Genetic diversity of Koala retrovirus (KoRV) *env* gene subtypes: Insights into northern and southern koala populations

Nishat Sarker[1,6*], Jessica Fabijan[2], Jennifer Seddon[1], Rachael Tarlinton[3], Helen Owen[1], Greg Simmons[1], Joshua Thia[4], Adam Mark Blanchard[7], Natasha Speight[2], Jasmeet Kaler[3], Richard David Emes [3,5], Lucy Woolford[2], Darren Trott[2], Farhid Hemmatzadeh[2], and Joanne Meers[1*]

[1]School of Veterinary Science, The University of Queensland, Australia
[2]School of Animal and Veterinary Sciences, The University of Adelaide, Australia
[3]School of Veterinary Medicine and Science, University of Nottingham, United Kingdom
[4]School of Biological Sciences, The University of Queensland, Australia
[5]Advanced Data Analysis Centre (ADAC), University of Nottingham, United Kingdom
[6]Laboratory Sciences & Services Division, International Centre for Diarrhoeal Disease Research, Bangladesh
[7]School of Animal, Rural and. Environmental Sciences, Nottingham Trent University, United Kingdom

*Corresponding authors:
(a) Professor Joanne Meers
Email: j.meers@uq.edu.au

(b) Nishat Sarker
Email: nishat.sarker@icddrb.org; n.sarker@uq.edu.au

Key words: KoRV, env subtype, koala, KoRV-B, genetic diversity

Repositories: Sequence reads are uploaded in NCBI database, Sequence Read Archive under accession number SRR8375764

Abstract

Koala retrovirus (KoRV) is a recently endogenised retrovirus associated with neoplasia and immunosuppression in koala populations. The virus is known to display sequence variability and to be present at varying prevalence in different populations, with animals in southern Australia displaying lower prevalence and viral loads than northern animals. This study used a PCR and next generation sequencing strategy to examine the diversity of the KoRV *env* gene in both proviral DNA and viral RNA forms in two distinct populations representative of the "northern" and "southern" koala genotypes. The current study demonstrated that the full range of KoRV subtypes is present across both populations, and in both healthy and sick animals. KoRV-A was the predominant proviral subtype in both populations, but there was marked diversity of DNA and RNA subtypes within individuals. Many of the northern animals displayed a higher RNA viral diversity than evident in their proviral DNA, indicating relatively higher replication efficiency of non-KoRV-A subtypes. The southern animals displayed a lower absolute copy number of KoRV than the northern animals as reported previously and a higher preponderance of KoRV-A in individual animals. These discrepancies in viral replication and diversity remain unexplained but may indicate relative protection of the southern population from KoRV replication due to either viral or host factors and may represent an important protective effect for the host in KoRV's ongoing entry into the koala genome.

1. Introduction

Koala retrovirus (KoRV) is reportedly the youngest endogenized retrovirus (virus integrated in the host's germline and inherited as part of its host's genome), having been integrated in the koala genome only about 22,200–49,900 years ago (Ishida et al., 2015). The low genetic diversity of the long terminal repeat (LTR) regions of KoRV provirus sequences suggests that endogenous KoRV probably arose as part of a single outbreak (Ishida et al., 2015). KoRV is also evident in an apparently exogenous (horizontally infectious) form (Hobbs et al., 2017), with many koalas possessing high levels of KoRV RNA in plasma reflecting active viral replication. KoRV is found at a proviral prevalence of 100% in koala populations in northern regions of Australia (Tarlinton et al., 2005; Simmons et al., 2012) and at much lower prevalence (15-25%) in southern populations (Simmons et al., 2012; Legione et al., 2017). This range of prevalence has led to suggestions that KoRV is currently spreading throughout the Australian koala population following a 'northern to southern' transmission wave (Fiebig et al., 2006; Tarlinton et al., 2008). Other recent work has indicated that, in southern populations that have apparently low KoRV prevalence, KoRV-negative animals may in fact have defective versions of the virus, missing the *pol* and *env* gene portions most commonly used in virus detection studies (Tarlinton et al., 2017).

The outcome of natural KoRV infection is variable, and little is known about the mechanisms of pathogenesis. KoRV is associated with neoplasia and immunosuppression leading to clinical chlamydial disease (Canfield et al., 1988; Hanger et al., 2000; Xu et al., 2013; Fabijan et al., 2017; Gonzalez-Astudillo et al., 2017; Nyari et al., 2017; Burnard et al., 2018). For retroviruses in other species, mutation or recombination events in *env* genes play a significant role in pathogenicity, such as immunosuppression induced by subtypes of feline leukaemia virus (FeLV) (Overbaugh et al., 1988; Anderson et al., 2000; Chandhasin et al., 2005a; Chandhasin et al., 2005b).

Classification of KoRV subtypes (in line with similar naming schemes used for better studied gammaretroviruses such as FeLV) is based around the nucleotide sequence of the *env* gene, which encodes the surface protein (SU) and transmembrane protein (TM) of the virus. The 5' end of this gene, known as the hypervariable region, is of particular importance in subtype classification. This region of the *env* gene encodes the protein most exposed to the host's immune response as it is external to the virus membrane and is therefore typically the most variable portion of a retrovirus. This region of the virus also determines host receptor specificity (and therefore cellular tropism) and is one of the major determinants of pathogenicity in FeLV (Bolin and Levy, 2011). The classification system of KoRV is based around phylogenetic groupings of *env* gene nucleotide sequences. For some subtypes (though not all) receptor binding differences have also been determined. The generally accepted classifications are: KoRV-A (Hanger et al., 2000), which binds to the sodium-dependent phosphate transporter Pit-1, KoRV-B and J which bind to the thiamine transporter encoded by THTR1 (Xu et al., 2013), KoRV-C, KoRV-D (Shojima et al., 2013), KoRV-E, KoRV-F (Xu et al., 2015), KoRV-G, KoRV-H, and KoRV-I (Chappell et al., 2017). The receptor usage of subtypes C, D, E, F, G, H and I have not been determined. KoRV-A is found in every KoRV-positive koala and is considered the endogenous version of KoRV from which other subtypes have arisen (Chappell et al., 2017; Hobbs et al., 2017). The other subtypes of KoRV are possibly not germ line transmitted, as they were present in only low copy number in the koala reference genome animal, and as such were considered putative somatic insertions (Hobbs et al 2017). The same study also reported that KoRV-D and KoRV-E were present only as defective viruses and the authors hypothesised that these subtypes may be transmitted with a replication competent "helper" virus as has been documented for other retroviruses (Hobbs et al., 2017). KoRV-B is thought to be more pathogenic than KoRV-A, having been reported at an increased prevalence in animals with chlamydiosis or neoplasia than in healthy animals (Waugh et al., 2017). KoRV B and J isolates have also been reported to have variable numbers of copies of repeat regions (these are present as single copies in all KoRV A variants) in their LTRs. These types of repeats are known to enhance replication efficiency in other retroviruses such as FeLV (Xu et al., 2013; Chaban et al., 2017; Waugh et al., 2017).

