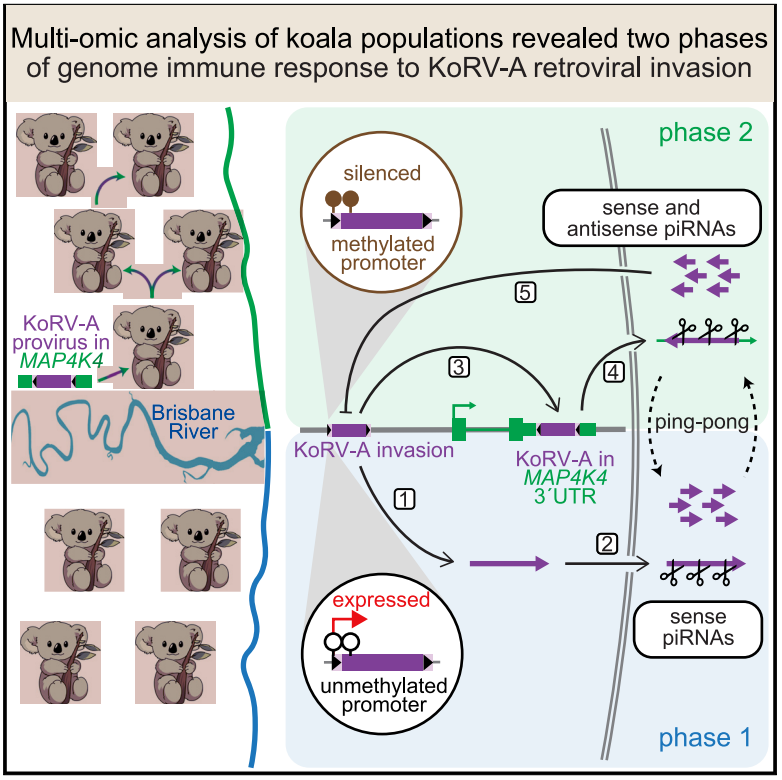
Cell



# Evolution of KoRV-A transcriptional silencing in wild koalas

## Graphical abstract




## Highlights

- The KoRV-A retrovirus is transducing koala germline cells and modifying the genome
- A subpopulation of koalas silences KoRV-A, suppressing replication
- These koalas produce antisense KoRV-A piRNAs, which guide silencing
- A genic UTR proviral insertion produces KoRV-A piRNAs and is sweeping to fixation

## Authors

Tianxiong Yu, Michaela B.J. Blyton, Milky Abajorga, ..., Keith Chappell, Zhiping Weng, William E. Theurkauf

## Correspondence

jeremy.luban@umassmed.edu (J.L.), k.chappell@uq.edu.au (K.C.), zhiping.weng@umassmed.edu (Z.W.), william.theurkauf@umassmed.edu (W.E.T.)

## In brief

Koala retrovirus-A is currently sweeping through wild koalas while infecting germ cells. In this work, the evolution of a sequence-specific genome defense system that tames this recent genome invader is captured.

Yu et al., 2025, Cell *188*, 2081–2093
April 17, 2025 
https://doi.org/10.1016/j.cell.2025.02.006

CellPress

Cell

CellPress

Article

# Evolution of KoRV-A transcriptional silencing in wild koalas

Tianxiong Yu,[1] Michaela B.J. Blyton,[2] Milky Abajorga,[3] Birgit S. Koppetsch,[3] Samantha Ho,[3] Bo Xu,[4] Zhongren Hu,[4] Jeremy Luban,[3,*] Keith Chappell,[2,5,*] Zhiping Weng,[1,*] and William E. Theurkauf[3,6,*]

[1]Department of Genomics and Computational Biology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA
[2]School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, QLD, Australia
[3]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA
[4]The School of Life Sciences and Technology, Tongji University, Shanghai 200092, China
[5]Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Brisbane, QLD, Australia
[6]Lead contact
*Correspondence: jeremy.luban@umassmed.edu (J.L.), k.chappell@uq.edu.au (K.C.), zhiping.weng@umassmed.edu (Z.W.), william.theurkauf@umassmed.edu (W.E.T.)
https://doi.org/10.1016/j.cell.2025.02.006

## SUMMARY

Koala retrovirus-A (KoRV-A) is spreading through wild koalas in a north-to-south wave while transducing the germ line, modifying the inherited genome as it transitions to an endogenous retrovirus. Previously, we found that KoRV-A is expressed in the germ line, but unspliced genomic transcripts are processed into sense-strand PIWI-interacting RNAs (piRNAs), which may provide an initial "innate" form of post-transcriptional silencing. Here, we show that this initial post-transcriptional response is prevalent south of the Brisbane River, whereas KoRV-A expression is suppressed, promoters are methylated, and sense and antisense piRNAs are equally abundant in a subpopulation of animals north of the river. These animals share a KoRV-A provirus in the *MAP4K4* gene's 3′ UTR that is spreading through northern koalas and produces hybrid transcripts that are processed into antisense piRNAs, which guide transcriptional silencing. We speculate that this provirus triggers adaptive transcriptional silencing of KoRV-A and is sweeping to fixation.

