step analysis of KoRV LTR ends, next to the host DNA flanking KoRV. The first round selection targeted the A region of the LTR end and its output was used for subsequent identification of the B region. The LTR ends of all sequences were trimmed off and only sequences longer than 4 bp were considered. Using a sequence clustering approach, unique vs shared integration sites were sorted into clusters. Sorting also included singleton clusters and non-singleton clusters. The consensus of each non-singleton cluster was computed using multiple sequence alignment. These consensus sequences and singleton sequences were queried against wallaby genomic scaffolds and koala Illumina Hiseq reads to determine whether they represented KoRV flanking sequences. At the same time extension products into the KoRV genome were identified.

Initially, sequences containing either of the two A regions in the KoRV LTR end (5A or 3A in Figure 1B) were identified. For this step, optimal local pairwise sequence alignments (Smith-Waterman, EMBOSS) were computed between each sample sequence and the A region in either 5' or 3' LTR end. Sequences were kept for further analysis if they could be aligned to at least 20 bp of the 30 bp 5A segment with at least 90% identity, or if they could be aligned to at least 43 of the 63 bp 3A segment with at least 90% identity (Table 2). Sequences not passing these criteria were discarded as artifacts. The LTR ends of all sequences meeting these criteria were trimmed to the last 19 bp and then used for further analyses.

From these sequences, B segments of either 3' or 5' LTR ends were identified (3B or 5B in Figure. 1B). For this step, optimal local sequence alignments were computed between each of the trimmed sequence and the B segment in either the 3' or the 5' LTR end. Only sequences that could be aligned to at least 12 bp of the 19 bp long B segment (3B or 5B) with at least 80% (Table 2) identity were selected. The last 19 bp of LTR ends were trimmed from all sequences meeting the selection criteria, leaving LTR free KoRV flanks or KoRV genomic DNA adjacent to the LTR.

All sequences that contained the A region, but for which the B region was not detected using the pairwise alignment strategy, were then subjected to another test. Specifically, these sequences were used as queries for two separate local database searches using BLAST (48). Such sequences represent LTRs that have suffered deletions at the end, a common occurrence in proviruses. One search was against HiSeq sequencing data of a koala from Queensland, Australia with 100X coverage. The data represent raw Illumina sequences and are not annotated or assembled. After adaptor and quality trimming, 6.469 billion reads from this koala, with a mean length of 78 bp, were used for this step. Sequences were considered KoRV integration sites when their non-LTR portion could be aligned with greater than 90% identity to the koala reads over 60% length of the sample sequence. A second search was against the Tammar wallaby (*Macropus eugenii*) genome (GenBank: ABQO000000000.2), which represents the closest related species to koala for which a genome has been assembled (46). Although the wallaby and koala lineages diverged more than 50 Mya (47), we expected that some of the koala genomic DNA (flanking KoRV) could be aligned to the homologous wallaby regions. Sequences with at least 70% identity over 50% length of the sample sequence to the wallaby genome were therefore considered to be KoRV integration sites. For the sequences with a match to the wallaby scaffolds or the koala data, the LTR sequences were trimmed and were then concatenated with the KoRV flanks (obtained in previous steps) for further analysis.

3.3.9 Sorting of sequences representing different integration sites

All sequences with matches to the different segments of the 3’ and 5’LTR ends and/or to wallaby scaffolds or koala HiSeq data from each of the enrichment techniques were collected. The sequences matching 3’ and 5’ LTR ends were kept separate, resulting in a total of six different data sets for further analysis (two data sets each for the PEC, SPEX and capture). LTR ends had been removed from all sequences in these data sets. Before using these sequences to identify shared and unique integration sites, KoRV flank sequences shorter than 4 bp (the typical length of a KoRV target site duplication) were defined as “short insertion sites” (Table 3) and were excluded from further analysis; only KoRV flanks of 4 bp or longer (representing the length of target site duplication as identified in Ishida, Y., et al 2015) were used for further analysis. At this stage, the PEC data had 392 5’ flank sequences and 2,347 3’ flank sequences; the SPEX data 6521 5’ flank sequences and 9,200 3’ flank sequences; and hybridization capture 1,158 5’ flank sequences and 28 3’ flank sequences.

A clustering approach was used to sort all sequences in each of the six data sets into groups of similar sequences; each cluster representing a unique integration site. Sequences that did not share significant similarity with any other sequences in the input file were called singletons. For each of the six data sets, all-against-all BLAST comparisons were run, and the BLAST output was used as input for clustering using TRIBE-MCL (48), separately for each data set. Different combinations of E-values (all against all BLAST) and inflation values (TRIBE-MCL) were used for this step and the optimal parameter combination for each data set was evaluated. For all combinations of E-values and inflation values, multiple sequence alignments were computed for all clusters using MAFFT v7.127b (49). To assess the quality of the clustering, alignments of the 30 largest clusters of each clustering result were visualized in jalview (50) and were checked by eye. An alignment was considered high quality if the total number of mismatches and gaps in every sequence of the alignment was no more than

10% of the sequence length. If all 30 clusters were evaluated to be of high quality, the sequence was further analyzed. The parameter combinations for optimal clustering and related all against all BLAST are listed in Table 4.

Singletons and non-singleton clusters containing sequences derived from a single individual koala were considered to represent unique integration sites. Clusters containing sequences shared by more than one koala were considered to represent shared integration sites. A consensus sequence was computed from the alignment of each non-singleton cluster. Singletons and consensus sequences were then further evaluated first by computing pairwise alignments between these sequences and the *gag* or *env* part of KoRV genome (Figure 1A) (GenBank: AF151794.2). The sequences that could be aligned to the KoRV genes with at least 90% identity and of any length were categorized as primer extension or flank capture within the KoRV genome. The LTR sequences at the 5' and 3' ends of the KoRV genome are identical or nearly so and therefore 50% of the PCR products should extend into the KoRV genome (Figure 1A). Sequences that could not be mapped to KoRV genome were potential KoRV integration sites and were evaluated further. For such sequences, a length filtering was performed with threshold of 15 bp, since this is the minimum length that can be effectively identified in BLAST. The sequences longer than 15 bp were first used as query in BLAST to search against the koala shotgun Hiseq data; they were also mapped to wallaby genome (GenBank: ABQO000000000.2) in Geneious version 6.18 (http://www.geneious.com, 51). Identified sequences for either one of the two computations were considered to be KoRV integration sites. Sequences shorter than 15 bp are too short for efficient mapping or BLAST; however, because they contained an LTR end, were included in the KoRV specific enrichment statistics (Table 3), although they were not further analyzed.

3.3.10 Pairing of 5' and 3' integration site to one KoRV provirus

Ishida, et al 2015 identified the length of the retroviral target site duplication (a stretch of host DNA directly adjacent to retrovirus which is duplicated during retroviral integration) for KoRV to be 4 bp. Based on this target site duplication length (Figure 3), all 5' and 3' integration sites were examined for shared 4 bp target site identity. Only flanks longer than 16 bp were used for matching 5' and 3' integration sites. The minimum 28 bp (32 bp minus the 4 bp target site duplication) combined length discriminated true wallaby matches from non-significant blastn results.

The paired 5'-3' integration sites were 1) mapped against the wallaby genome using the mapping tool in Geneious using default settings, where only the paired 5'-3' integration sites that could be mapped to the wallaby genome with over 70% of their total length were scored as positively identified; 2) used as query to search in the Hiseq data (a Queensland wild koala) using BLAST. Here, only the paired 5’-3’ integration sites that could be aligned with over 90% identity with the koala Hiseq reads were considered positive.