This study explored the evolutionary patterns of KoRV *env* gene subtypes in two koala populations. Patterns of KoRV genetic diversity were investigated in one "northern" genotype in South-East

Queensland (QLD) with a KoRV prevalence of 100% and one “southern” genotype in the Mount Lofty Ranges, South Australia (SA), with an unknown prevalence (Figure 1). Patterns of KoRV *env* subtype diversity were compared in paired DNA and RNA samples in a subset of koalas from both populations to understand *env* gene variation in integrated proviral genome (DNA) and in circulating virus (RNA). Further, this study examined the possible relationship of *env* gene subtypes with a diverse range of clinical diseases.

2. Results

This study assessed *env* gene diversity of both integrated KoRV provirus DNA and expressed plasma viral RNA. Thirty-three “northern” (Queensland, QLD) koalas were assessed, comprising 28 with paired DNA and RNA samples and five with only DNA as plasma was not available. Twenty-eight “southern” (South Australia, SA) koalas were included, comprising five with paired DNA and RNA samples, and 23 with only DNA samples because plasma was not available (10 animals) or the RNA sample was negative in the *env* gene RT-PCR (13 animals). The overall summary of the demographic details and clinical status of the animals is shown in Table 1 and details are in supplementary file 2.

After quality evaluation and filtration, an average of 22719 total reads were generated for each provirus DNA sample, ranging between 6169 to 59558 total reads. However, the number of total reads was higher from amplified RNA, averaging 104894 total reads with a range of 12717 to 245827.

Table 1: Overall details of study samples (percentages of the total number of animals are given in parentheses). Not all information was available for all koalas.

Variables	Characters	Queensland	South Australia
Sex	Male	n= 20 (60.6%)	n= 14 (50%)
	Female	n= 13 (39.4%)	n= 14 (50%)
	Total	**n= 33**	**n= 28**
Age group	Juvenile	n = 2 (6.3%)	n = 3 (11.5%)
	Young adult	n = 8 (25%)	n = 14 (53.8%)
	Adult	n = 22 (68.8%)	n = 9 (34.6%)
	Total	**n = 32**	**n = 26**
Body condition score	0 to <3	n = 15 (46.9%)	n = 4 (15.4%)
	≥ 3 to 5	n = 17 (53.1%)	n = 22 (84.6%)
	Total	**n = 32**	**n = 26**
Clinical status	Healthy	n = 11 (33.3%)	n = 5 (17.9%)
	Neoplasia	n = 8 (24.2%)	n = 5 (17.9%)
	Oxalate nephrosis	negative	n = 5 (17.9%)
	Chlamydiosis	n = 16 (48.5%)	n = 11 (39.3%)

2.1 *Env* subtypes

From deep sequencing, a total of 169 unique sequences were generated after sequence validation from all samples of both populations. Sequence reads are available under Sequence Read Archive accession number SRR8375764. The sequence alignment is shown in supplementary file 3. The Bayesian phylogenetic tree (Figure 2) showed high genetic diversity in the KoRV *env* gene at the population level. The identified sequences were grouped with previously recognized subtypes A, B, D, and I. Subtypes B and I were monophyletic in the tree with posterior probability support of 1.0. Subtype A formed a well-supported monophyletic clade. In contrast, subtype D exhibited multiple sub-clades with relatively long branches, high posterior probability values and sequences that were divergent in the hyper-variable region of the env receptor binding domain (RBD) (Supplementary Files 3 and 4). This subtype D grouping, with posterior probability support of 0.93, contained reference sequences that had been previously identified as subtypes F, G and H in addition to previously assigned subtype D sequences (Chappell et al., 2017). We identified 13 distinct sub-clades within the subtype D grouping; each of these sub-clades was strongly supported with posterior probabilities of 0.99-1.0. Two of these sub-clades contained only KoRV-G or KoRV-H sequences, and for consistency with the literature, we retained these names despite their phylogenetic placement as sub-clades within subtype D. The remaining 11 sub-clades within subtype D were designated D1 to D11. Sequences that had previously been designated KoRV-F by different authors (Xu et al., 2015; Chappell et al., 2017) belonged to two different subtype D sub-clades in our analysis, with the Chappell et al. (2017) sequence strongly clustering with sub-clade D3 and the Xu et al. (2015) sequence clustering with moderate support (posterior probability 0.83) with D9. Previously assigned KoRV-D sequences clustered either with sub-clade D1 (Shojima et al., 2013) or D4 (Chappell et al., 2017). Sequences clustering in sub-clades D2, D5, D6, D7, D8, 10 and D11 were not matched or clustered with any reference sequences. Subtypes C and E, which have been previously described (Miyazawa et al., 2011; Shojima et al., 2013; Xu et al., 2015), clustered together close to but distinct from group D.

Sequences in this study were assigned to all of the afore-mentioned subtypes except for subtypes C, E and H, which were not found in any of our samples. A total of 63 sequences were found from the KoRV-A subtype, 22 from subtype B, 16 from subtype I, and 68 from subtype D. Within the subtype D sequences, most (22) sequences were within the D1 sub-clade, with one sequence in each of D4 and D10, two in each of D6, D8 and KoRV-G, four sequences in each of D2 and D7, five in D3, six in D5, eight sequences in D9, and 11 sequences in the D11 sub-clade.

The average read count of each unique sequence was highly diverse between animals, between RNA and DNA forms in the same animal and within the two koala populations (QLD and SA). The number of different KoRV DNA sequences within an individual, indicative of the number of provirus insertions, was significantly higher (Mann Whitney U, p< 0.0001) in QLD individuals (median 77, range 63-100) in comparison to SA koalas median 59, range 43 – 74).

