**Supplementary Methods**

**Study sites and sample collection**

Koala faecal samples were collected between April and September 2016, from 20 locations (Fig. 1) that span the koala’s current wild geographic range. The habitats at the sites varied between cleared farmland with semi-isolated paddock trees, open woodland and tall eucalypt forest. In general, faecal samples were collected from 8-11 individuals per site with the exception of Port Macquarie (NSW) where three animals were sampled and Cape Otway (Victoria) where 18 koalas were sampled. Fresh faecal pellets were collected from each koala by placing a plastic mat (2 m by 3 m) directly beneath them. Faecal pellets were collected from the mat and immediately frozen at -18 °C. All samples were collected less than 24 hours post-production and 98 % were collected less than 10 hours after production. Latitude and longitude were recorded at the location of each sampled koala using a Garmin eTrex 20x handheld GPS. These coordinates were used to calculate the geographic centre of each study site.

To determine how KoRV subtype profiles recovered from faecal DNA compared to those obtained from plasma RNA, paired blood and faecal samples were collected from eight captive koalas housed at Currumbin Wildlife Sanctuary, Dreamworld on the Gold Coast and Paradise Country Wildlife Park. Up to 3 mls of blood was drawn by a qualified veterinarian from the cephalic vein of each koala during conscious restraint. The blood was then transferred to an EDTA coated tube to prevent clotting. Faecal samples were collected concurrently directly from the koalas’ cloacas into sterile 15 ml tubes. The samples were stored on ice for approximately 3 to 5 hours during transport to the laboratory. The blood was then centrifuged at 16 000 g for 10 minutes and the plasma collected by pipetting. The plasma and faeces were then stored at -80 °C.

The fieldwork was carried out under Western Sydney University ethic approval (A11253) and with appropriate permits from the New South Wales (SL101722), Queensland (WITK17277716), Victorian (10007714) and South Australian State governments (U26533-1). Collection of samples from captive koalas was carried out under University of Queensland ethic approval (AE36153).

**Molecular techniques**

*Nucleic acid Extraction*

Total DNA was extracted from surface washes of the faecal samples following the general approach of Wedrowicz et al. (1). A single faecal pellet from each koala was placed in 3 mls of 1% phosphate buffered saline and gently rotated for 15 minutes on a rotatory mixer. The pellet was then removed and discarded. The phosphate buffered saline wash was centrifuged for 15 minutes at 4 000 g and the supernatant removed. DNA was extracted from the wash pellet by Digsol/proteinase K digestion following the method of Bruford et al. (2) and eluted in 50 ul of 1x Tris-EDTA. PCR inhibitors were then removed from the extracted DNA using the OneStep PCR inhibitor removal kit (Zymo) following the manufacturer’s instructions.

Prior to RNA extraction the plasma samples were passed through a 0.45 µm membrane filter to remove any remaining koala cells. Total RNA was extracted from the plasma samples using the High Pure Viral Nucleic Acid Kit (roche) according to the manufacturer’s instructions with modification as outlined in the supplementary methods. First strand cDNA was then created from
the RNA extracts using Superscript III (Invitrogen) and 32 µM random pentadecemer primers (3) following the manufacturer’s instructions. Following synthesis, the cDNA was treated with 1 µl of RNase H (New England Biolabs) and incubated at 37°C of 20 minutes.

Quantitative PCR

The presence of the KoRV pol gene in a faecal sample was inferred from the successful amplification of a 110 bp fragment by quantitative PCR (qPCR) run in triplicate using previously designed primers (4). The presence of the KoRV env gene was inferred from the amplification of a 97 bp (in the case of KoRV-A) fragment using degenerate primers designed from the alignment of KoRV-A and KoRV-D env sequences (5) (KoRV-RBD(A,D).F: CTCACTGCAWCGGCCTCYCAACAGGC; KoRV-RBD(A,D).R: GGGATAGCTACATCCCAGGGTTYC). A 123 bp fragment of the koala β-actin gene was also amplified in triplicate as a positive control (4). Each fragment was amplified in a 10 µl reaction containing 1X SYBR™ Green PCR Master Mix, each primer at 500 nM and 2 µl of template.

Tenfold dilutions of the DNA were first used as template, with the amplification repeated with 100-fold dilutions for those samples that showed evidence of inhibition or that failed to amplify for β-actin.

Amplification was performed using a QuantStudio 6 Flex Real-Time PCR system using the standard comparative CT protocol with an annealing temperature of 60°C. Quality control and standard curves were then completed as per the supplementary methods.

env hypervariable domain sequencing

An approximately 500 bp region of the KoRV env gene containing the previously identified hypervariable domain was amplified and sequenced from all KoRV positive samples (identified by qPCR). Of these samples, the env hypervariable region could be amplified in 64 using the protocol of Chappell et al. (5), with the number of PCR cycles varied between 28 and 40 depending on KoRV DNA quantity. Adequate amplification could only be obtained for the remaining 57 faecal wash samples and the plasma cDNA using the nested polymerase chain reaction (PCR) protocol described in the supplementary methods. For the faecal DNA, either 10 (n =91) or 100 (n = 30) fold dilutions were used as PCR template to overcome the presence of any remaining PCR inhibitors in the extractions.