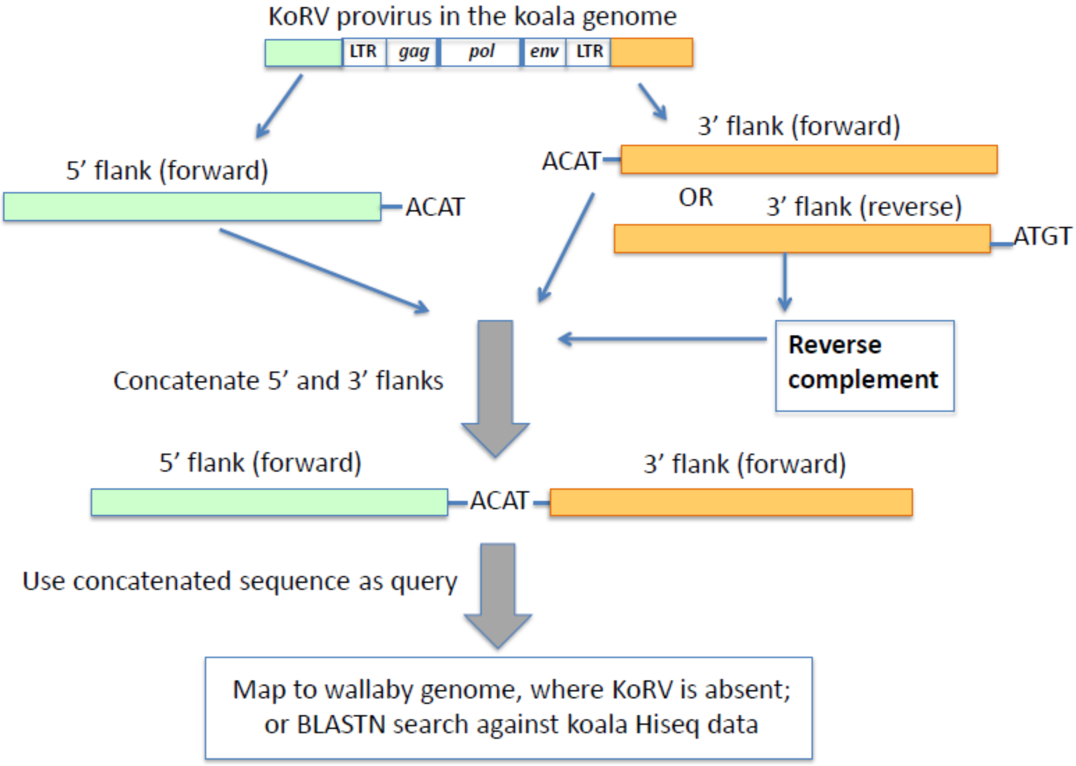

KoRV provirus in the koala genome
LTR
gag
pol
env
LTR
5’ flank (forward)
ACAT
ACAT
3’ flank (forward)
OR
3’ flank (reverse)
ATGT
Reverse complement
Concatenate 5’ and 3’ flanks
5’ flank (forward)
ACAT
3’ flank (forward)
Use concatenated sequence as query
Map to wallaby genome, where KoRV is absent; or BLASTN search against koala Hiseq data

Figure 3. Pairing of 5’ and 3’ integration sites. The first 4 bp beyond the KoRV 5’ LTR is the target site duplication (eg. ACAT in this figure), and the same 4 bp is found at the beginning of a 3’ flank (Ishida et al. 2015). One copy of the target site duplication was trimmed off and the 2 flanks were concatenated. The paired 5’-3’ integration sites were then screened against the wallaby draft genome and koala Hiseq genomic sequences.

3.3.11 Statistical analysis of shared integration sites

Statistical tests were performed to check if the occurrences of KoRV at sampled integration sites increased as the samples became younger among the 10 museum koala samples. Two logistic regression models were employed: one for 5’ integration sites and one for 3’ integration sites. Both models had the same structure. The occurrence was considered (binary: 1=presence, 0=absence) as the response variable and time as a continuous fixed effect. Because results were qualitatively similar irrespective of expressing "time" as rank or directly as years, for the sake of simplicity, only the latter was reported. The identity of koalas and of insertion sites were considered as two Gaussian random effects, making this logistic regression a Generalised Mixed effect Model (GLMM). The GLMM was fitted using the function HLfit from the R package spaMM 1.4.1 (52), considering a Binomial error structure.

The effect of time was tested by performing an asymptotic Likelihood Ratio Test (LRT) using the function anova.HLfit from the same package.

4 Results

NGS sequencing post enrichment by all three tested methods generated hundreds of thousands to millions of reads. After the pre-processing steps, 1,129,772 sequences from the PEC approach were available for further analysis, 690,626 from SPEX, and 11,585,210 from hybridization capture.

4.1 Single primer extension

Using SPEX to target the 5’ LTR flanks, 66 integration sites unique to a single koala, and 15 integration sites shared by more than one koala were identified across the 10 koala samples. Integration sites derived from consensus sequences generated from sequence clusters with at least 4 bp of sequence flanking the KoRV LTR. An additional 15,822 sequences were too short (less than 4 bp) for further biological interpretation. A total of 212 sequences contained only the KoRV genome, *env* to 5’ LTR. This is a consequence of the identical primer binding sites in the 5’ and 3’ LTRs (Figure 1A), since KoRV 5’ and 3’ LTRs are identical or nearly so (12). Thus, approximately 50% of the sequences are expected to extend from the LTR into the virus rather into the host flanking region. Sequences that extended into KoRV were categorized separately but included in the total enrichment efficiency evaluation. SPEX also identified 182 unique and 28 shared 3’ LTR flanks; an additional 1,527 sequences were too short to further analyze and 223 were found extending into the KoRV genome (Table 3).

4.2 Primer extension capture

PEC designed to identify flanking regions 5’ of integration sites detected 126 unique and 17 shared integration sites; an additional 496 sequences were too short to further characterize and 135 sequences extended into the KoRV genome. PEC targeting regions downstream of 3’ LTR integration sites identified 538 unique and 134 shared integration sites; an additional 1,806 sequences were too short to characterize further and 1,406 sequences extended into the KoRV genome (Table 3).

4.3 Hybridization capture

Using the 5’ LTR region as bait, 862 unique and 25 shared 5’ flanking regions were identified. An additional 191 sequences were too short to further characterize, while 151 sequences extended into the KoRV genome. Additionally, 24 unique and no shared integration sites were identified by hybridization using the 3’ LTR as bait. The strong bias towards the 5’ integration sites has been observed previously (11) although it is unclear why the preferential LTR enrichment occurs. Additionally, 41 sequences were too short to further characterize and 14 sequences extended into the KoRV genome (Table 3).

4.4 Summary of computational data processing

At each step of our bioinformatics pipeline, we recorded for each experiment the number of sequences that met our screening criteria. Additional information like mean length, minimum length and maximum length of sequences was also computed at each step (Table S3). Before any screening criteria were applied, PEC produced 6,956 million reads, SPEX produced 7,627 million, and hybridization capture produced 31,096 million. After pre-processing (including PCR duplicate removal) of this sequencing data, 16.24% of the initial sequencing reads were kept for PEC, 9.05% for SPEX and 37.25% for hybridization capture. Clonality was more prevalent for SPEX than for either PEC or hybridization capture.

After the first round of LTR end identification, 31,787 (2.67%) LTR positive sequences were identified for PEC, 142577 (19.94%) for SPEX and 5,648 (0.0483%) for hybridization capture. Sequences passing the second round of LTR end selection were 5,692 for PEC, 31,941 for SPEX, and 1,503 for hybridization capture. No KoRV flanks were detected in negative controls, extraction or PCR controls lacking template, for any experiment.

4.5 Cross-technique comparisons

Efficiency of target enrichment for each technique was calculated as the total number of identified integration sites divided by the total number of sequences after removal of clonality. The total number of identified integration sites included KoRV flanking sequences (including sequences shorter or longer than 4 bp) and reads extending into the KoRV genome. Sequences extending into the KoRV genome are not the desired target but because of the identical or nearly identical sequences of the 5’ and 3’ LTRs all such sequences represent correctly targeted enriched sequences.

As shown in Table 2, PEC enriched the highest total number of 3’ integration sites, 531, whereas hybridization capture enriched the most 5’ integration sites, 762. As a percentage of the total sequences retrieved, SPEX achieved the highest target enrichment efficiency (4.684%). Both PEC and hybridization capture exhibited lower enrichment percentages (0.554% and 0.0135% respectively).

Due to a phenomenon known as CapFlank (53), koala genome sequences near the integration sites may be enriched together with KoRV flanks by concatenation of library molecules on the baits. To estimate the number of such target flanks, after PCR duplicate removal all sequences were screened against the wallaby genome using BLAST. Hybridization capture exhibited the lowest efficiency of on-target enrichment (0.0135%, Table 3) and highest ratio of CapFlank enrichment (16.409%), while SPEX achieved the highest efficiency of on-target enrichment (4.684%) and lowest ratio of CapFlank enrichment (0.226%).