## INTRODUCTION

Integration into the host genome is an essential step in retroviral replication, complicating therapeutic interventions and potentially disrupting gene function.[1–5] By contrast, retroviruses have also driven adaptive evolution by rewiring regulatory circuits and providing proteins that have been co-opted for essential host functions and currently compose approximately 8% of the human genome.[6–10] While endogenization of infectious retroviruses has had a major role in the evolution of genome organization and function, retroviral infection of the germ line is very rare on a generational time scale, and the endogenization process is not understood.[11–13] Koala retrovirus subtype A (KoRV-A) is a gamma retrovirus that infects somatic and germline cells of wild koalas, leading to horizontal (direct contact) and vertical (parents to offspring) transmission.[14,15] The germline prevalence of KoRV-A is highest in northern regions of Australia and lowest in the far south.[15–17] This north-to-south sweep provides a unique opportunity to directly analyze the endogenization of an infectious retrovirus and the host response to genome invasion.

The 23- to 32-nt-long PIWI-interacting RNAs (piRNAs) are expressed in the germ line and have a conserved role in repressing established transposable elements, including endogenous retroviruses (ERVs).[18–25] Antisense piRNAs transcriptionally silence target elements by binding nascent transcripts and directing repressive DNA and chromatin modifications and post-transcriptionally silencing targets through sequence-specific transcript cleavage.[19,21,23] Sense piRNAs cleave complementary transcripts, generating precursors for antisense piRNA and driving processing through ping-pong amplification and phased biogenesis.[23,26–29] The most abundant piRNAs are derived from chromosomal domains composed of nested transposon fragments, termed piRNA clusters, suggesting sequence-specific silencing is acquired through transposition into one of these specialized loci, which leads to "adaptive" genome immunity.[18,20,21,30–33] The initial response of germ cells to a new mobile genetic element is not well understood. Our analysis of two koalas from south of the Brisbane River indicated that KoRV-A is initially expressed, but the resulting unspliced retroviral transcripts are processed into sense piRNAs.[34] Spliced KoRV-A transcripts encoding the envelope, by contrast, evade processing and appear indistinguishable from other protein-coding genes.[34,35] We also found that selective piRNA processing of unspliced endogenous retroviral transcripts is conserved from flies to placental mammals.[34] Host factors are co-opted by retroviruses to promote nuclear export of unspliced transcripts in somatic cells, and many of these factors are required for nuclear export of unspliced piRNA precursors in the germ

line.[36,37] Based on these observations, we propose that intron retention and recruitment of host factors, which is essential to retroviral replication in the soma, generates a sequence-independent "molecular pattern" that is recognized as non-self in germ cells, triggering processing into sense piRNAs.[34] This initial response thus post-transcriptionally suppresses viral replication by consuming genomic transcripts and simultaneously "primes" the germ cells for the transition to transcriptional silencing by generating sense piRNAs.[34]

In this study, we investigated the genome, transcriptome, and piRNA profiles of eight koalas from the region surrounding Brisbane. Consistent with our earlier observations, KoRV-A was expressed at relatively high levels, and unspliced KoRV-A transcripts were processed into sense-biased piRNAs in all of the animals south of the Brisbane River.[34] By striking contrast, KoRV-A expression was 10-fold lower in animals north of the Brisbane River, sense and antisense piRNAs were of roughly equal abundance, and KoRV-A promoter DNA was methylated. Significantly, the animals that transcriptionally silence KoRV-A carry an antisense provirus in the 3′ UTR of the *MAP4K4* gene, and the resulting hybrid transcripts are processed into antisense piRNAs. This provirus is sweeping through the northern koala population and is linked to low haplotype variation, consistent with positive selection. McClintock proposed that transposon mobilization could generate beneficial genetic diversity and drive adaptive evolution.[38] We propose that the integration of a KoRV-A provirus into the *MAP4K4* gene initiates adaptive transcriptional silencing of KoRV-A, stabilizing the germline genome. We speculate that this enhances reproductive fitness, which is driving the provirus to fixation.

## RESULTS

### KoRV-A is a dynamic genome invader

KoRV-A is actively invading the germline genome and exhibits significant heterogeneity among different koala populations and individuals of the same population.[34,39–41] To gain insight into the impact of KoRV-A on genome organization and functions, we collected testis, ovary, liver, and brain tissue samples from three koalas from the Sunshine Coast and five koalas from Currumbin, which lie immediately north and south of the Brisbane River, toward the northern end of the natural range (Figure 1A; Table S1), and sequenced their genomes (DNA sequencing [DNA-seq]) and transcriptomes (RNA sequencing [RNA-seq]; see STAR Methods).