Sequencing libraries were prepared from the env PCR products using the Nextera XT sample preparation kit (Illumina) following the manufacturer’s protocol. Each library was uniquely tagged using the Nextera XT indexing primers (Illumina). Paired-end sequencing was performed at the Australian Centre of Ecogenomics, on the Illumina Nextseq using the version 2 reagent kit for 300 cycles.

Bioinformatic processing and statistical analysis

KoRV incidence and copy number per cell

KoRV incidence was inferred from the proportion of β-actin positive samples that were also positive for KoRV pol or env as determined by qPCR. The number of KoRV gene copies per koala cell was estimated for each sample from the number of KoRV pol or env and β-actin molecules per qPCR reaction. The number of KoRV pol, env and β-actin molecules present in each reaction were calculated from the average CT values and standard curves using the appropriate standard equations. To estimate the number of β-actin copies per koala cell, the number of fragments
produced from the koala genome (NCBI assembly: GCA_900166895.1) with the qPCR primers was determined in CLC genomics workbench 20. This analysis produced 7 fragments or 14 copies per diploid cell. Therefore, the KoRV gene copies per cell were estimated as:

\[ \text{KoRV gene copies per reaction} = \frac{\text{KoRV gene copies per reaction}}{\text{beta-actin copies per reaction/14}} \]

**Geographic distribution of KoRV subtypes**

The geographic distribution of KoRV subtypes was determined from the env Illumina deep sequencing data. The raw reads were merged, quality filtered and rarefied to 10,000 reads per sample in CLC genomics workbench 20 as described in the supplementary methods. The rarefied reads were then de novo clustered across all samples at 97% similarity in QIIME 2 (6). Clusters containing only a single read across the dataset were discarded. To determine if the representative KoRV sequences were putatively functional, they were translated in silico with those containing missense and frameshift mutations or large deletions designated as non-functional as described in the supplementary methods. Subtype designations for the intact representative KoRV sequences were assigned by comparison to a set of reference subtype sequences in a three-step process as described in the supplementary methods. The geographic distribution of the subtypes was then inferred from the representative sequence by sample table generated from QIIME 2 clustering.

**env sequence richness, genetic diversity and differentiation**

To assess how the genetic diversity of KoRV varied within and among populations over the koala’s geographic range both the intact and non-functional representative KoRV env sequences were analysed to capture the complete KoRV “genetic pool” within which recombination is presumed to be prolific. Sequence incidence and within koala abundances were calculated from the representative sequence by sample table generated from the QIIME 2 clustering. Network diagrams of sequence sharing were constructed in NetDraw. Rarefaction curves and 95% confidence intervals of the number of unique sequences detected with increasing sampling effort were constructed using the package ‘rich’ (7) in R (8).

When calculating the sequence similarities, the hypervariable portion of the env nucleotide sequences (corresponding to aa81-143 of KoRV A numbering) was excluded as it may be of non-KoRV origin in some subtypes (9). The pairwise number of nucleotide differences between sequences were calculated in GenAlEx 6.5 (10, 11). Analysis of Molecular Variance (AMOVA) and pairwise PhiPT values were also calculated using GenAlEx 6.5 (10, 11) with significance determined by 9999 permutations of the dataset. The AMOVA was carried out using three different underlying versions of the data. In the first analysis, each sequence detected in a population was included once in the distance matrix, with sequences found in multiple populations included once for each population. This analysis tested sequence differentiation among populations. In the second analysis, the incidence of each sequence within a population was accounted for by including each sequence found within a koala once (e.g., if a sequence was detected in 5 koalas in a population it was included in the distance matrix 5 times for that population). This analysis also measured variation between koalas within populations in the sequences they carried. In the third analysis, the values in the distance matrix from the second analysis were weighted by the within koala relative abundance of the sequences (multiplied by 1000). Thus, this analysis accounted for within koala relative abundance and incidence. However, the weighting of the distance matrix reduces the validity of the permutation test of significance and thus significance values associated with this analysis should be viewed with caution.

The geographic distance between populations was calculated from the decimal latitude and longitude values for the centre of each site. Mantle tests were then performed in GenAlEx to test for
associations between geographic and genetic distances. Generalised and linear regression models were fitted with the stats package in R using appropriate link functions. Regression models that included random effects were fitted using the lme4 package (12) in R with significance determined using the lmerTest package (13).