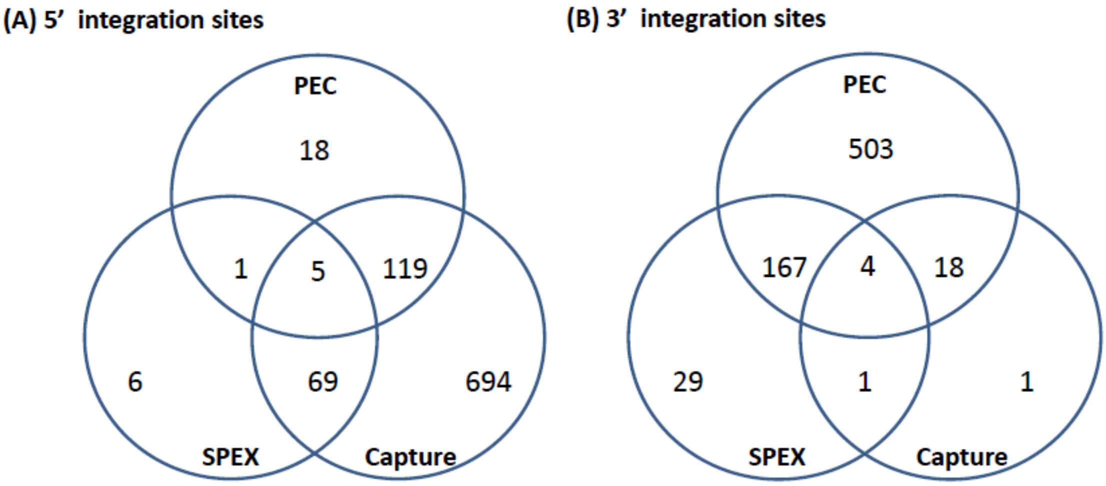


Figure 4. Venn diagrams of KoRV integration sites found by different methods. (A) For 5' integration sites, HC (hybridization capture) yielded the highest total number of integration sites (887), and covered 91.3% of the integration sites found by SPEX and 86.7% of the integration sites found by PEC. (B) For 3' integration sites , PEC yielded the highest total number of integration sites (672), and covered 81.4% of the integration sites found by SPEX and 91.7% of the integration sites found by HC (capture hybridization).

As illustrated in Figure 4, for the 5' LTR integration sites, hybridization capture yielded the highest total number of integration sites, 887, and contained 91.3% of the integration sites identified in the SPEX data set and 86.7% of the integration sites identified in PEC data set. The 3' LTR integration data followed a different profile with PEC generating the highest total number of integration sites, 672, containing 81.4% of the integration sites in the SPEX data set and 91.7% of the integration sites in the hybridization capture data set.

4.6 Shared and unique integration sites

After identical integration sites across the data sets generated by the 3 techniques were combined, 52 shared and 865 unique 5' KoRV host flanks could be identified. Shared integration sites accounted for 5.7% of the total identified using 5’ flanking host sequences, a similar percentage as estimated in previous studies (11). Among the 3’ flanking regions, 146 shared and 570 unique integration sites were identified, with shared sites accounting for 20.4% of total integration sites identified using 3’ host genomic sequences.

4.7 Pairing of 5’ and 3’ flanking regions to identify individual proviral integration sites

KoRV typically produces a 4 bp target site duplication upstream and downstream of its integration site (12). By comparing the 4 bp target site duplication, 1,690 5’ and 3’ host flanking regions were screened in the koala genome to identify potential paired flanking regions. Sixty three pairs of 5’ and 3’ KoRV integration sites were identified as originating from a same proviral loci. Of these 63 pairs, 40 were derived from a single koala (Supplement List 1), whereas 23 matches were identified by pairing 5’ and 3’ flanks identified in different koalas (Supplement list 2).

4.8 Statistical modeling of shared KoRV integration sites among 10 koalas

The proportion of 5’ integration sites that were shared with other koalas was significantly higher in more recently collected koala specimens than in specimens collected further in the past. This was true both when the influence of the identity of the koala and of the insertion site were accounted for in a statistical model (time effect in the GLMM: LRT=5.06, df=1, pv=0.0024), and on raw mean occurrence frequencies pooled across insertion sites (Spearman correlation test, rho=0.75, pv=0.033; Fig. 5A). For the 3’ data set, the increase with time in raw mean occurrence frequencies pooled across integration sites did not reach significance (Spearman correlation test, rho=0.57, pv=0.15) due to the low prevalence of integration sites (7.53 %) observed in the koala sampled in 1960 (Figure 5B). However, similar to the 5’ data set, the prevalence of shared integration sites did increase with time when controlling for the effect of koalas and insertion site (time effect in the GLMM: LRT=5.53, df=1, pv=0.019).

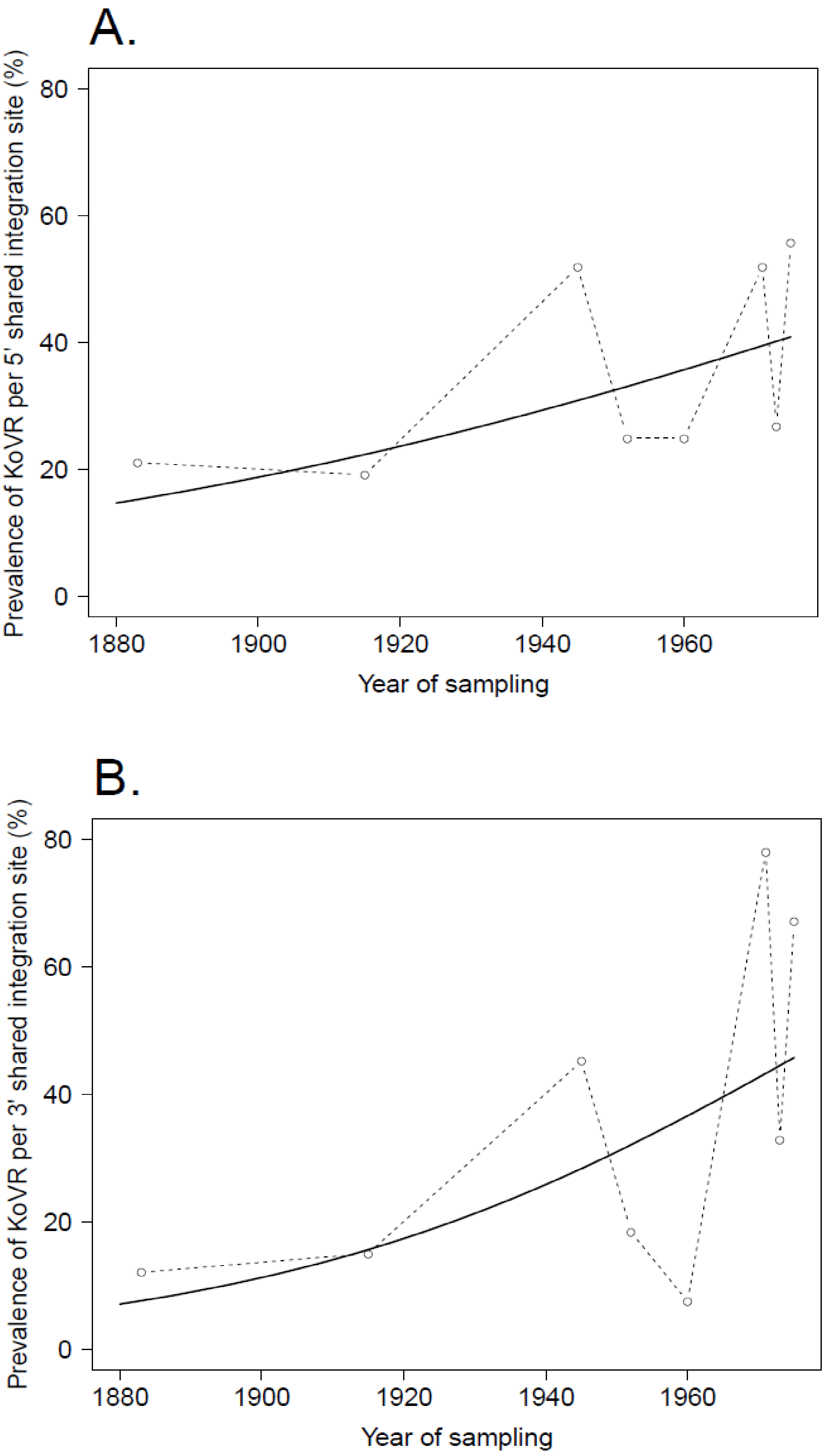


Figure 5. The proportion of KoRV integration sites that are shared among koalas may be increasing over time. The horizontal axis shows the year of collection of museum koala samples

screened for KoRV. The vertical axis shows the proportion of KoRV integration sites within a koala sample that were also detected in other koalas. Dots connected by dashed lines represent the mean prevalence for each year of sampling. The full line represents the prediction from the statistical analysis (Generalised Mixed effect Model): it shows that every 20 years, the odd for shared KoRV integration sites increased by 1.26 times for the 5' data set (LRT=5.06, df=1, pv=0.0024) and by 1.87 times for the 3' data set (LRT=5.53, df=1, pv=0.019), among the ten koala specimens examined.

5 Discussion

The currently available software for identifying viral integration sites using NGS data require an assembled host genome as a reference, e.g., SLOPE (54), VirusFinder (55) and VirusSeq (56). For the koala, however, no assembled genome but only raw sequence reads averaging 98 bp in length are available. We therefore established a customized computational pipeline that was largely reference-independent but made use of the Illumina Hiseq reads of koala and of assembled scaffolds of wallaby (the closest relative to the koala with a genome assembled).