The read count details of unique sequences in individual animals are available in Supplementary File 5 and the relative percentage levels of each subtype (group of sequences) are available in Supplementary File 6. There was variation among individuals in overall read count and so read counts of subtypes were converted to the relative percentage of each subtype of the total reads for that individual with following equation:

> Relative percentage of each specific subtype (for each individual) = (total number of unique sequence reads of the subtype / total read count of all subtypes) x 100

2.2 Env subtype abundance in QLD and SA koala population

As shown in Figure 3, the absolute read count values of *env* gene subtypes in the DNA of QLD koalas were significantly higher (p value <0.0001) in comparison to SA koalas. Each koala had multiple *env* subtypes in their genome (Figure 4). In the QLD animals, KoRV-A, KoRV-B, and KoRV-D (sub-clades 10 and 11) were present in proviral DNA form at some level in all individuals (supplementary file 5 and 6) KoRV-A was the dominant subtype (present at >40% of reads) in proviral DNA in 30 of the 33 QLD animals (Figure 4A). The exceptions were two animals where the D1 subclade was in higher abundance and one koala with the D6 subclade as the highest abundance. For the SA koalas, subtypes A, B and sub-clade D10 were present in the DNA of all koalas, while sub-clades D1 (27/28) and D11 (27/28) were also represented in the majority of koalas. KoRV-A dominated the proviral DNA subtypes in all SA koalas with a much higher relative percentage than in the QLD animals. The median Shannon diversity index was also significantly lower in the SA than the QLD animals (Mann Whitney U (<0.0001).

The subtypes present in the RNA of individual animals differed from those present in the DNA. The median Shannon diversity index was lower for DNA than RNA samples for both populations but did not reach significance (Mann Whitney U test, QLD p=0.4 and SA p=0.5). For the QLD animals subtypes A, B, D2, D3 and D5 were more abundant in the DNA samples and D10 and G more abundant in the RNA samples (FDR p values >0.005). All QLD (n= 28) and SA koalas (n=5) had subtypes or subclades A, B, D1, D10 and D11 at some level in their viral RNA; additionally all SA koalas also had subclades D2 and D3.

All five SA koalas showed a very high relative percentage of KoRV-A (92.5-99.9%) in viral RNA. In contrast, none of the RNA samples for QLD koalas showed KoRV-A to be the most abundant subtype, with KoRV-B (n=8), D11 (n=7), D1 (n=3), D3 (n=2), KoRV-I (n=1), D2 (n=1), D5 (n=2), D6 (n=2), D7 (n=1) or D8 (n=1) the most abundant subtypes or subclades within individual koalas (supplementary file 6)

2.3 Distribution differences between viral DNA and RNA of env subtypes

The difference in the distribution of subtypes between viral DNA and RNA of individual koalas was striking, in particular amongst the QLD koalas (Figure 4A and 5 A,C). In some koalas, the predominant viral RNA subtype formed only a very minor proportion of the proviral DNA subtype distribution. As examples, koalas Q2 and Q27 had an overwhelming predominance of subtype/subclades D2 and D8 in their viral RNA, comprising 85% and 88% of their RNA subtype distribution, respectively, whereas these two subtypes comprised only 13% and 8%, respectively, of the proviral DNA subtype distribution in these koalas. Within individual koalas, it is clear that some KoRV proviral subtypes have very high rates of expression while others are poorly expressed.

These results probably reflect a greater replication rate (and overall viral load) in the QLD animals with viral diversity increasing in the RNA form of the virus (a greater number and range of non KoRV-A subtypes being produced). In the SA animals where the viral load (and presumably the replication rate) is lower this difference in viral diversity is not seen, with these animals continuing to display a higher preponderance of the ancestral A subtype (both when compared to the QLD animals and when RNA and DNA forms within the SA animals are compared).

2.4 KoRV-A and KoRV-B status based on conventional PCR

Conventional PCR of DNA using KoRV-A and KoRV-B specific primers demonstrated a 100% prevalence of KoRV-A in QLD koalas and 96.4% in SA koalas, while the KoRV-B prevalence was 48.5% in QLD and 0% in SA koalas. This is in contrast to the MiSeq deep sequencing results where all animals in both populations were positive for both subtypes. There was a significant differences ($p<0.0001$) between a positive test for KoRV B with conventional PCR and a Miseq read count of >2700 for KoRV B. With one exception, koalas with raw read counts below 2700 were negative by KoRV-B specific conventional PCR.

2.5 Subtype correspondence with clinical status of respective koalas

Amongst the 33 QLD koalas, 11 were clinically healthy, 13 had chlamydiosis, four had neoplasia, four had both chlamydiosis and neoplasia and one had a non-neoplastic hepatic mass (Table 1 and Supplementary File 2). All DNA and RNA subtypes, including the putative pathogenic KoRV-B, were found in both healthy and diseased animals. Of the 28 SA koalas, 12 had chlamydiosis, five had neoplasia, five had oxalate nephrosis (a genetic kidney disease not commonly found in QLD animals), and six were healthy. As with the QLD animals, all SA koalas had both KoRV-A, KoRV-B and subclades of KoRV-D.

There were too few animals (particularly in the SA population) with RNA for a sensible analysis of disease status vs viral subtypes. For proviral DNA there was no clear association between the abundance of any particular subtype and any particular disease syndrome. Graphs of subtypes A, B, combined D and I versus disease categories of healthy, neoplasia, oxalate nephrosis and chlamydiosis for each population are presented in supplementary file 7. There was a trend towards healthy animals (in both the QLD and SA populations) and the oxalate nephrosis animals in the SA population (this disease is thought to have a genetic basic) having a lower viral diversity (with a greater preponderance of KoRV A) than those with neoplasia or chlamydiosis (the diseases thought to be associated with KoRV infection) though a major confounding factor for more robust analysis here was the number of animals with multiple disease syndromes and the small number of animals in some disease categories in each population.

3. Discussion

Despite the high prevalence of KoRV and its potential impact on the health of koalas, there are few reports available about KoRV genetic diversity in the Australian koala population. Most of the information about KoRV diversity comes from studies in overseas captive koalas (Miyazawa et al., 2011; Shojima et al., 2013; Xu et al., 2013; Xu et al., 2015) or wild South-East Queensland (SE QLD) koalas (Chappell et al., 2017), all of which are of the "northern" or mixed genotypes. Here, we made a substantial contribution to knowledge in this field by investigating KoRV *env* gene diversity in diseased and healthy koalas from both "northern" and "southern" genotype populations (SE QLD and Mt Lofty Ranges, SA) highlighting the differences in abundance of KoRV subtypes at both DNA and RNA level between these populations, with the southern animals demonstrating both a lower viral load, a reduced viral diversity and a greater preponderance of KoRV A abundance. The paired DNA-RNA samples in individual koalas also demonstrated that the abundance of different DNA and RNA subtypes within individual koalas do not correspond to each other, with a trend (though not significant) towards a higher diversity in the RNA samples, indicating variable expression of proviral DNA subtypes.