REFERENCES

1. F. Wedrowicz, M. Karsa, J. Mosse, F. E. Hogan, Reliable genotyping of the koala (Phascolarctos cinereus) using DNA isolated from a single faecal pellet. *Molecular ecology resources* **13**, 634-641 (2013).
2. M. W. Bruford, O. Hanotte, J. K. Y. Brookfield, T. Burke, "Multilocus and single-locus DNA fingerprinting" in Molecular Genetic Analysis of Populations: A Practical Approach, A. R. Hoelzel, Ed. (IRL Press, Oxford, 1998), pp. 287-336.
3. M. Stangegaard, I. Høgh Dufva, M. J. B. Dufva, Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. **40**, 649-657 (2006).
4. M. E. H. Kayesh et al., Molecular dynamics of koala retrovirus infection in captive koalas in Japan. *Archives of virology* **164**, 757-765 (2019).
5. K. Chappell et al., Phylogenetic diversity of koala retrovirus within a wild koala population. *Journal of virology* **91**, e01820-01816 (2017).
6. E. Bolyen et al. (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. (PeerJ Preprints).
7. J.-P. J. D. Rossi, rich: an R package to analyse species richness. **3**, 112-120 (2011).
8. R-Core-Team (2012) R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria).
9. Y. Anai et al., Infectious endogenous retroviruses in cats and emergence of recombinant viruses. **86**, 8634-8644 (2012).
10. R. Peakall, P. E. Smouse, GenAIEx 6: genetic analysis in Exel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295 (2006).
11. R. Peakall, P. E. Smouse, GenAIEx 6.5: genetic analysis in Exel. Population genetic software for teaching and research - an update. *Bioinformatics* **28**, 2537–2539 (2012).
12. D. Bates, M. Maechler, B. Bolker, lme4: Linear mixed-effects models using S4 classes. (2012).
13. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen (2013) lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).
Fig S1: First three dimensions of the principal components analysis of the intact KoRV *env* representative sequences. Subtype B and A/B are not included as they clustered separately to all the other subtypes. The variation explained by each axis is shown in brackets. Colours represent the final designation of the sequences after manual inspection of the translated amino acid sequences. As such, it should be noted that not all designations match the principal components clusters (e.g., D/F intermediates). Cyan = A, dark grey = C, red = D, blue = D divergent, cream = A/D intermediates, orange = F, light green = D/F intermediates, purple pink = I, yellow = K, light pink = L, light grey = M, grey green = other previously identified subtype reference sequences.
Fig S2: *In silico* translated amino acid sequences for KoRV-L *env* hypervariable region aligned against the reference sequences for the previously identified subtypes.

Fig S3: *In silico* translated amino acid sequences for KoRV-M *env* hypervariable region aligned against the reference sequences for the previously identified subtypes.

Fig S4: *In silico* translated amino acid sequences for KoRV *env* hypervariable region A/B intermediates aligned against the A and B reference sequences. Black vertical lines among amino acids indicate approximate breakpoints between A and B motifs.
Fig S5. Pairwise PhiPT for northern populations against the Euclidian geographic distance between those populations.

Fig S6: The within koala sequence abundance in relation to sequence incidence (a) within and (b) among populations.
Figure S7: Rank order of KoRV-A insertions in the koala reference genome. The graph shows each annotated KoRV-A insertion, ordered by increasing sequence divergence. Sequences 31-54 showing divergence >12% are all short LTR elements.

Fig S8: First two principal components of the pairwise Fst values between the twenty study populations based on 13 microsatellite markers (AF31, AF34, Pcin07, Pcin08, Pcin10, Pcin14, Pcin15, Pcin20, Pcin21, Pcin23, Phci19, Phci27, Phci5). Proportion of variation explained by each component shown in the axis labels.
### Supplementary Tables

#### Table S1: Pairwise PhiPT (below diagonal) and significance (above diagonal) for northern populations based on the distances among the unique sequences detected in each population.  