Given the degraded state of DNA in the museum specimens, many of the captured or extended molecules either did not extend beyond the LTR or extended only a few bases into the flank. However, such sequences still represent successful targeted enrichment even if they did not provide extensive integration site information. Primers closer to the ends of the LTRs may have retrieved more and longer integration site data. However, polymorphisms within the ends of the LTRs would likely have caused the loss in the ability of all three methods to identify integration sites, due to the reduced ability of the mismatching oligonucleotides to bind. The distance from the 5' LTR, 37 bp, compared with the 3' LTR, 70 bp, may explain why capture yielded an overabundance of 5' flanking regions as compared to 3' flanking regions. But distance alone is not the explanation as both PEC and SPEX yielded more 3' integration sites overall even though the oligonucleotides were identically positioned. Of note, both techniques that involve extension from a primer (SPEX and PEC) were biased toward the 3' integration sites whereas techniques that did not extend from a primer (hybridization capture or genome-walking) were not. Further analysis will be required to determine the underlying mechanisms generating this bias. Of note, several koala samples in the current study overlap with those examined by PCR (around 100 bp amplifications) in Avila-Arcos et al. 2012 (Table 1). Several samples in that study failed to yield PCR products but were successful here likely because shorter sequences, less than 100 bp, are easily retrieved by the methods applied in the current study.

Hybridization capture found the greatest number of 5' integration sites which included all integration sites identified by SPEX and 87.68% of the integration sites identified by PEC (Figure 4). In contrast, for the 3' LTRs, PEC yielded the most integration sites including 91.28% and 94.12% of the integration sites identified by SPEX and hybridization capture respectively. Considering the output of the methods, the most reliable and comprehensive screening of museum DNA for sequences flanking a target would be achieved by performing PEC and hybridization capture in combination. Both methods covered the full diversity of integration sites identified by SPEX. However, PEC and hybridization capture each retrieved integration sites unique to the method and had reciprocal biases in retrieving 5' and 3' integration sites. It should also be considered that because not all integration sites could be paired for 5' and 3' LTRs, it is clear that not all integration sites present in the samples were retrieved, even when combining all methods. The strong biases towards the 5' or 3' integration sites may prevent such comprehensive analysis from historical samples without very high sequence coverage depth for example, Illumina HiSeq sequencing.

Querying of concatenated 5' and 3' flanks on either side of an integration site yielded 63 matches using the wallaby genome as a reference. The success rate would likely improve upon the availability of an assembled koala reference genome (genome data available to this project was represented by unassembled raw reads of 98 bp average length). Among 63 paired flanking sequences, 40 matches were between 5' and 3' flanks derived from the same individual koala. Twenty three pairs were identified by matching the 5' and 3' flanking sites from different koala individuals. This result demonstrates that although many integration sites were identified per koala, saturation was not achieved and some integration sites were missed. Considering that there are an estimated 165 KoRV copies per haploid genome in Queensland koalas (7), saturation would have required identification of 1,650 5’ and 3’ integration sites for the 10 koalas for which sequences could be obtained. The average may be an overestimate or underestimate as it was determined by qPCR. However, for aDNA, reaching saturation would be challenging for most samples due to the poor and variable condition of the samples regardless of the actual copy number of KoRV.

KoRV integrations demonstrate significant increased sharing of integration sites among museum koalas in the more recently collected samples. While DNA degradation may alter the detection of both shared and non-shared integrations, modern and historical koalas demonstrated a strong bias against shared integration sites. Moreover, the ancient DNA samples from the current data set did not demonstrate a linear pattern of poorer sample performance based on age. Therefore, data suggests that the proportion of KoRVs shared across koalas has increased in over a period of 110 years. As some of the samples, particularly the oldest were from New South Wales and the younger samples from Queensland, the results could also be explained by geographical differences in specific KoRV integrations. The more commonly shared integrations may represent older KoRV integrations that endogenized earlier and that have had more time for drift to increase their frequency in the population and their geographic extent within the koala population. The methods described here should facilitate the characterization of target flanking sequences of any kind from modern and historical samples.

4.6 REFERENCES

1. Boeke,J.D. & Stoye,J.P. (1997) in *Retroviruses*, eds. Coffin,J.M., Hughes,S.H. & Varmus,H.E. *Cold Spring Harbor Lab. Press*, Plainview, NY, pp. 343–435.

2. Bromham,L. (2002) The human zoo: retroviruses in the human genome. *Trends Ecol. Evol.,* 17, 91-97.

3. Pontius,J.U., Mullikin,J.C., et al. (2007). Initial sequence and comparative analysis of the cat genome. *Genome Res.,* 17(11): 1675-1689.

4. Khodosevich,K., Lebedev,L., Sverdolv,E. (Oct 2002). Endogenous retroviruses and human evolution. *Comp. Funct. Genomics,* (3): 494–98. Doi:10.1002/cfg216.

5. Gifford,R., Tristem,M. (May 2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, 26 (3): 291–315.

6. Tarlinton,R.E., Meers,J., Hanger,J., and Young,P.R. (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(3): 783–787.

7. Tarlinton,R.E., Meers,J., and Young,P.R. (2006) Retroviral invasion of the koala genome. *Nature,* 442(7098): 79–81.

8. Simmons,G.S., Young, P.R., Hanger,J.J., Jones, K., Clarke,D.T.W., McKee,J.J., and Meers,J. (2012) Prevalence of koala retrovirus in geographically diverse populations in Australia. *Aust. Vet. J.,* 90(10): 404–409.

9. Tarlinton,R.E., Meers,J., and Young,P.R. (2008) Biology and evolution of the endogenous koala retrovirus. *Cell. Mol. Life Sci.,* 65: 3413–3421.

10. Ávila-Arcos,M.C., Ho, S.Y.W., et al. (2013) One Hundred Twenty Years of Koala Retrovirus Evolution Determined from Museum Skins. *Mol. Biol. Evol.*, 30(2): 299-304.

11. Tsangaras,K, Siracusa,MC, Nikolaidis,N, Ishida,Y, Cui,P, et al. (2014) Hybridization Capture Reveals Evolution and Conservation across the Entire Koala Retrovirus Genome. *PLoS ONE*, 9(4): e95633.

12. Ishida,Y., Zhao,K., Greenwood,A.D. and Roca,A.L. (2015). Proliferation of Endogenous Retroviruses in the Early Stages of a Host Germ Line Invasion. *Mol. Biol. Evol.,* 32(1): 109-120.

13. Taruscio,D., Manuelidis,L. (1991) Integration site preferences of endogenous retroviruses. *Chromosoma,* 101:141-156.

14. Maxfield,L. F., Fraize, C. D., et al. (2005) Relationship between retroviral DNA-integration-site selection and host cell transcription. *Proc. Natl Acad. Sci. USA,* 102(5): 1436-1441.

15. Lewinski,M.K., Yamashita,M., Emerman,M., Ciuffi,A., Marshall,H., et al. (2006) Retroviral DNA Integration: Viral and Cellular Determinants of Target-Site Selection. *PLoS Pathog,* 2(6): e60. doi:10.1371/journal.ppat.0020060

16. Santoni,F.A., Hartley,O., Luban, J. (2010) Deciphering the Code for Retroviral Integration Target Site Selection. *PLoS Comput Biol,* 6(11): e1001008. doi:10.1371/journal.pcbi.1001008

17. Mitchell,R.S., Beitzel,B.F., Schroder,A.R.W., Shinn,P., Chen,H., et al. (2004) Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site Preferences. *PLoS Biol* 2(8): e234. doi: 10.1371/journal.pbio.0020234

18. Blikstad,V., Benachenhou,F., et al. (2008). Evolution of human endogenous retroviral sequences: a conceptual account. . *Cell. Mol. Life Sci.,* 65(21): 3348-3365.

19. Nowrouzi,A., Dittrich,M., et al. (2006). Genome-wide mapping of foamy virus vector integrations into a human cell line. *J. Gen.Virol.,* 87(5): 1339-1347.

20. Bushman,F., Lewinski,M., et al. (2005). Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Micro.,* 3(11): 848-858.

21. Moalic,Y., Blanchard,Y., et al. (2006). Porcine Endogenous Retrovirus Integration Sites in the Human Genome: Features in Common with Those of Murine Leukemia Virus. *J. Virol.,*80(22): 10980-10988.

22. Schmidt,M., Schwarzwaelder,K.,et al. (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nat. Meth.,* 4(12): 1051-1057.