Our study demonstrated that the full range of KoRV subtypes was present in both northern and southern koala populations, and in both DNA and RNA forms of the virus. The finding of KoRV-B in all southern animals studied was unexpected and is in contrast to recent PCR-based studies of KoRV-B prevalence which have reported varying (Waugh et al., 2017) or absent (Legione et al., 2017) prevalence rates of KoRV-B in "southern" animals.

Phylogenetic analysis of the KoRV *env* genes in this study found four major subtypes; three were strongly supported monophyletic clades clustering with previously designated as A, B and I. The fourth subtype was the large paraphyletic group D, which this study classified into 13 sub-clades comprising previously designated subtypes G and H and newly designated subtypes D1 to D11. Two reference sequences that had previously been designated KoRV-F clustered with two distinct group D subtypes (KoRV-D3 and KoRV-D9). The paraphyletic nature of the subtype D grouping highlights the difficulties of assigning KoRV subtypes. Rather than following convention and designating our newly identified sequences as further alphabetical subtypes (KoRV-K, L, M, N, etc), we recognised that these sequences belong to a large phylogenetic grouping and should not be classified as distinct lettered subtypes, but rather as sub-clades of subtype D. We cannot entirely rule out PCR related recombination of differing loci or PCR based errors in the sequences (particularly for the KoRV-D group) though the parameters set for including sequences in subsequent phylogenetic analysis (sequences present in at least two animals, a minimum read count of four and a 99% clustering threshold) will have removed sequences that appeared only once in the data.

KoRV-A, KoRV-B, and KoRV-D sub-clades 1, 10 and 11 were highly prevalent in individuals in this study, while KoRV-C, E and H were not identified. KoRV-C was identified at a Japanese zoo from captive koalas (Shojima et al., 2013) and to date has not been found in any wild koala (Chappell et al., 2017; Hobbs et al., 2017). KoRV-E was identified from a zoo in USA (Xu et al., 2015) and was also not found by Chappell et al. (2017), although a defective form is present in the reference genome animal (Hobbs et al., 2017). KoRV-H is rare, having been found in only one animal in viral RNA form (Chappell et al., 2017); in our study, KoRV-H sequences clustered within a larger KoRV-D clade, so it is possible that this subtype may exist in other geographic ranges. Overall these data highlight the extreme intra-animal variability of KoRV with many subtypes being reported in only a small number of animals.

This study is not able to distinguish between endogenous (incorporated into the genome and vertically transmitted) and exogenous (horizontally transmitted) virus. Indeed, in a newly integrated virus like KoRV, this distinction may not be very helpful as there is no reason why the virus cannot be both vertically and horizontally transmitted. The original demonstration of KoRV as an endogenous virus (Tarlinton et al 2006) did not use methods that would distinguish the different KoRV subtypes, though subsequent analysis of the variants present in the reference genome animal (of the northern genotype) (Hobbs et al., 2017) indicated that KoRV-A is endogenous in this animal (present at high copy number) while variants B-I are likely present only as low copy number somatic cell insertions (and so likely not vertically transmitted) though to date this has only been examined in this one individual. A number of sequencing efforts from museum specimens have only demonstrated KoRV-A and not the other variants in historical specimens (though DNA quality is an issue in these specimens). There has been limited sequencing of KoRV strains outside of the variable region of *env*; analysis of the koala reference genome indicated that variants D and E were defective in the source animal (Hobbs et al., 2017). This does not mean that these variants are not horizontally transmitted as there are multiple examples of retroviruses in other species (notably cats and chickens) where defective viruses are transmitted alongside replication competent “helper viruses”. KoRV-A might represent remnants of ancestral germ-line infections by exogenous retroviruses with other forms of the virus representing those still active due to continual reinfection or retro-transposition in cis within germ-line cells as reported in other retroviral systems (Boeke and Stoye, 1997; Belshaw et al., 2005). This theory is potentially supported by the phylogenetic pattern evident in the KoRV-A isolates in this study, with the long branch lengths obvious within the KoRV-A cluster (supplementary figure 4) consistent with very closely related endogenous proviruses that have diverged post integration.

KoRV-A is at any rate present in all KoRV positive koalas and is consistent with being endogenous (Xu et al., 2015; Legione et al., 2017; Waugh et al., 2017). This study confirms KoRV-A as being present in all KoRV positive koalas as previously reported by many groups. It also highlights the previously described lower viral load in southern animals (Legione et al., 2017, Simmons et al., 2012). This lower viral load corresponds with a reduced viral diversity and a higher relative percentage abundance of KoRV-A in southern koalas. In addition, in the northern animals, the relative percentage of KoRV-A was much higher in the integrated proviral DNA than in the viral RNA, with other subtypes variable among all samples. This may reflect a relatively greater viral diversity in animals with higher viral loads due to the greater rate of mutation in actively replicating virus. This phenomenon is well described in other retroviruses such as HIV where viral diversity increases with viral replication (Theys et al., 2018). Alternatively, some variants of KoRV-A (particularly endogenous loci) may not be very effectively transcribed, either directly or as a result of competition with high copy number of other transcribed subtypes. Another possibility is that a mutation of the provirus may disrupt DNA sequence elements from the promoter which are essential for transcription. Indeed, the original KoRV-A isolate does not replicate efficiently in cell culture, probably due to sequence changes in its LTR when compared with more replication competent clones (Shimode et al., 2014). It is also possible that transcription from individual KoRV-A loci is uneven with some highly transcribed loci responsible for the RNA detected. These particular loci may be less prevalent in QLD animals. Another potential confounding factor in blood samples is that the levels may not directly reflect viral transcription in other tissues (there is likely differential transcription in different tissues as has been demonstrated for many retroviruses, both endogenous and exogenous). However, replicating virus in any tissue likely produces virions spilling over in the blood.

Several other studies have reported a linkage between detection of KoRV-B provirus and neoplasia or chlamydial disease occurrence (Chaban et al., 2017; Waugh et al., 2017) (Xu et al., 2013). However, the current study does not support an association between the presence of particular virus subtypes in either DNA or RNA forms and the occurrence of disease. Previous studies on the association between KoRV-B subtype and disease were based on conventional PCR. The NGS approach used here is not reliant on sequence specific primers for each KoRV subtype and is therefore able to detect a more comprehensive range of subtypes in individual animals. The NGS approach was also more sensitive in detection of KoRV-B, with only 48.5% of QLD animals and no SA animals testing positive for KoRV-B with conventional PCR, in contrast to 100% of the same animals testing positive with the PCR and NGS approach. There was a significant association between higher read counts of KoRV-B in the NGS data and the likelihood of testing positive for KoRV-B on conventional PCR. These findings indicate that previously reported results for an association of KoRV-B with disease in animals might reflect an association between higher viral load and disease rather than the presence of the KoRV-B subtype per se. In terms of koala population management decisions, the findings also indicate that testing for KoRV-B via endpoint PCR, as has been adopted by some zoological collections, is probably not a useful screen for future neoplasia risk.