DNA-seq identified 73–89 and 72–109 KoRV-A proviruses in the germline genomes of the three Sunshine Coast koalas and the five Currumbin koalas, respectively (Figure 1B; Table S2). Although a few KoRV-A proviruses are shared between koalas, most are unique to one koala. Only one provirus is shared by all eight animals from the Sunshine Coast and Currumbin (Figure S1A). By sharp contrast with KoRV-A, all the assayed koalas share insertions of three other active ERVs—Ko.ERV.1, Ko.ERVL.1, and Ko.ERVK.14 (Figures S1B–S1E, far-right bars), indicating that they are earlier genome invaders.

KoRV-A frequently recombines with Phascolarctos endogenous retroelement (PhER), an ancient endogenous retrovirus, resulting in the formation of recombinant KoRV (recKoRV).[16,41] To identify the structure of recKoRV, we performed long-read DNA-seq (using the Nanopore platform) on one koala from each region. Consistent with previous studies,[41] recKoRV in Sunshine Coast and Currumbin koalas consists of the 5′ end of KoRV-A, followed by the 3′ portion of PhER, and terminating in the 3′ end of KoRV-A (Figure S1F). Recombination of the viruses was likely facilitated by microhomology.[41] The structure of the dominant form or recKoRV is identical in both populations, consistent with a common ancestor. However, very few proviruses are shared between animals, indicating that the recombinant is active (Figure S1A). Remarkably, recKoRV does not encode full-length viral proteins and is not competent to replicate autonomously. The recombinant retroviral genome could be packaged into infectious particles by hijacking proteins from functional endogenous viruses. New insertions could then be generated by soma to germline infection within an animal, through horizontal transfer followed by germline infection, or a combination of both mechanisms.

### KoRV-A is silenced in animals north of the Brisbane River

The Sunshine Coast and Currumbin are only 200 km apart, but RNA-seq analysis of testes tissue demonstrates that KoRV-A expression is more than 10-fold lower in Sunshine Coast koalas than Currumbin koalas (mean expression = 16.2 vs. 170.7 reads per kilobase per million mapped reads [RPKM]; t test $p = 0.02$; Figures 1C, 1D, and S2; Table S3). In addition, analysis of the genomes of our koalas and koalas from the Koala Genome Survey[42] indicates that animals north of Brisbane, including animals from the Sunshine Coast, carry fewer KoRV-A insertions than animals south of Brisbane, with the change in average provirus number mapping to the Brisbane River (Figure 4C). KoRV-A expression in the liver and ovary, by contrast, is comparable in animals from the Sunshine Coast and Currumbin (1.4-fold lower in the liver and 1.9-fold higher in the ovary of Sunshine Coast koalas than Currumbin koalas, t test $p = 0.34$ in the liver, and $p$ value in the ovary not available due to the small sample size; Figures 1D and S2B; Table S3). Three established ERVs (Ko.ERV.1, Ko.ERVL.1, and Ko.ERVK.14) are also expressed at similar levels in Sunshine Coast and Currumbin animals, across all three tissues assayed (Figures 1D and S2B; Table S3). Reduced steady-state KoRV-A expression is therefore specific to the testes of the Sunshine Coast animals.

Germline silencing of established endogenous retroviruses is guided by antisense piRNAs. To determine if animals north of the Brisbane River had acquired piRNA-guided silencing of KoRV-A, we analyzed piRNA expression in animals from north and south of the river (Figures 2 and S3; Table S4). Mature piRNAs carry 2′-O-methylated 3′ ends, rendering them resistant to oxidation.[43–45] To distinguish piRNAs from other small RNAs and non-specific RNA breakdown products, we therefore analyzed libraries prepared from oxidized samples. However, this precluded the use of miRNAs as internal normalization standards. In addition, analysis was necessarily restricted to whole tissue, and animal age and disease state varied. The resulting data, normalized to sequencing depth, thus does not allow