23. Ciuffi,A. and Barr,S.D. (2011). Identification of HIV integration sites in infected host genomic DNA. *Methods,* 53(1): 39-46.

24. Kustikova,O., Modlich,U., et al. (2009). Retroviral Insertion Site Analysis in Dominant Haematopoietic Clones. *Methods Mol. Biol.,* 506: 373-390.

25. Hüser,D., Gogol-Döring,A., et al. (2010). Integration Preferences of Wildtype AAV-2 for Consensus Rep-Binding Sites at Numerous Loci in the Human Genome. *PLoS Pathog,* 6(7): e1000985.

26. Willerslev, E., & Cooper, A. (2005). Ancient DNA. *Pro. R. Soc. Lond. [Biol],* 272(1558), 3–16. doi:10.1098/rspb.2004.2813

27. Pääbo,S., Poinar,H., Serre, D., Jaenicke-Després,V., Hebler,J., et al. (Dec 2004) Genetic analyses from ancient DNA. *Annu. Rev. Genet.,* 38,645 -679

28. Allentoft,M.E., Collins,M., et al. (Dec 2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci***.,** 7;279(1748):4724-33.

29. Briggs,A.W., Good,J.M., Green,R.E., Krause,J., Maricic,T., Stenzel,U., *et al.* (2009). Primer Extension Capture: Targeted Sequence Retrieval from Heavily Degraded DNA Sources. *J. Vis. Exp.* (31), e1573.

30. Brotherton,P., Endicott,P. et al. (2007). Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res.,* 35(17): 5717-5728.

31. Maricic,T., Whitten,M., et al. (2010). Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE,* 5(11): e14004.

32. Roca,A.L., Ishida,Y., Nikolaidis,N., Kolokotronis,S.O., Fratpietro,S., et al. (Sep 2009) Genetic variation at hair length candidate genes in elephants and the extinct woolly mammoth. *BMC Evol Biol.,* 11;9:232.

33. Wyatt,K.B., Campos,P.F., et al. (2008) Historical Mammal Extinction on Christmas Island (Indian Ocean) Correlates with Introduced Infectious Disease. *PLoS ONE*, 3(11): e3602. doi:10.1371/journal.pone.0003602.

34. Gilbert,M.T.P., Tomsho,L.P., et al. (2007). Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts. *Science*, 317(5846): 1927-1930.

35. Meyer,M. and Kircher,M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc.,* 2010(6): pdb.prot5448.

36. Kircher,M., Sawyer,S., et al. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.,* 40(1): e3.

37. Meyer, M., A. W. Briggs, et al. (2008). From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Res.,* 36(1): e5.

38. Der Sarkissian,C., et al. (2015) Ancient genomics. *Phil. Trans. R. Soc. B.,* 370: 20130387.

39. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, [S.l.], v. 17, n. 1, p. pp. 10-12, may. 2011. ISSN 2226-6089.

40. Bolger,A.M., Lohse,M., and Usadel,B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

41. Magoc,T., and Salzberg,S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics,* 27: 2957-2963.

42. Li,W., Jaroszewski,L., and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics,* 17:282-283.

43. Stajich,J.E., et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res,* 12, 1611-8.

44. Fields, B.N., David M. Knipe, D.M., and Howley,P.M. (1996) Fields Virology, Volume 1(3rd Edition) *Lippincott-Raven*, Philedelphia, PA. ASIN: B000TGXAAM.

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

46. Renfree,M.B., Papenfuss,A.T., Deakin,J.E., Lindsay,J., Heider,T., et al. (2011) Genome sequence of an Australian kangaroo,*Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. *Genome Biol.*, *12*(8), R81. doi:10.1186/gb-2011-12-8-r81

47. Meredith,R.W., Westerman,M., et al. (2009). A phylogeny of Diprotodontia (Marsupialia) based on sequences for five nuclear genes. *Mol. Phylogenet. Evol.* 51(3): 554-571.

48. Enright,A.J., Van Dongen,S., and Ouzounis,A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* 30 (7): 1575-1584.

49 Katoh,K., Misawa,K., Kuma,K., and Miyata,T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

50. Waterhouse,A.M., Procter, J.B., Martin,D.M.A, Clamp, M. and Barton, G. J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25 (9) 1189-1191.

51. Kearse,M., Moir,R., Wilson,A., Stones-Havas,S., Cheung,M., et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647-1649.

52. Rousset,F., and Ferdy,J.-B. (2014). Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography*, 37(8): 781-790.

53. Tsangaras,K., Wales,N., Sicheritz-Pontén,T., Rasmussen,S., Michaux,J., et al. (2014) Hybridization Capture Using Short PCR Products Enriches Small Genomes by *Cap*turing *Flank*ing Sequences (CapFlank). *PLoS ONE*, 9(10): e109101.

54. Duncavage,E.J., Magrini,V., Becker,N., Armstrong,J.R., Demeter,R.T., et al. (2011) Hybrid Capture and Next-Generation Sequencing Identify Viral Integration Sites from Formalin-Fixed, Paraffin-Embedded Tissue. *J. Mol. Diagn., s : JMD*, *13*(3), 325–333. doi:10.1016/j.jmoldx.2011.01.006

55. Wang,Q., Jia,P., Zhao,Z. (2013) VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. *PLoS ONE*, 8(5): e64465.

56. Chen,Y., Yao,H., et al. (2013). VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics*, 29(2): 266-267.

Figure legends

Figure 1. Experiment design for the identification of KoRV integration sites.

Panel A illustrates that the genome of the koala retrovirus (KoRV) has two identical long terminal repeats (LTRs) on both ends. The primers or baits can bind to both LTRs, so there should be two categories of products: A) products extending into the flanks from primer extension a; B) products extending into the middle of KoRV genome from primer extension b. In principle, there should be equal number of sequences for the two categories. Panel B indicates that the KoRV LTRs contain three components, U3, R and U5. For SPEX, primers were partially nested. All primers are 20 bp long and there is a 8 bp-overlap between the inner primers (3.1 and 5.1) and outer primers (3.2 and 5.2) respectively. To avoid known polymorphisms in the LTR, the 3' end of outer primers are 17 bp from the 5' end of LTR and 50 bp from the 3' end of LTR. Since the 5' LTR and 3' LTR of the same KoRV are identical products can also extend into the KoRV genome. The 5' and 3' flanks can be distinguished by their linked LTR end, with the 5' flank linked to 5' LTR and 3' flank linked to 3' LTR. Considering the longest deletion found at the end of LTR is 19 bp, the LTR end was divided into two segments for subsequent computational identification: the B region representing the last 19 bp of the LTR, and the A region representing the rest of LTR end.

Figure 2. Bioinformatic pipeline for identification of KoRV integration sites. The pipeline was run separately for each data set obtained by three different techniques. For the key steps, the number of sequences retained is indicated in parentheses for each technique in this order from left to right: PEC, SPEX and hybridization capture. After processing NGS reads, KoRV integration sites were identified in a two-step analysis of KoRV LTR ends, next to the host DNA flanking KoRV. The first round selection targeted the A region of the LTR end and its output was used for subsequent identification of the B region. The LTR ends of all sequences were trimmed off and only sequences longer than 4 bp were considered. Using a sequence clustering approach, unique vs shared integration sites were sorted into clusters. Sorting also included singleton clusters and non-singleton clusters. The consensus of each non-singleton cluster was computed using multiple sequence alignment. These consensus sequences and singleton sequences were queried against wallaby genomic scaffolds and koala Illumina Hiseq reads to determine whether they represented KoRV flanking sequences. At the same time extension products into the KoRV genome were identified.

Figure 3. Pairing of 5' and 3' integration sites. The first 4 bp beyond the KoRV 5' LTR is the target site duplication (eg. ACAT in this figure), and the same 4 bp is found at the beginning of a 3' flank (Ishida et al. 2015). One copy of the target site duplication was trimmed off and the 2 flanks were concatenated. The paired 5'-3' integration sites were then screened against the wallaby draft genome and koala Hiseq genomic sequences.

Figure 4. Venn diagrams of KoRV integration sites found by different methods. (A) For 5' integration sites, HC (hybridization capture) yielded the highest total number of integration sites (887), and covered 91.3% of the integration sites found by SPEX and 86.7% of the integration sites found by PEC. (B) For 3' integration sites , PEC yielded the highest total number of integration sites (672), and covered 81.4% of the integration sites found by SPEX and 91.7% of the integration sites found by HC (capture hybridization).