Indeed, this study does not provide convincing evidence for an association of any particular virus subtype with a particular disease syndrome, although a trend towards a reduced viral diversity and an increased preponderance of KoRV-A is evident in healthy animals and in the SA animals the oxalate nephrosis animals (this is a genetic disease seen predominantly in the southern population). As with the differences in viral diversity between the populations this probably reflects the previously reported relationship between higher viral load (and therefore sequence diversity), neoplasia and clinical chlamydiosis in KoRV affected animals (Tarlinton et al., 2005). Important caveats where are the small numbers of animals in some disease categories, the differences in disease patterns between the two populations and the numbers of animals with multiple diseases which will have confounded this analysis. These confounding factors also made more appropriate statistical analysis techniques for this type of data (like multivariate modelling) inaccurate.

It still remains unexplained why the South Australian animals have a lower level of KoRV replication and a reduced level of abundance of non-KoRV-A subtypes. This study only looked at *env* diversity and there are other factors that can affect viral replication efficiency. In particular the LTR sequences of retroviruses are known to be major determinants of replication efficiency (Pantginis et al., 1997; Chandhasin et al., 2004) and variations in KoRV-B/J isolate LTRs have been reported previously (Shimode et al., 2014) Hobbs et al 2017) that appear to affect replication efficiency. It is also possible that the South Australian population has defects in the receptor for one or more variants of KoRV (KoRV-A and B are known to use different receptors) affecting the efficiency of viral re-infection in these animals, although preliminary analysis of unpublished transcriptome sequences from the two populations would indicate that this is not the case. Other unpublished data indicate that the SA animals may have a defective form of the virus, which is missing most of the *gag*, *pol* and *env* genes (Tarlinton et al., 2017) and it is possible that this defective virus inhibits replication of the full length virus as has been reported for some other retroviruses (Boeke and Stoye, 1997).

Overall, this study analysed KoRV *env* gene diversity in paired samples of provirus DNA and viral RNA within individual koalas from two different zones of koala habitat representing northern and southern koala populations. The identified sequences significantly enhance the number of *env* gene sequences known for KoRV and highlighted significant variation between the abundance of transcribed variants of KoRV present in the RNA of individuals when compared with the provirus complement in the DNA. This probably reflects differential transcription efficiency of different loci and subtypes. KoRV-A is the likely ancestral version of KoRV, with other variants likely generated via mutations, deletions, or recombination events. These other subtypes have now become the predominant transcribed form of KoRV in the Queensland population. It remains unexplained why the southern animals display such lower viral loads and reduced viral diversity than the northern population, along with such a different disease pattern, however this study highlights that this is not as simple as the presence or absence of particular virus subtypes as has been previously hypothesised.

4. Methods

4.1 Sample collection and preparation

In South-East QLD, animals were sourced from Moggill Koala Hospital, Australia Zoo Wildlife Hospital, RSPCA Wacol and Sea World Paradise Country. South Australian (Mount Lofty Ranges) samples were collected from Fauna Rescue of South Australia (Figure 1, Table 1). Blood (2-3 ml) was collected from live and clinically healthy captive koalas using a sterile 22-gauge butterfly catheter and 5 ml syringe. Wild koalas hospitalised due to disease or serious injury following trauma or animal attack were euthanased and necropsied. Koalas were anaesthetised with 0.25 ml Zoletil (Virbac) intramuscularly. Euthanasia was performed with an intravenous injection of pentobarbitone. Immediately following euthanasia, 10 – 15 ml of blood was withdrawn from a femoral vein or by cardiac puncture into EDTA tubes.

DNA was extracted from 100 µl EDTA whole blood using Qiagen DNeasy Blood & Tissue Kit according to manufacturer's (Qiagen) instructions. A 1-2 ml aliquot of blood was centrifuged at 3000 g for 5 mins and 200 µl of plasma was removed and added to 300 µl of RNAlater stabilisation agent (Qiagen) within 15 min of blood collection. RNA was extracted using Qiagen QIAmp Viral RNA mini kit with on-column Qiagen RNase free DNase steps. Briefly, 140 µl of RNAlater diluted plasma was suspended in 560 µl viral lysis buffer containing carrier RNA and extraction continued following the extraction kit procedures and finally eluted in 30 µl water. The extracted RNA samples were stored at -80°C until required.

4.2 KoRV-A and KoRV-B real-time and conventional PCR

To test for the presence of KoRV-A and B, we used real time and conventional PCR. Initially, KoRV positivity of the extracted DNA and RNA was initially assessed with a real-time PCR of the KoRV *pol* gene using published primers and probe (Tarlinton et al., 2005). DNA and RNA samples were amplified using TaqMan gene expression master mix (Applied Biosystem) and SuperScript® III One-Step RT-PCR System with Platinum® Taq DNA Polymerase (Invitrogen) respectively, following manufacturers' instructions, in a BioRad CFX 96. Samples were considered KoRV-positive if the CT value was $\leq$35.

Conventional PCR of the *env* gene was performed using published primers to specifically amplify each of the KoRV-A and KoRV-B env subtypes (Waugh et al., 2017) as a preliminary assessment of KoRV subtype prevalence. Primers used in this study shown in Table 2. The Qiagen HotStartTaq Plus Master Mix kit was used for PCR of DNA samples following the manufacturer's instructions with 35 cycles of amplification and an annealing temperature of 51°C. KoRV-A and KoRV-B positive samples were directly purified with ExoSAP-IT (Thermo Fisher Scientific), following manufacturer's directions and Sanger sequenced to validate the amplification of this subtype. Sequencing was undertaken using Big Dye Terminators (ThermoFisher Scientific) at the Animal Genetics Laboratory, University of Queensland. Sequences were subjected to BLAST analysis through the NCBI database to determine the percentage of homology to known subtypes.

Table 2: Primers used in this study for PCR

Region	**Forward**	**Reverse**	**Reference**
Pol	TTGGAGGAGGAATACCGATTACAC	GCCAGTCCCATACCTGCCTT	(Tarlinton et al., 2005)
Env KoRV-A specific	TCCTGGGAACTGGAAAAGAC	GGGTTCCCCAAGTGATCTG	(Waugh et al., 2017)
Env KoRV-B specific	TCCTGGGAACTGGAAAAGAC	GGCGCAGACTGTTGAGATTC	(Waugh et al., 2017)

4.3 Sample preparation for Illumina sequencing

Previously published oligonucleotide primers flanking the hypervariable region of the *env* gene (Chappell et al., 2017) were used to amplify a 500 bp fragment of target sequence by PCR. The primers contained the Illumina adaptor sequences (italics) ligated to *env* gene complementary regions.