| Population            | Sample Size | Na   | Ne  | F   | Ho   | He  | uHe  | F   |
|-----------------------|-------------|------|-----|-----|------|-----|------|-----|
| Mikkira               | 10          | 2.46 | 1.87| 0.66| 0.44 | 0.41| 0.43 | -0.08|
| Kangaroo Is.          | 10          | 3.08 | 2.03| 0.79| 0.49 | 0.45| 0.47 | -0.09|
| Adelaide Hills        | 10          | 2.46 | 1.90| 0.66| 0.46 | 0.40| 0.42 | -0.18|
| Bessiebelle           | 9           | 3.31 | 2.35| 0.91| 0.56 | 0.51| 0.54 | -0.08|
| Cape Otway            | 18          | 3.62 | 2.28| 0.85| 0.50 | 0.47| 0.48 | -0.08|
| Strzelecki            | 10          | 3.15 | 2.22| 0.80| 0.44 | 0.44| 0.44 | 0.00 |
| Strz blue             | 10          | 3.46 | 2.53| 0.91| 0.51 | 0.49| 0.52 | 0.02 |
| Boho Sth              | 10          | 3.15 | 2.15| 0.79| 0.45 | 0.43| 0.46 | -0.07|
| Ulupna                | 10          | 3.15 | 2.37| 0.82| 0.48 | 0.46| 0.48 | -0.04|
| Monaro                | 8           | 4.77 | 3.23| 1.27| 0.70 | 0.65| 0.69 | -0.11|
| Mountain Lagoon       | 8           | 4.38 | 2.77| 1.14| 0.65 | 0.60| 0.64 | -0.10|
| Port Macquarie        | 3           | 3.00 | 2.54| 0.96| 0.59 | 0.57| 0.68 | -0.01|
| Gunnedah              | 10          | 4.00 | 2.58| 1.03| 0.55 | 0.55| 0.57 | -0.01|
| Nowendoc              | 10          | 4.77 | 3.03| 1.19| 0.69 | 0.60| 0.63 | -0.16|
| Mt. Byron             | 10          | 5.46 | 3.69| 1.38| 0.72 | 0.67| 0.70 | -0.09|
| SW QLD                | 10          | 5.38 | 3.58| 1.36| 0.66 | 0.67| 0.71 | 0.00 |
| Nth Stradbroke Is.    | 11          | 3.38 | 2.12| 0.86| 0.48 | 0.49| 0.51 | 0.03 |
| Clermont              | 10          | 4.46 | 3.32| 1.26| 0.65 | 0.66| 0.69 | 0.00 |
| St. Bee’s Is.         | 9           | 4.62 | 2.86| 1.16| 0.56 | 0.60| 0.63 | 0.09 |
| Magnetic Is.          | 10          | 4.23 | 2.85| 1.10| 0.62 | 0.57| 0.60 | -0.08|

1. Na = No. of Different Alleles; 2. Ne = No. of Effective Alleles = 1 / (∑pi²); 3. I = Shannon’s Information Index = -1 x Sum (pi x Ln (pi)); 4. Ho = Observed Heterozygosity = No. of Heterozygotes /N; 5. He = Expected Heterozygosity = 1 - ∑pi²; 6. uHe = Unbiased Expected Heterozygosity = (2N / (2N-1)) * He; 7. F = Fixation Index = (He - Ho) / He = 1 - (Ho / He); Where pi is the frequency of the ith allele for the population and N = sample size.

#### Table S2: Measures of koala genetic diversity for the twenty study populations based on 13 microsatellite markers (AF31, AF34, Pcin07, Pcin08, Pcin10, Pcin14, Pcin15, Pcin20, Pcin21, Pcin23, Phci19, Phci27, Phci5).

| Population            | Sample Size | Na¹ | Ne² | F³  | Ho⁴ | He⁵ | uHe⁶ | F⁷   |
|-----------------------|-------------|-----|-----|-----|-----|-----|------|------|
| SW QLD                | 10          | 0.01| 0.01| 0.01| 0.01| 0.01| 0.01 | 0.01 |
| St. Bee’s Is.         | 0.053       | 0.081| 0.036| 0.216| 0.005| 0.003| 0.001| 0.123| 0.248| 0.002|
| Port Macquarie        | 0.090       | 0.020| 0.130| 0.175| 0.006| 0.084| 0.001| 0.114| 0.192| 0.004|
| Nth Stradbroke Is.    | 0.131       | 0.023| 0.020| 0.018| 0.001| 0.002| 0.001| 0.003| 0.012| 0.001|
| Nowendoc              | 0.044       | 0.004| 0.013| 0.032| 0.012| 0.053| 0.001| 0.016| 0.542| 0.001|
| Blue Mountains        | 0.078       | 0.031| 0.049| 0.074| 0.025| 0.001| 0.001| 0.001| 0.003| 0.001|
| Mt. Byron             | 0.054       | 0.032| 0.029| 0.062| 0.019| 0.049| 0.001| 0.001| 0.004| 0.001|
| Monaro                | 0.159       | 0.117| 0.150| 0.172| 0.124| 0.153| 0.138| 0.001| 0.001| 0.001|
| Magnetic Is.          | 0.069       | 0.006| 0.018| 0.034| 0.018| 0.037| 0.033| 0.124| 0.069| 0.001|
| Gunnedah              | 0.066       | 0.003| 0.011| 0.028| -0.002| 0.027| 0.027| 0.121| 0.006| 0.001|
| Clermont              | 0.096       | 0.034| 0.049| 0.062| 0.041| 0.061| 0.062| 0.141| 0.042| 0.047|

1. Background colours for PhiPT values indicate the magnitude of differentiation. For p values blue shading indicates significant differentiation, while white shading indicates non-significant differentiation.