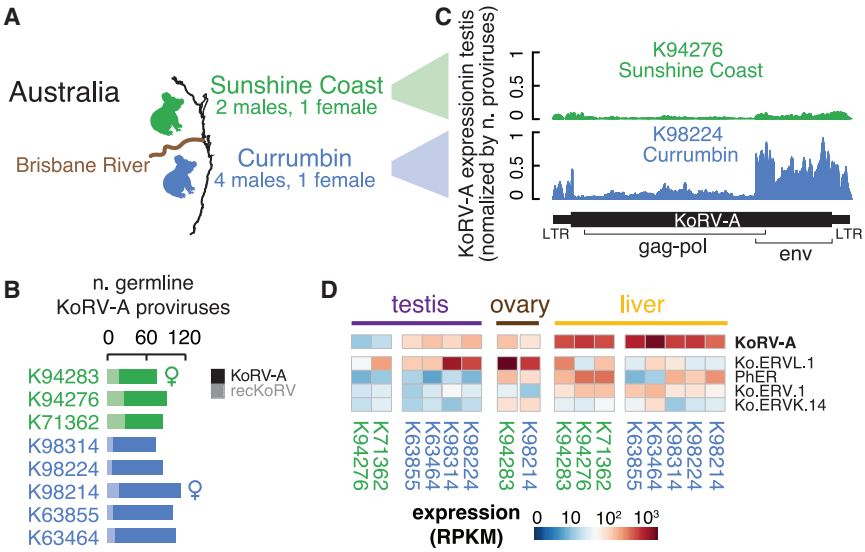



**Figure 1. KoRV-A genomic insertions and expression in koala tissues**

(A) Map showing the location of the Sunshine Coast and Currumbin, relative to the Brisbane River. Tissue was collected from two males and one female from the Sunshine Coast (green) and four males and one female from Currumbin (blue). Geographical color coding (green-blue) is maintained throughout the study.

(B) Number of mappable full-length KoRV-A proviruses (dark shading) and of recKoRV recombinant proviruses (lighter shading) in the germ line of each koala. The numbers of full-length KoRV-A and recKoRV proviruses for K94276 and K98224 were directly determined from long-read DNA sequencing, while those for the other six koalas were estimated based on read coverage over the KoRV-A consensus sequence. Number of unique and shared proviruses in the assayed koalas is shown in Figure S1.

(C) KoRV-A expression is suppressed in testes of Sunshine Coast koala K94276 relative to Currumbin koala K98224, determined by RNA-seq. Data from the other koalas are shown in Figure S4A. Signal was normalized to RNA-seq library sequencing depth and the number of KoRV-A proviruses per koala.

(D) Expression KoRV-A relative to four active endogenous EVS (in RPKM) in testes, liver, and ovary.

See also Figures S1 and S2.

reliable comparisons of absolute piRNA abundance but does reveal piRNA length and strand bias. This analysis shows that sense and antisense piRNAs targeting established elements (e.g., Ko.ERV.1) are equally abundant in koalas from north and south of the Brisbane River (Figures 2C and S3C; Table S4). In addition, the expression of 19 annotated protein-coding piRNA pathway genes and 371 piRNA clusters, which produce piRNA precursors, is expressed at similar levels in Sunshine Coast and Currumbin koalas (definition of statistical significance: t test $p < 0.05$ and fold change > 2; Table S3). By striking contrast, KoRV-A piRNAs are strongly sense-strand biased in Currumbin koalas, while sense and antisense KoRV-A piRNAs are equally abundant in animals from the Sunshine Coast (Figures 2A, 2B, S3A, and S3B; Table S4).

piRNA length is determined by the PIWI protein to which they are bound,[46–48] and antisense piRNAs targeting established endogenous koala retroviruses peak at 29 nt in all of the koalas analyzed (Figures 2C and S3C). KoRV-A piRNAs are sense-strand biased in animals from Currumbin, but antisense piRNAs are produced and show a distinct bimodal length distribution, with the major peak at 25 nt and a smaller 29 nt peak (Figures 2B, S3B, S3D, and S3E). By contrast, antisense KoRV-A piRNAs in the Sunshine Coast animals show a single peak centered at 29 nt, which is comparable to piRNAs targeting established endogenous retroviruses (Figures 2B, S3B, S3D, and S3E). The shift in KoRV-A piRNA strand bias and antisense piRNA length in the Sunshine Coast animals is therefore consistent with a transition to PIWI-piRNA complexes capable of guiding transcriptional silencing.

## KoRV-A promoter methylation

Antisense piRNAs guide PIWI proteins to genomic targets and promote transcriptional silencing by recruiting machinery that establishes repressive epigenetic marks, including cytosine DNA methylation.[21,23,49,50] We therefore used long-read DNA-seq on the Nanopore platform to estimate cytosine methylation levels in genomic DNA isolated from the testes of a Sunshine Coast (K94276) and a Currumbin koala (K98224; Figures 3 and S4; Table S5). In both Sunshine Coast and Currumbin koalas, the promoters of expressed protein-coding genes (expression levels $\geq$ 0.1 RPKM) are mostly unmethylated, whereas the promoters of silent protein-coding genes (levels < 0.1 RPKM) are mostly methylated (Figures S4A and S4B; median methylation level = 4.4% vs. 74.8%, Wilcoxon rank-sum test $p$ values < $2.2 \times 10^{-16}$). In addition, the promoters of genic piRNA clusters are unmethylated, and the promoters of intergenic piRNA clusters are methylated (Figures S4A and S4B), consistent with observations in mouse testis.[51] Promoter methylation levels for three established ERVs are also similarly high in the two koalas (Figures 3C and S4C). Due to the large number of Ko.ERVL.1 insertions (186 in the Sunshine Coast koala and 187 in the Currumbin koala), the small methylation difference between the two koalas (median = 82.6% vs. 77.7%) is statistically significant (Wilcoxon rank-sum test $p$ value = $1.3 \times 10^{-3}$). However, when we restricted our comparison to the Ko.ERVL.1 insertions shared by the two koalas ($n = 55$), the methylation levels were no longer significantly different (Figure S4D).