The Qiagen HotStartTaq Plus Master Mix kit was used to amplify from DNA and the Qiagen OneStep RT-PCR kit was used to amplify from RNA, both following the manufacturer's instructions with an annealing temperature of 58°C and 35 rounds of amplification. We adopted recently established deep sequencing methodology for analysis of the *env* gene hypervariable region, such that consistency was retained between current and previous findings (Chappell et al., 2017). Samples were prepared following the Illumina 16S Metagenomic Sequencing Library Preparation guidelines. The purification and sequencing of PCR amplicons was performed at the Ramaciotti Centre for Genomics (University of New South Wales, Sydney, Australia). Purification was performed using Agencourt AMPure XP beads (Beckman Coulter, USA) and purified DNA was indexed with unique 8 bp barcodes using the Illumina Nextera XT 384 sample Index Kit A-D (Illumina FC-131-1002) following standard PCR conditions. Indexed amplicons were pooled together in equimolar concentrations and sequenced on the MiSeq Sequencing System (Illumina, USA) using paired end sequencing with V3 300bp following manufacturer's protocols.

4.4 Sequence assembly

The overlapped forward and reverse reads of the Illumina next generation sequencing (NGS) were assembled using the OL assembly method in the *dDocent* pipeline (Puritz et al., 2014). The reads were trimmed using Trimmomatic (Bolger et al., 2014) and assembled using *Rainbow* (Chong et al., 2012) and *CD-HIT* (Fu et al., 2012). A series of optimization assemblies were run to assess how the number of contigs would be affected by parameter choice in *dDocent* and the effect of clustering threshold: (a) the minimum number of samples in which a sequence had to be represented (1−10 samples); and (b) the clustering threshold (80−98%). These preliminary optimization runs indicated that the level of clustering did not have a substantial effect on the total number of unique contigs assembled. What did have a major impact was the number of individuals required to represent a sequence: when this was 1, the number of contigs that could be assembled from a single individual was substantially larger then when assemblies required a sequence to be found in ≥2 samples. This result implicates considerable subtype sequence variation, but it is hard to resolve this from technical or sequencing error that may generate false variation. Therefore, the final assembly included the parameter selection of: (i) sequences present in at least two samples, (ii) a minimum read count of four and (iii) a 99% clustering threshold. Graphical view of optimisation assemblies are shown in supplementary file 1.

The representative sequences were aligned with KoRV-A (AF151794) and KoRV-B (KC779547.1) using the *ClustalW* alignment in the program *BioEdit* (Hall, 1999) to identify the presence of any anomalous contigs. Sequences that failed to show homology against reference sequences were removed from further analysis. The putative sequences were mapped by *BWA* (Li and Durbin, 2009) with the following parameters: match score = 1, mismatch penalty = 4, gap open penalty = 15. Finally, *SAMTools*

(Li et al., 2009) was used to filter the alignment bam files (for a MapQ score of 30) and to extract the counts of reads mapped to each contig.

4.5 Phylogenetic analysis

The representative unique sequences were imported into the Geneious v11.0.4 software package (https://www.geneious.com/) and combined with previously published KoRV env sequences; KoRV-A (AF151794, KX587957.1 and KP792565.1), KoRV-B (KX588002.1, KX588011.1, KX588027.1, KC779547.1, AB822553.1, KX588031.1, KX588053.1), KoRV-C (AB828005.1, KP792564.1), KoRV-D (KX587952.1, KX587991.1, KX588043.1, KX587993.1, KX587972.1, KX587972.1, AB828004.1, KX587997.1), KoRV-E (KU533853.1), KoRV-F (KX588025.1, KX588028.1, KX587994.1, KU533852.1), KoRV-G (KX587961.1 and KX587998.1), KoRV-H (KX588036.1 and KX587979.1), and KoRV-I (KX587976.1 and KX588021.1) were used. Moreover, four other sequences from the viruses previously determined to be the most closely related to KoRV (Simmons et al., 2014) in other were used as outgroups to root the KoRV phylogeny: two Gibbon ape leukaemia virus (GALV) sequences, (KT724047.1, KT724048.1) and two *Melomys burtoni* retrovirus (MbRV) sequences (KF572486.1, KF572485.1). Sequences were aligned using *ClustalW* alignment with a gap opening cost of 15 and a gap extension cost of 7 as implemented in Geneious 11.0.4. The alignments were further edited by hand to fill the blanks at the beginning and end.

A Bayesian phylogenetic tree was determined from the aligned reads using the Geneious plugin of MrBayes 3.2.6 (Huelsenbeck and Ronquist, 2001) with a chain length of 10,100,000, a subsampling frequency of 2000 and a burn-in-length of 1,100,100, with all others parameters set at defaults. Sequences were manually allocated to a KoRV subtype based on clustering with previously identified reference subtypes and phylogenetic topology.

4.6 Statistical analysis

The comparison of the number of unique proviral sequences and read counts between QLD and SA populations was statistically evaluated through non-parametric Mann Whitney U test. The concordance of the NGS results with the conventional PCR testing for KoRV-B was assessed through Pearson's chi-squared tests. The taxonomic count data was analysed for statistically significant differences, between QLD vs SA provirus, and DNA vs RNA for both SA and QLD samples, in R (Ihaka and Gentleman, 1996) using the EdgeR wrapper (Robinson et al., 2010) as part of the phyloseq package (McMurdie and Holmes, 2013). Diversity statistics were calculated using vegan (Dixon, 2003) and differences were assessed for significance using Mann-Whitney U tests in Prism 8.01 (GraphPad Software Inc. USA).

FUNDING INFORMATION

This project and scholarship for NS were funded by the Queensland Department of the Environment and Heritage Koala Research Grant Programme 2012. NS was also supported by a Keith Mackie Lucas Travel Scholarship from the University of Queensland.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

ETHICAL STATEMENTS

Ethical approval for this study was granted by the University of Queensland Animal Ethics Committee, permit number ANFRA/SVS/461/12 and ANRFA/SVS/445/15, the Queensland Government Department of Environment and Heritage Protection permit number WISP11989112, University of Adelaide Animal Ethics Committee permit number S-2013-198 and South Australian Government Department of Environment, Water and Natural Resources Scientific Research Permit Y26054.