Figure 5. The proportion of KoRV integration sites that are shared among koalas may be increasing over time. The horizontal axis shows the year of collection of museum koala samples screened for KoRV. The vertical axis shows the proportion of KoRV integration sites within a koala sample that were also detected in other koalas. Dots connected by dashed lines represent the mean prevalence for each year of sampling. The full line represents the prediction from the statistical analysis (Generalised Mixed effect Model): it shows that every 20 years, the odd for shared KoRV integration sites increased by 1.26 times for the 5' data set (LRT=5.06, df=1, pv=0.0024) and by 1.87 times for the 3' data set (LRT=5.53, df=1, pv=0.019), among the ten koala specimens examined.

3.7 Supplementary material

3.7.1 Paired 5' and 3' integration sites from the same koala

```
>P_3rU_6s8_cluster_446_cluster_937__C_5fU_S8.120509_S8.334494
gaattagaaatcgaggagatgcccatcaattggggaatggctgaacaagtcatggtatAAGGAATTATGATGTGATG
TTGTCTTCATGATCCCATTTAGAGTTTTCTTGGCAAAGA
>P_3rU_cluster_564_S_3fS_cluster_29_cluster_977_cluster_4_cluster_n1_cluster_28__C_5f
U_2S3_S_5rS_cluster_7_C_5rS_S3.746665_S4.29352_S_S9.36057_S1.30664_S5.12084_S1
0.22389
gaacccacaaaaagacagaatgaaacaaatatccagcccaatacagcctggatgATTGATCAGAAGGGTGTATC
GCGCCAGGCTGGGAGCACAG
>S_3rU_S17.112853_P_3rU_8S2__P_5fS_cluster_35
GTAAATTGGTCTgaggctggatttgaactcagatcctcctgact
>P_3rU_S8.18869_S8.19758_S8.24709__P_5fU_cluster_54_C_5fU_2S8
AGTGGCTTGCCCCGGGGCACACAGCTAGTATGTATCGGAGGCTGGATTTGAACTC
AGGT
>P_3rS_cluster_175.maffT1s4-6s8__P_5fU_cluster_54_C_5fU_2S8
aggaaactgaggcaaagttaagtgatttgccgggtcacacagctagTATGTATCGGAGGCTGGATTTGAACT
CAGGT
>P_3rU_S2.133819_S2.154135__P_5fU_S2.2639_C_5fU_S2.23072_S2.547057
GACAGCATTTTCCATCCTGCGGCCTCTGGAGATGTCTTAGATCCTTGTATTGCTGA
>P_3rS_S10.119113_7S2_P_3fU_S2.18246_S_3fS_cluster_68_S_3rU_S15.6385__P_5fU_S
2.2639_C_5fU_S2.23072_S2.547057
tatgcttcaatctacattcaaactccgtagttctttctttggatgtggatagcatttttcatcatgaggcctttggagatgtcttagatccttgT
ATTGCTGA
>P_3rS_cluster_31_cluster_252_P_3fS_S2.18623_S2.14939_S_3rU_2S17__P_5fU_S2.2639
```

```
_C_5fU_S2.23072_S2.547057
cttttccgatttgacagcattttccatcctgaggcctctggagatgtcttagatccttgTATTGCTGA
>P_3rU_2S8_P_3fU_7S8__C_5fU_3S8
ccgatctttgcctgaggttgcacacttggtcatatttgagctcaggcaactggggttaagtgacttgcccagagtcacaagtctgaggtcg
gatttgaa
>P_3rU_S8.12674_S8.39898__C_5fU_S8.950860
gacacctcggtgtgtctcagtttcctcatctataaTATGAGCTGGAAAAGGAAATGGCACACTACTCTA
GTATCTTTGCCAAGAAAACCC
>P_3rU_7S3__C_5fU_S3.480683
TATTGTGTCATGTAACCATATAGTATCTCATTGTAAACAAATCATTACATGCACAC
ATTCCCACCAATGCATGCTGGACTTCCTGACAAAGTACAACATGCTCACCTGCCA
ACACTTGCTTGTTAAGACCTGTCAGTGGACTTGACCACTGATGGGTTATAGCTTG
CATTT
>P_3rU_cluster_231__C_5fU_2S8
attcagcctcagacactttccagctgtgtgaccctgggcaagtcagttAACCCCGTCTGCCTCAGTTTCTCCATC
CATAAAATGAGCTGGAGAAAGAAATGGCAAACCACTCCAGGATCTTTGCCAAGA
AATCCCCAAATGGGG
>P_3rU_S8.36760__C_5fU_S8.958388
CATATCAAATGACTTGTCCACAGTCACACAGC
>P_3rU_S8.36760__C_5rU_4S8
CATATCAAATGACTTGTCCACAGTCACACAGCTAGTTTCAGAGGTGAGATTTGAA
>P_3rU_cluster_446_cluster_937__P_5rU_3S8_C_5rU_S8.15867
gaattagaaatcgaggagatgcccatcaattggggaatggctgaacaagtcatggtatAAGGAATTATGATGTGA
>P_3rU_cluster_446_cluster_937__C_5rU_S8.170262
gaattagaaatcgaggagatgcccatcaattggggaatggctgaacaagtcatggtatATGAATGTAATGGAATACT
ATCGTGCTATAAGAAAGA
>P_3rS_8S3_S8.47689__S_5rU_cluster_541_C_5rU_S8.50529
ATACACATAaattagagataagaggcagagttgcacagtcatcagcctcactttctc
>P_3rU_4S8_P_3fU_S8.70215__C_5rU_S8.1022487
ttaaagacagcaatcttctgtctattcttcaagattggtttCAGGAAACTTTAACTGGGGGCTGGAAAAGCA
TCCCATCATTTCTGACTTCCATCCTCCTTCTACTG
>P_3rU_6s3_P_3fS_3S2_2S9_S8.5461_S5.1115_2S4_3S3_S_3fS_cluster_29_cluster_977_c
luster_4_cluster_n1_cluster_28__S_5rS_clsuter_cp1_C_5fU_2S7_P_5fU_cluster_38_C_5rS_
S7.10526_2S10_2S3_S4.17612_S5.4886_S9.1880_S1.11056_2S8_S2.123759
gaacccacaaaaagacagaatgaaacaaatatccagcccaatacagcctggatgATTGATCAGAAGGGTGTATC
GCGCCTTGCTAGGAGCGCAGTGCAGCGCGGTGTGGGCGCACAGGCTGCAGCAAA
CCTGGAGCAGGCCTCAGACTGAATCATGGGCAGCTG
>P_3rU_6s3_P_3fS_3S2_2S9_S8.5461_S5.1115_2S4_3S3_S_3fS_cluster_29_cluster_977_c
luster_4_cluster_n1_cluster_28__S_5rU_S20.31638_C_5rU_S3.50689
gaacccacaaaaagacagaatgaaacaaatatccagcccaatacagcctggatgCTTGATCGGA
>P_3rS_cluster_31_cluster_252_P_3fU_3S2_S_3rS_2S17_S15.6385_S_3fS_cluster_68__S_
5rU_S18.27629_C_5fU_S2.1235481_C_5rU_S2.466916
GACAGCATTTTCCATCCTGAGGCCTCTGGAGATGTCTTAGATCCTTGTATTGCTGA
GAAGGGTTAAGTCTATTAATATTAGTACCTAACTGATTATGTTATTCTCTTCTTGA
GCCAAATCTGATGAGAGTAAGGTTCAAACAATGCTAATATCCGTC
>P_3rS_S5.47458_6S8__S_5rS_cluster_43_C_5fU_S8.854973
ATAAATGAGGAACCTGATATCCaaagaactgaaatgacttaccaaggtcacacagctgatgagtagcagaagca
agaagagaaacaaaatcttctgattcccaggttcctgccac
```

```
>P_3rU_cluster_231__C_5rU_S8.503819
attcagcctcagacactttccagctgtgtgaccctgggcaagtcagttAACCCCGTCTGCCTCAGTTTCC
>P_3rU_S8.12674_S8.39898__C_5rU_S8.911698
gacacctcggtgtgtctcagtttcctcatctataaAATGAGCTGGAGAAGGAAATGACAAACCACTCTA
GTATCTTTGCCAAGAAAACCCCAAATGAGATCA
>P_5fU_cluster_54.maffT3s8_C_5fU_S8.30495_S8.561570__P_3fU_S8.88040_S8.3563
ACCTGAGTTCAAATCCAGCCTCCGATACATACTAGCTGTGCGACCCGGGGCAAGC
CACT
>P_5fU_cluster_54.maffT3s8_C_5fU_S8.30495_S8.561570__P_3fU_6S8_3fU_S8.105580
ACCTGAGTTCAAATCCAGCCTCCGATACATACTAGctgtgtgacccggggcaagccact
>P_5fU_S2.26104_C_5fU_S2.82531__P_3fU_S2.2953
CTGGGAGTTAGGGAGGACCTGAGTTCAAATCCAGCCTCAGACACATAACACTTA
GCATATGTGATG
>C_5fU_S2.410158__S_3fU_S18.5221_P_3fU_S2.310
CATTTTTATTTATTCATACATACTTCCAATCATCAATATGAGAACCATTTTATGTG
CAATACATTGTGCTTTCCAGAACAGTGGAGCCAATTCCCAGCTCCACCAACAATG
CATCAGTG
>C_5fU_S8.589094__C_3fU_S8.917292_P_3fU_S8.49463
CCCAAGAAGTGGTATTGCTGGATCAAAGGGTATGCAGTTTTATAGCCCTTTGGGC
ATAGTTCCAAAT
>C_5fU_S3.480683__P_3fU_S3.28583_S8.82168_S3.90397_S3.64168_S_3fU_S20.10443
AAATGCAAGCTATAACCCATCAGTGGTCAAGTCCACTGACAGGTCTTAACAAGCA
AGTGTTGGCAGGTGAGCATGTTGTACTTTGTCAGGAAGTCCAGCATGCATTGGTG
GGAATGTGTGCATGTAATGATTTGTTTACAATGAGATACTATATGGTTAcatgacacaat
attgtgtcatcaaaattttgatgcaggggaaaaactcacaaatttacaataaatt
>C_5fU_S8.92543_S8.302794__P_3fU_S8.42372
CCCCATTTGGGGATTTCTTGGCAAAGATCCTGGAGTGGTTTGCCATTTCTTTCTCC
AGCTCATTTTATGGATGGAGAAACTGAGGCAGACGGGGTTAACTGACTTGCCCAG
GGTCATACAACTAGGAAGTGTCTGAGGCCAGATTTGAATCCAAGAAGATAAGTC
CTCCTGACTCCGGGTTTGGCAGTCTGTCCACTATGAC
>C_5fU_S3.102282_S3.954929__S_3fU_cluster_195_P_3fU_S3.39389
ATATTATATTCCATGCCCAGCGGTCCTTTAATGTAGaagctgctaaaacttgtgttatcctgattgtgttt
ccactatacttgaattgtttctttcttgcagcttgtaatatat
>C_5fU_S8.242952_S8.249062__S_3fU_S17.14523_C_3fU_S8.70423_C_3rU_cluster_339
TGCTTGGAACCATCGGTTATAGcaaatggagaagttttgaactctgtggataagttcacttacctcggtagtgtacta
>P_5rU_cluster_19_C_5rU_S8.10722__S_3fU_S17.14523_C_3fU_S8.70423_C_3rU_cluster
_339
TTTCTCCACCAGCTGGCACCACATCATCTATGCTTGGAACCATCGGTTatagcaaatgga
gaagttttgaactctgtggataagttcacttacctcggtagtgtacta
>C_5rU_2S3__S_3fU_4S20_P_3fU_S3.74019
GGCCGGTGCTCTATTCACTGTGCCACCTAGATGCCCCTGAAGAATATATTTTAGG
CATATAAATGTGTAT
>C_5rU_S2.906334__P_3fS_5S3_S2.110884_S8.68922
GGTCACCCAGCTAGTAAATATCTGAGGCCAGATTTGAGTCTTTCTGACTTCAGGC
CCTGCACTTTATTCACTGTGCCACCTAGATGAGATCG
>S_5rU_cluster_645.maffT2s20_C_5rU_S3.16848__S_3fU_S20.29912
gaccagactgattagaagcataactacaaaattctgattattgCAATAACCCTTAAGTATAATATTCCAATTA
AGACATCCAAGAGTCACATTTAAATATTGCACTCTTC
```

```
>C_5rU_S8.108028__P_3fU_S8.380
GAATGGGGCACAGTAGTAATAACATTAAAAAGACACACAACTTTGAGAGAATTA
AGGACTTTGATCAACCTAATGACTAACCACAGTTCCAG
>C_5rU_S8.490159__P_3fU_S8.380
GAAGGAGCAGAAATAACATTAAAAAGACACACAACTTTGAGAGAATTAAGGACT
TTGATCAACCTAATGACTAACCACAGTTCCAG
>C_5rU_S8.503819__P_3fU_S8.42372
GGAAACTGAGGCAGACGGGGTTAACTGACTTGCCCAGGGTCATACAACTAGGAA
GTGTCTGAGGCCAGATTTGAATCCAAGAAGATAAGTCCTCCTGACTCCGGGTTTG
GCAGTCTGTCCACTATGACA
```