By striking contrast, almost all of the KoRV-A promoters in the Sunshine Coast koala are highly methylated (median = 87.8%; first and third quartiles = 84.2% and 90.0%; Figures 3A and 3B; Table S5), while KoRV-A promoter methylation shows significantly lower (Wilcoxon rank-sum test $p$ value < $2.2 \times 10^{-16}$) and highly variable in the Currumbin koala (median = 59.8%; first and third quartiles = 46.9% and 71.5%; Figures 3A and 3B; Table S5). The two koalas share two full-length KoRV-A insertions, with both loci heterozygous for the provirus, and these shared proviruses show higher promoter methylation in the Sunshine Coast

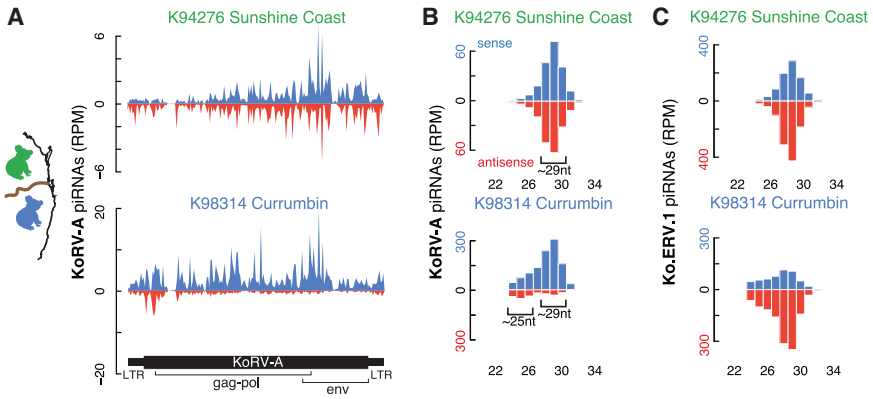



**Figure 2. Characteristics of KoRV-A and ERV piRNAs in koala testes**

(A) Sense and antisense piRNAs mapping to KoRV-A in testes from one Sunshine Coast koala and one Currumbin koala. Sunshine Coast koala testis produces similar amounts of sense and antisense KoRV-A piRNAs, contrasting with the sense-strand bias observed in Currumbin koala testis. A snapshot depicting the origin of these koalas is shown on the left. Oxidized small RNA-seq libraries are used. Data from unoxidized libraries and the other koalas are shown in Figures S5–S7.

(B) Bar plots illustrate the abundance and size distribution of sense and antisense piRNAs from KoRV-A and Ko.ERV.1 in the testes of one Sunshine Coast koala and one Currumbin koala. Notably, antisense piRNAs in Sunshine Coast koala testis are predominantly 29 nt, while in Currumbin koalas, they are divided into two sets based on size: 25 and 29 nt.

(C) Similar to (B) but for an established endogenous retrovirus Ko.ERV.1. Ko.ERV.1 piRNAs show similar sense-antisense ratio and size distribution in Sunshine Coast and Currumbin koala testes.

See also Figure S3.

koala than the Currumbin koala (Figures 3D, S4E, and S4F). The Sunshine Coast koalas have therefore acquired the ability to transcriptionally silence KoRV-A proviruses that are spread throughout the genome.

## A potential trigger for KoRV-A silencing

Antisense piRNAs guide transcriptional silencing of endogenous retroviruses and other mobile genetic elements, and clusters are the dominant source of piRNAs, supporting a model in which transposition of a mobile element into a cluster triggers silencing.[19,21,52,53] Consistent with this model, a subset of clusters expressed during pre-pachytene stages of mammalian spermatogenesis are enriched for endogenous retroviruses, and a full-length copy of the established ERV1 retrotransposon is in an annotated piRNA cluster in koalas.[19,34,54] However, short and long-read DNA-seq analysis of two Sunshine Coast koalas that silence KoRV-A failed to identify KoRV-A proviruses in piRNA clusters.