AUTHOR CONTRIBUTIONS

N.S. performed DNA and RNA extraction, laboratory experiments, data analysis and drafted manuscript. J. Meers, J.M.S., G.S. and H.O. helped in laboratory experiment set up, data interpretation and manuscript preparation. J. T. helped in bioinformatics analysis. R.D.E and R.T edited the manuscript. J.F. and N. Speight helped in sample collection and reviewing manuscript. J.K. and A.B.M reviewed the statistical analysis and edited the manuscript. F.H, D.T. and L.W. reviewed the manuscript. All authors read and approved the final manuscript.

REFERENCES

Anderson, M.M., Lauring, A.S., Burns, C.C., Overbaugh, J., 2000. Identification of a cellular cofactor required for infection by feline leukemia virus. Science 287, 1828-1830.

Belshaw, R., Katzourakis, A., Pačes, J., Burt, A., Tristem, M., 2005. High Copy Number in Human Endogenous Retrovirus Families is Associated with Copying Mechanisms in Addition to Reinfection. Molecular Biology and Evolution 22, 814-817.

Boeke, J.D., Stoye, J.P. 1997. Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements, In: Coffin, J.M., Hughes, S.H., Varmus, H.E. (Eds.) Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Bolin, L.L., Levy, L.S., 2011. Viral determinants of FeLV infection and pathogenesis: lessons learned from analysis of a natural cohort. Viruses 3, 1681-1698.

Burnard, D., Gillett, A., Polkinghorne, A., 2018. Chlamydia pecorum in Joint Tissue and Synovial Fluid of a Koala (Phascolarctos cinereus) with Arthritis. Journal of wildlife diseases.

Canfield, P.J., Sabine, J.M., Love, D.N., 1988. Virus particles associated with leukaemia in a koala. Australian veterinary journal 65, 327-328.

Chaban, B., Ong, V.A., Hanger, J., Timms, P., 2017. Molecular dynamics and mode of transmission of Koala Retrovirus (KoRV) as it invades and spreads through a wild Queensland koala population. Journal of virology.

Chandhasin, C., Coan, P.N., Levy, L.S., 2005a. Subtle mutational changes in the SU protein of a natural feline leukemia virus subgroup A isolate alter disease spectrum. J Virol 79, 1351-1360.

Chandhasin, C., Coan, P.N., Pandrea, I., Grant, C.K., Lobelle-Rich, P.A., Puetter, A., Levy, L.S., 2005b. Unique long terminal repeat and surface glycoprotein gene sequences of feline leukemia virus as determinants of disease outcome. J Virol 79, 5278-5287.

Chandhasin, C., Lobelle-Rich, P., Levy, L.S., 2004. Feline leukaemia virus LTR variation and disease association in a geographical and temporal cluster. J Gen Virol 85, 2937-2942.

Chappell, K.J., Brealey, J.C., Amarilla, A.A., Watterson, D., Hulse, L., Palmieri, C., Johnston, S.D., Holmes, E.C., Meers, J., Young, P.R., 2017. Phylogenetic Diversity of Koala Retrovirus within a Wild Koala Population. Journal of virology 91.

Chong, Z., Ruan, J., Wu, C.I., 2012. Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. Bioinformatics 28, 2732-2737.

Dixon, P., 2003. VEGAN, A Package of R Functions for Community Ecology. Journal of Vegetation Science 14, 927-930.

Fabijan, J., Woolford, L., Lathe, S., Simmons, G., Hemmatzadeh, F., Trott, D.J., Speight, N., 2017. Lymphoma, Koala Retrovirus Infection and Reproductive Chlamydiosis in a Koala (Phascolarctos cinereus). Journal of Comparative Pathology 157, 188-192.

Fiebig, U., Hartmann, M.G., Bannert, N., Kurth, R., Denner, J., 2006. Transspecies transmission of the endogenous koala retrovirus. Journal of virology 80, 5651-5654.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152.

Gonzalez-Astudillo, V., Allavena, R., McKinnon, A., Larkin, R., Henning, J., 2017. Decline causes of Koalas in South East Queensland, Australia: a 17-year retrospective study of mortality and morbidity. Sci Rep 7, 42587.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41, 95-98.

Hanger, J.J., Bromham, L.D., McKee, J.J., O'Brien, T.M., Robinson, W.F., 2000. The nucleotide sequence of koala (Phascolarctos cinereus) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus. Journal of virology 74, 4264-4272.

Hobbs, M., King, A., Salinas, R., Chen, Z., Tsangaras, K., Greenwood, A.D., Johnson, R.N., Belov, K., Wilkins, M.R., Timms, P., 2017. Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. Sci Rep 7, 15838.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755.

Ihaka, R., Gentleman, R., 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5, 299-314.

Ishida, Y., Zhao, K., Greenwood, A.D., Roca, A.L., 2015. Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. Molecular biology and evolution 32, 109-120.

Legione, A.R., Patterson, J.L., Whiteley, P., Firestone, S.M., Curnick, M., Bodley, K., Lynch, M., Gilkerson, J.R., Sansom, F.M., Devlin, J.M., 2017. Koala retrovirus genotyping analyses reveal a low prevalence of KoRV-A in Victorian koalas and an association with clinical disease. J Med Microbiol 66, 236-244.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

McMurdie, P.J., Holmes, S., 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217.

Miyazawa, T., Shojima, T., Yoshikawa, R., Ohata, T., 2011. Isolation of koala retroviruses from koalas in Japan. The Journal of veterinary medical science / the Japanese Society of Veterinary Science 73, 65-70.

Nyari, S., Waugh, C.A., Dong, J., Quigley, B.L., Hanger, J., Loader, J., Polkinghorne, A., Timms, P., 2017. Epidemiology of chlamydial infection and disease in a free-ranging koala (Phascolarctos cinereus) population. PloS one 12, e0190114.

Overbaugh, J., Donahue, P.R., Quackenbush, S.L., Hoover, E.A., Mullins, J.I., 1988. Molecular cloning of a feline leukemia virus that induces fatal immunodeficiency disease in cats. Science 239, 906-910.

Pantginis, J., Beaty, R.M., Levy, L.S., Lenz, J., 1997. The feline leukemia virus long terminal repeat contains a potent genetic determinant of T-cell lymphomagenicity. J Virol 71, 9786-9791.

Puritz, J.B., Hollenbeck, C.M., Gold, J.R., 2014. dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2, e431.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 26, 139-140.

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

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

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

Simmons, G.S., Young, P.R., Hanger, J.J., Jones, K., Clarke, D., McKee, J.J., Meers, J., 2012. Prevalence of koala retrovirus in geographically diverse populations in Australia. Australian veterinary journal 90, 404-409.

Tarlinton, R., Meers, J., Hanger, J., Young, P., 2005. Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. The Journal of general virology 86, 783-787.