3.7.2 Paired 5' and 3' integration sites from different koalas

```
>P_3rS_S2.74718_S3.57590_11S8__C_5fU_3S8
TATTTGAGCTCAGGCaactggggttaagtgacttgcccagagtcacaagtctgaggtcggatttgaa
>P_3rU_4S8_P_3fU_S8.70215__C_5fU_S4.684182_5fU_S4.1661030
ttaaagacagcaatcttctgtctattcttcaagattggtttGCCATTTCCTTCTCCAGCTCATTTTATGGAT
>P_3rU_4S8_P_3fU_S8.70215__C_5fU_S4.1683056
ttaaagacagcaatcttctgtctattcttcaagattggtttCCTCATCTGTAAAATGGGGATAATAACA
>P_3rU_S3.26225__C_5fU_S8.589094
AATTGAGGAACTATGCCCAAAGGGCTATAAAACTGCATACCCTTTGATCCAGCAA
TACCACTTCTTGGG
>P_3rU_S8.97760__C_5fU_S3.480683
ACTGAGCCATATAACCATATAGTATCTCATTGTAAACAAATCATTACATGCACAC
ATTCCCACCAATGCATGCTGGACTTCCTGACAAAGTACAACATGCTCACCTGCCA
ACACTTGCTTGTTAAGACCTGTCAGTGGACTTGACCACTGATGGGTTATAGCTTG
CATTT
>P_3rS_S5.105847_5S8__C_5fU_S4.1481035
TGAATGAACATTTTCTCTACTCCGCCATCTTGGCTCCACCCCCC
>C_3rU_S8.1018030_P_3rS_S4.143028_6S3_S2.99154_17S8_P_3fU_2S8_S_3fU_S20.580
49__P_5rU_S9.30145
ggctgattcggactcaggtgagtcttccaactccagggctggcactctatccattcccccacctacctgccctccccacATTCCTT
CAAACCCTCTGTC
>P_3rS_S3.41285_S4.156475_S5.66477_S7.113364_S8.73375_S9.5274_36S2__S_5rU_S18
.20949_C_5rU_S2.111871
AGGAGGTTGAACCAGATGACCTCTGGGGTCTCTTTTAGCC
>S_3rU_S17.112853_P_3rU_8S2__S_5rU_S23.5584_2S20_C_5rS_3S3_S7.4344
GTAAATTGGTCTGAGGCTGGATTTGAACTCAGGTC
>P_3rU_4S8_P_3fU_S8.70215__C_5rU_S4.1173169
ttaaagacagcaatcttctgtctattcttcaagattggtttGTCATTTCCTTCTCCAGCTCATG
>P_3rS_S2.74718_S3.57590_11S8__C_5rU_S9.964451
TATTTGAGCTCAGGCAACTGGGGTTAAGTGACTTGCCAGATCGGAAGAGCGT
>C_5fU_S2.547057__P_3fU_S4.4833
TCAGCAATACAAGGATCTAAGACATCTCCAA
>C_5fU_S8.1066112_S8.1022169_S8.1055816__S_3fU_S20.2502
ttcaaatccgacctcagacttgtgactctgggcaagtcacttaaccccagttgcctCAGATCCAATTCACAT
>C_5fU_S4.684182_5fU_S4.1661030__S_3fU_S17.64559_P_3fU_S8.56518
```

```
ATCCATAAAATGAGCTGGAGAAGGAAATGGCAAACTAGTTTCCCAGTCTATTTCA
TCCGGAGTTGT
>C_5fU_S4.1683056__S_3fU_S17.64559_P_3fU_S8.56518
TGTTATTATCCCCATTTTACAGATGAGGAAACTAGTTTCCCAGTCTATTTCATCCG
GAGTTGT
>C_5fU_S2.729164__S_3fU_4S20_P_3fU_S3.74019
CTTTATTCACTGTGCCACCTAGATGCCCCTGAAGAATATATTTTAGGCATATAAAT
GTGTAT
>C_5fU_S4.1481035__P_3fU_S8.33956
GGGGGGTGGAGCCAAGATGGCGGAGTagaagatgaggaaaaggctcacagaagg
>S_5rU_cluster_620_C_5rU_2S8_P_5rU_S8.78052__P_3fU_S3.22152
agaggtccaggcaaagagCTATGATCCCCGGTTTCTGCTTTCCTTCTAGTTAAATCGGA
>C_5rU_S2.906334__S_3fU_4S20_P_3fU_S3.74019
GGTCACCCAGCTAGTAAATATCTGAGGCCAGATTTGAGTCTTTCTGACTTCAGGC
CCTGCACTTTATTCACTGTGCCACCTAGATGCCCCTGAAGAATATATTTTAGGCAT
ATAAATGTGTAT
>C_5rU_2S3__P_3fS_5S3_S2.110884_S8.68922
GGCCGGTGCTCTATTCACTGTGCCACCTAGATGAGATCG
>C_5rU_S4.1173169__S_3fU_S17.64559_P_3fU_S8.56518
CATGAGCTGGAGAAGGAAATGACAAACTAGTTTCCCAGTCTATTTCATCCGGAGT
TGT
>S_5rU_S23.5584_2S20_C_5rS_3S3_S7.4344__P_3fS_8s2_1s5
GACCTGAGTTCAAATCCAGCCTCAGACAAATTCCAAGAAAAAGTTAGCTCTTTCC
CCTTCCTCCCCCTCCTGTGCCAT
>P_5rU_S4.11115__S_3fU_S20.10939_P_3fU_S3.74062
TTCAAGCAGAAGACGGCATACGAGATAGAGGCGGTGACTGGAGTTCAGACGTGT
GCTTTGCCGATCTGAAGACACAGTAGTCAATGACTATAGTAGTCTTC
```