We therefore speculated that a KoRV-A provirus located outside a cluster could trigger silencing and assumed that a proviral trigger for silencing would be shared by both Sunshine Coast males and absent in the Currumbin males. The two Sunshine Coast koalas carry KoRV-A proviruses at 167 different genomic locations, but only three are present in both animals and absent from the Currumbin animals, reflecting the dynamic nature of the KoRV-A genome invasion (Figure 4A). We further speculated that a provirus leading to KoRV-A silencing in the germ line would enhance genome stability and provide a reproductive advantage, driving the spread of KoRV-A through the population. We therefore analyzed the distribution of the three candidate "trigger" proviruses in koalas from populations surrounding Brisbane, using DNA-seq data generated by the Koala Genome Survey (Figure 4; Table S6).[42] One of the proviruses present in both Sunshine Coast males was not detected in any of the sequenced koalas, and a second shared provirus was present in only one additional animal (Figure 4A). By striking contrast, the third provirus, which maps downstream of the *MAP4K4* gene, was present in 27 of 83 sequenced koalas north of the Brisbane River and is homozygous in four of these animals. This provirus is not present in any of the 101 koalas south of the Brisbane River (Figure 4A). In addition, a recent analysis of nine additional koalas from the region surrounding Brisbane identified the *MAP4K4* provirus in 5 of 6 animals north of the Brisbane River, while this provirus was not detected in the four animals south of the river.[40]

To determine if over-representation of the *MAP4K4* provirus in koalas north of the Brisbane River is statistically significant, we compared the population frequency of this provirus to all of the proviruses carried in the two Sunshine Coast koalas. As shown in Figure S5A, the *MAP4K4* provirus ranks third in population frequency with a Z test $p$ value of $1.1 \times 10^{-3}$. Two proviruses are present at a higher frequency than the *MAP4K4* provirus, but one is present in animals that do not silence KoRV-A, and the other is only in one of the Sunshine Coast animals. These proviruses are therefore unlinked to KoRV-A silencing but may provide distinct advantages to the host.

Significant over-representation of the *MAP4K4* provirus in the northern population is consistent with positive selection, which should lead to co-segregation of linked genetic variants. We therefore estimated haplotype variation associated with the *MAP4K4* provirus and the number of segregating sites ($nS_L$), a measure of haplotype homozygosity (see STAR Methods).[55] Figure 4D shows haplotype variation at the 3′ end of the *MAP4K4* gene in northern koalas, with chromosomes carrying the provirus indicated (box, top tracks). A window of low variation surrounds the provirus, also reflected by an $nS_L$ of 0.81, which is in the top 99th percentile among all polymorphic sites in the same contig (Figure S5B). We also determined the germline prevalence of the provirus as a function of geographic location. In the map in Figure 4B, circles are placed in the location of the sampled populations, circle size reflects sample size, and the fraction of animals carrying the *MAP4K4* provirus is indicated by filled sectors. The location of the Brisbane River is also shown. Within the sequenced animals, the provirus is only detected north of the Brisbane River, with population frequency highest near Brisbane and dropping moving north and