Tarlinton, R., Meers, J., Young, P., 2008. Biology and evolution of the endogenous koala retrovirus. Cellular and molecular life sciences : CMLS 65, 3413-3421.

Tarlinton, R.E., Sarker, N., Fabijan, J., Dottorini, T., Woolford, L., Meers, J., Simmons, G., Owen, H., Seddon, J., Hemmatzedah, F., Trott, D., Speight, N., Emes, R., 2017. Differential and defective expression of Koala Retrovirus reveal complexity of host and virus evolution. bioRxiv.

Theys, K., Libin, P., Pineda-Peña, A.-C., Nowé, A., Vandamme, A.-M., Abecasis, A.B., 2018. The impact of HIV-1 within-host evolution on transmission dynamics. Current Opinion in Virology 28, 92-101.

Waugh, C.A., Hanger, J., Loader, J., King, A., Hobbs, M., Johnson, R., Timms, P., 2017. Infection with koala retrovirus subgroup B (KoRV-B), but not KoRV-A, is associated with chlamydial disease in free-ranging koalas (Phascolarctos cinereus). Sci Rep 7, 134.

Xu, W., Gorman, K., Santiago, J.C., Kluska, K., Eiden, M.V., 2015. Genetic diversity of koala retroviral envelopes. Viruses 7, 1258-1270.

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

Figures:

Fig. 1: Location of sample collection site. Red dot showing the sample collection cities. From Queensland, samples were collected from Gold Coast, Brisbane, and Sunshine Coast and from South Australia, koalas were collected from Mount Lofty region. Map was adapted from Australian Koala Foundation (AKF) website.

Fig. 2: Phylogenetic tree from aligned KoRV sequences (including 169 newly identified in this study and 24 previously published sequences with two sequences from GALV and MbRV (used as outgroups to root the tree) generated through Geneious implemented Bayesian approach. Previously published KoRV env sequences; KoRV-A (AF151794, KX587957.1 and KP792565.1), KoRV-B (KX588002.1, KX588011.1, KX588027.1, KC779547.1, AB822553.1, KX588031.1, KX588053.1), KoRV-C (AB828005.1, KP792564.1), KoRV-D (KX587952.1, KX587991.1, KX588043.1, KX587993.1, KX587972.1, KX587972.1, AB828004.1, KX587997.1), KoRV-E (KU533853.1), KoRV-F (KX588025.1, KX588028.1, KX587994.1, KU533852.1), KoRV-G (KX587961.1 and KX587998.1), KoRV-H (KX588036.1 and KX587979.1), and KoRV-I (KX587976.1 and KX588021.1) were used. Outgroup sequences: Gibbon ape leukaemia virus (GALV) sequences, (KT724047.1, KT724048.1) and *Melomys burton*i retrovirus (MbRV) sequences (KF572486.1, KF572485.1). Bayesian value are shown. Paraphyletic KoRV-D has multiple subclades numbered D1 to D11 and also includes previously designated KoRV-G and KoRV-H. Clades color and weight are marked as gradient following posterior probabilities values

Fig. 3: Comparison of the read counts of KoRV env subtypes A, B, I and D (including subclades) in the proviral DNA form within QLD (n = 33) and SA (n = 28) koala populations. Mean read counts with one standard deviation error bars are shown. Although all SA animals had KoRV-B and D10, their lower level read counts are not observable at this scale.

Fig. 4: Genetic diversity of KoRV *env* subtypes among paired DNA and RNA samples was illustrated through the relative percentage of total reads of (A) QLD and (B) SA koala populations. Colors indicate

the different subtypes. (A) Among 33 QLD koalas, 28 were present in both DNA and RNA forms and (B) among 28 SA koalas, 5 had both DNA and RNA forms.

Fig. 5: Percentage relative abundance of viral subtypes A (green), B (orange), I (red) and combined D (purple). Compared between QLD and SA animals for A) DNA, B) RNA, C) RNA and DNA for paired QLD samples, and D) RNA and DNA for paired SA samples. Median and interquartile ranges displayed.

Supplementary Files:

File 1 (figure): Preliminary optimisation assemblies using *dDocent*. Two parameters were explored: (1) the minimum number of samples required to represent a sequence in the assembly (colours, see legend), and (2) the clustering thresholds used to group reads into similar sequences. The clear effect of singletons (only requiring a sequence to be represented in one individual) on the number of contigs probably arises from two competing (non-mutually exclusive) hypotheses: firstly, large variation exists within and between individuals; and secondly, technical and sequencing error introduces sequence variation. This problem appears to disappear when the number of samples was ≥2.

File 2: Details of koala samples used for analysis of *env* gene diversity using 16S Metagenomics sequencing.

File 3 (figure): Alignment of unique Koala Retrovirus (KoRV) *env* sequences from koalas sampled in Queensland and South Australia.

File 4 (figure): Simple view of Bayesian phylogenetic tree. Unique sequences were aligned with previously published KoRV env sequences; KoRV-A (AF151794, KX587957.1 and KP792565.1), KoRV-B (KX588002.1, KX588011.1, KX588027.1, KC779547.1, AB822553.1, KX588031.1, KX588053.1), KoRV-C (AB828005.1, KP792564.1), KoRV-D (KX587952.1, KX587991.1, KX588043.1, KX587993.1, KX587972.1, KX587972.1, AB828004.1, KX587997.1), KoRV-E (KU533853.1), KoRV-F (KX588025.1, KX588028.1, KX587994.1, KU533852.1), KoRV-G (KX587961.1 and KX587998.1), KoRV-H (KX588036.1 and KX587979.1), and KoRV-I (KX587976.1 and KX588021.1) using ClustalW alignment programme. Moreover, four other sequences were used as outgroups to root the KoRV phylogeny: two Gibbon ape leukaemia virus (GALV) sequences, (KT724047.1, KT724048.1) and two *Melomys burtoni* retrovirus (MbRV) sequences (KF572486.1, KF572485.1). Alignments were further edited by hand to fill the blanks at the beginning and end. Phylogenetic tree was determined from the aligned reads using the Geneious plugin of MrBayes 3.2.6.

File 5: The read count of unique KoRV *env* sequences from PCR and next-generation sequencing of individual koala DNA and RNA samples.

File 6: The relative percentage of each subtype of KoRV *env* gene in individual DNA and RNA samples of koalas.

File 7 (figure) 7: Pecentage abundance of each major subtype of KoRV (A, B, Combined D, I) in the DNA of individuals compared with disease status (Healthy, Neoplasia, Oxalate Nephrosis, Chlamydiosis). Median and interquartile range displayed.