Table S1. Indexing primers for indexed Illumina library construction

index primer sequence*	**Experiment Method**
CAAGCAGAAGACGGCATACGAGATgccatctGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATaacctggGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATctaacggGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATagaggcgGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATccgcaagGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATctccgccGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATacgtccaGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATcatggttGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATcttcctgGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATaggtatgGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATggattggGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATacgccggGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATgcggcaaGTGACTGGAGTTCAGACGTGT	PEC and hybridzation capture
CAAGCAGAAGACGGCATACGAGATtgatagGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATtatacgGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATcgatgaGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATatacacGTGACTGGAGTTCAGACGTGT	SPEX

CAAGCAGAAGACGGCATACGAGATatagcgGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATtgttcaGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATagatacGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATtagctgGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATgtatgtGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATggctcaGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATcatgctGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATtcatcgGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATcatctaGTGACTGGAGTTCAGACGTGT	SPEX
CAAGCAGAAGACGGCATACGAGATgtcacaGTGACTGGAGTTCAGACGTGT	SPEX

*** The Illumina indice (6-7bp long) are embeded in the primers in lower letters**

Table S2. Primers and baits used in the experiments

Primer name	**Sequence in 5' to 3'direction**
SPEX_Primer_3.1	Biotin- ATTTGCATCCGGAGTTGTGT
SPEX_Primer_5.1	Biotin- CGGAATGATTTCTGCCTCAT
SPEX_Primer_3.2	Biotin- AGTTGTGTTCGCGTTGATCC
SPEX_Primer_5.2	Biotin- TTCCATACTCCACGGAATGA
SPEX-2R_illumina P7	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATATATGGGIIGGGIIGGGIIGGG
SPEX-2F_5_illumina P5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGGAATGATTTCTGCCTCAT
SPEX-2F_3_illumina P5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTGTGTTCGCGTTGATCC
HybCap_KoRV3LTR_F	Biotin- TCAAGGACATCC*GATTTGCATCCGGAGTTGTGTTCGCGTTGATCC
HybCap_KoRV5LTR_R	Biotin- TCAAGGACATCC*GTTCCATACTCCACGGAATGATTTCTGCCTCAT
PEC_Primer_3.1	Biotin- CAAGGACATCC*GATTTGCATCCGGAGTTGTGT
PEC_Primer_3.2	Biotin- CAAGGACATCC*GAGTTGTGTTCGCGTTGATCC
PEC_Primer_5.1	Biotin- CAAGGACATCC*GCGGAATGATTTCTGCCTCAT
PEC_Primer_5.2	Biotin- CAAGGACATCC*GTTCCATACTCCACGGAATGA

I = deoxyinosine.

*** phosphorothioate bond to render the oligonucelotides resistant to nuclease degradation**

Chapter V
Concluding remarks

Concluding remarks

ERVs are diverse and exist in high copies in vertebrate genomes. Despite the evolutionary and medical importance of ERVs study, the methodology for comprehensive profiling of retroviral genomic changes and integration sites are far from sufficient or efficient, especially for the challenging task of retrieving them from historical or ancient DNA samples, which can obscure the answers to important questions about endogenization and evolution of ERVs.

Unlike most ERVs, the koala retrovirus (KoRV) is in transition from an exogenous to an endogenous state. The full KoRV genome retrieval from six museum koalas described in **Chapter II** of the thesis and the comparison of three sequence target enrichment techniques for the comprehensive profiling of KoRV integration sites from 10 museum koalas described in **Chapter III** of the thesis indicate that hybridization capture is currently the best experimental approach to ancient ERV analysis. And the bioinformatic pipeline in **Chapter III** of the thesis characterized over 1000 KoRV integration sites from the 10 museum koalas, which provides the ERV researchers an efficient *in silico* tool for EVR identification without a host reference genome. Furthermore, comparison of the KoRV flanks characterized in **Chapter III** to previously described integration sites from other koalas suggest that the proportion of KoRV integration sites shared among unrelated koalas may have increased over the past 140 years.

Foamy retroviruses have been previously identified as undergoing co-evolution at higher taxanomic levels, but their co-evolution and co-divergence more recently among sloths is not well studied due to lack of both retroviral and host sequence information from extinct sloths. Therefore, I modified (**Chapter IV**) hybridization capture technique to allow for enrichment of a diversity of target-related sequences and established an efficient pipeline based on optimization of MITObim (baiting and iterative mapping) for re-construction of ancient DNA sequence when only distant extant relative can be used as genetic reference. I applied these methods for illumina targeted sequencing of full mitochondrial genomes (mitogenomes) and partial polymerase gene of SloEFV (sloth endogenous foamy virus), from two extinct and three extant sloth species. The mitochondrial capture results produced a strongly supported phylogeny for extinct ground and living tree sloths that conflicts with recent morphological analyses. And comparison of SloEFV *pol* gene tree to the mitochondrial phylogeny of both extant and extinct sloths demonstrates multiple complex invasions of SloEFV into the ancestral sloth germline line followed by subsequent introgressions across different sloth lineages.

In this thesis, different modifications of hybridization capture were tested, including long bait (500bp) (**Chapter II**), short bait (30-40bp) (**Chapter III**), medium size bait (200-

300bp) (**Chapter IV**) as well as the triplication of pre- and post- capture amplification (**Chapter IV**). Each of these modifications was effective for a specific ERV study.

Rapid progress in the development of sequencing technologies will offer the scientific community new opportunities to sequence DNA and RNA sequences in longer reads, with less starting material, and at lower cost. However, before sequencing cost drops to a point where whole genome sequencing of a vertebrate is commonplace, targeted sequencing will still be the needed for molecular retrovirology. Similarly, until computational power is robust enough to sift through unenriched data, including highly degraded aDNA, targeted enrichment will have enormous computational benefits by reducing the complexity of sequence data sets. Therefore, target sequence enrichment technique will still need development and will likely extend in two directions:1) longer target DNA fragments to enable investigation of large indels and structural mutations of target DNA , and to link the polymorphism apart beyond the illumina sequencing read length, and to assign integration sites to related proviral insert; 2) higher sensitivity allowing for retrieval of sequence information from degraded genetic samples, e.g. aDNA. The high throughput sequencing era provides a rich resources for retroviral investigation *in silico*. Various computational tools have been developed: RetroTector, RepBase, RepeatMasker and HERVd. Unfortunately, they all have limitations. For example, the most widely used program RetroTector is only efficient for detecting relatively genome-intact proviral loci and often fail to detect novel retroviruses which share low similarity with known retroviruses. Advancement in bioinformatics will need to solve these problems and more computational tools will be developed with higher sensitivity and less CPU-time consuming.

With the development of sequencing technologies, target sequence enrichment techniques and bioinformatic tools, the molecular invesitigation of ERVs would eventually reach single proviral locus resolution. More novel EVRs and more structural variations will be found (Hayward et al 2015).

Reference

Hayward, A., C. K. Cornwallis, et al. (2015). "Pan-vertebrate comparative genomics unmasks retrovirus macroevolution." Proceedings of the National Academy of Sciences **112**(2): 464-469.

For reasons of data protection, the curriculum vitae is not published in the electronic version.

For reasons of data protection, the curriculum vitae is not published in the electronic version.

For reasons of data protection, the curriculum vitae is not published in the electronic version.