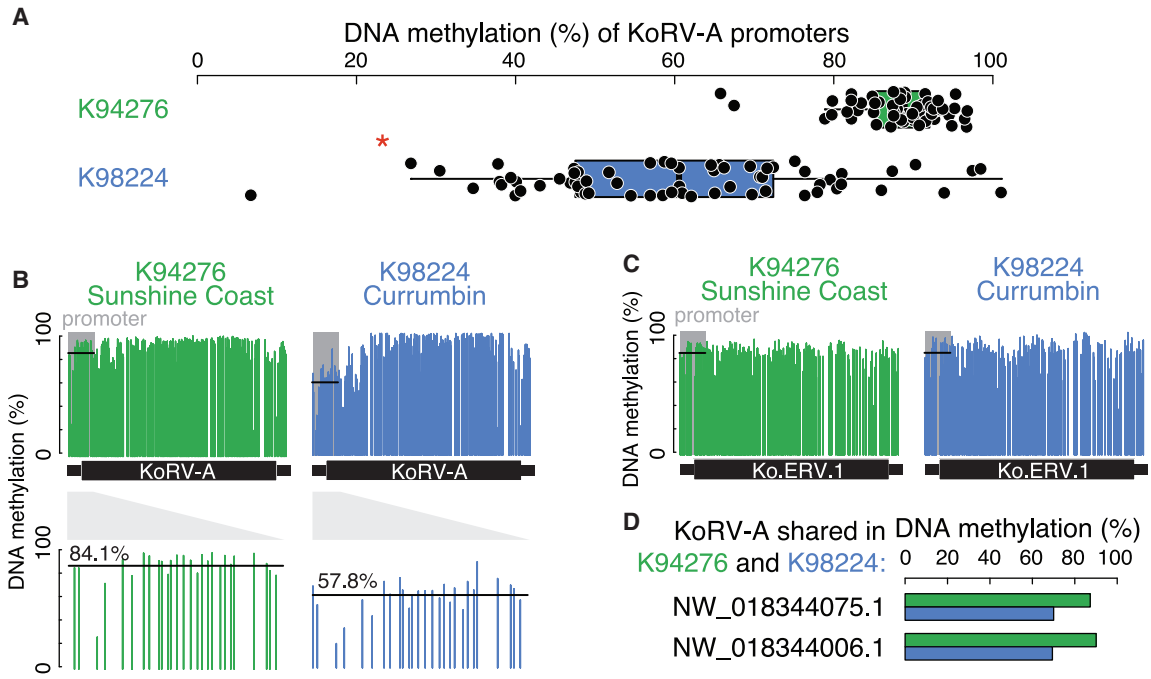



**Figure 3. DNA methylation levels of KoRV-A and Ko.ERV.1 in the two koala testes**

(A) Boxplots demonstrate DNA methylation levels in the promoters of full-length KoRV-A proviruses. Many full-length KoRV-A provirus promoters in Currumbin are DNA hypo-methylated in testis, whereas almost all full-length KoRV-A provirus promoters in Sunshine Coast are DNA hyper-methylated.

(B) DNA methylation levels of KoRV-A in testes of one Sunshine Coast koala and one Currumbin koala reveal that while other regions of KoRV-A are consistently DNA hyper-methylated in both Sunshine Coast and Currumbin koala testes. The promoter region of KoRV-A (marked in gray rectangles and zoomed in) exhibits higher DNA methylation in Sunshine Coast koala testis, leading to KoRV-A silencing, compared with Currumbin koala testis. Average promoter methylation levels are marked with horizontal lines.

(C) Same as (B) but for DNA methylation levels of Ko.ERV.1. Ko.ERV.1 is consistently DNA hyper-methylated in both Sunshine Coast and Currumbin koala testes.

(D) Bar plots show the DNA methylation level of two full-length KoRV-A provirus promoters shared in the Sunshine Coast koala and the Currumbin koala. Although the loci of the two KoRV-A insertions are the same in Currumbin and Sunshine Coast koala, the DNA methylation level is lower in Sunshine Coast koala testis, suggesting the difference is koala-specific instead of loci-specific.

See also Figure S4.

west. The *MAP4K4* provirus thus appears to have originated in an animal near Brisbane and is spreading through the northern population, with the Brisbane River forming a geographic boundary to efficient transmission. KoRV-A provirus copy number is significantly lower in animals north of the river, consistent with a role for the provirus in suppressing KoRV-A replication (Figure 4C).

To further explore a potential link between the *MAP4K4* provirus and KoRV-A silencing, we obtained testes samples from six additional animals near Brisbane, assayed for the presence of this provirus by PCR, and determined retrovirus and gene expression by RNA-seq (Tables S1 and S2). Three of the animals were heterozygous, and three were negative for the *MAP4K4* provirus (Figure 5A). Strikingly, KoRV-A was expressed at roughly 9-fold lower levels in the heterozygous animals relative to the animals that did not carry this provirus (14.2 vs. 134.3 RPKM; t test $p$ value = 0.04; Figures 5B and 5C; Table S2). Other retrotransposons, by contrast, are expressed at similar levels in all six animals (fold change < 2.15; Figure 5C; Table S2). Taken together, these observations are consistent with positive selection of the *MAP4K4* provirus, driven by KoRV-A silencing.

The *MAP4K4* provirus is not present in any of the animals from south of the Brisbane River that were sequenced by the Koala Genome Consortium, but two of the animals that carry the provirus and silence KoRV-A analyzed here were from immediately south of the river (Table S1). The provirus is in the same location and orientation as in these koalas and animals north of the river, indicating a common origin. The river thus appears to represent a "soft" barrier to migration that has recently been crossed by animals harboring this provirus. This may anticipate spread of the provirus through the southern population.

## The *MAP4K4* provirus produces antisense piRNAs

Antisense piRNAs guide transcriptional silencing, and the *MAP4K4* provirus is downstream of the *MAP4K4* gene in the antisense direction relative to transcription. Readthrough transcription from the gene could therefore extend through the provirus, generating the precursors for antisense piRNAs. As shown in Figure 6, long RNA-seq reads map to the unique sequences at the 5′ and 3′ junctions of the provirus and genome. Significantly, piRNAs also map to both junctions, with KoRV-A portions of these reads in the antisense orientation (Figure 6). Hybrid transcripts produced by readthrough transcription from the *MAP4K4*

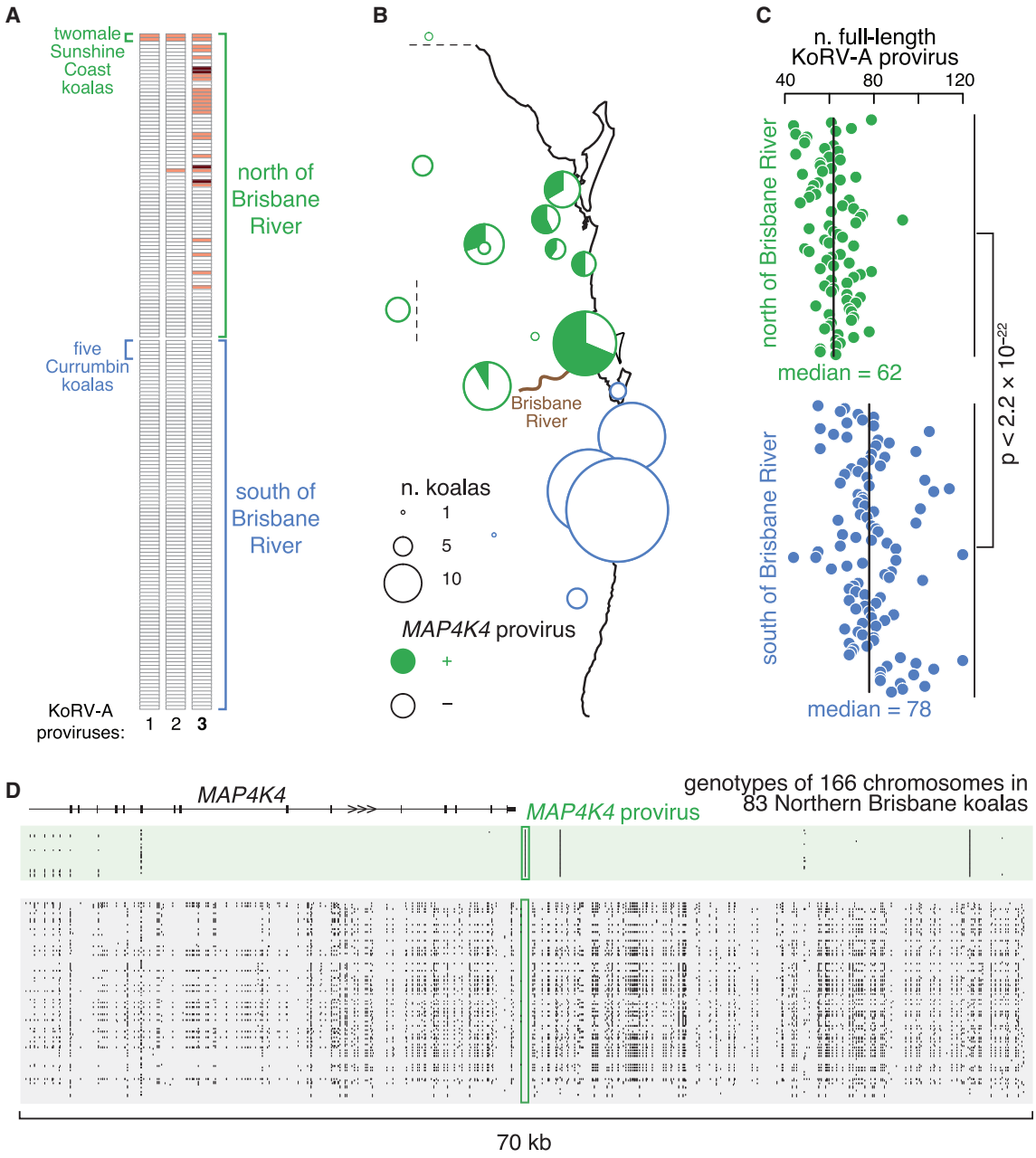



**Figure 4. Distribution of KoRV-A proviruses in koala populations**

(A) Population distribution of three proviral insertions shared by the two Sunshine Coast koalas that silence KoRV-A. Genomic sequence data from 83 koalas north of the Brisbane River (green) and 101 koalas south of the Brisbane River (blue) were analyzed for the presence of all three proviruses. Only the KoRV-A provirus located in contig NW_018343981.1 (*MAP4K4* KoRV-A provirus) was widespread, appearing in 27 of 83 koalas north of the Brisbane River. This provirus was not detected in koalas south of the Brisbane River.

(B) Pie charts show the percentage of koalas with the *MAP4K4* KoRV-A provirus in each koala population. The size of the pies reflects the size of the koala populations. The *MAP4K4* KoRV-A provirus is most abundant in the koala population at Sunshine Coast, suggesting it as the founder population. The *MAP4K4* provirus is spreading in all directions except southward, where the Brisbane River lies.

(C) The number of full-length KoRV-A proviruses in 83 koalas north of the Brisbane River and 101 koalas south of the Brisbane River. Median numbers are indicated.

(D) Genotypes of 31 chromosomes with the *MAP4K4* KoRV-A provirus and 135 chromosomes without the *MAP4K4* KoRV-A provirus. A 70 kbp region centered at the *MAP4K4* provirus is shown. Minor alleles at the northern Brisbane koala population are colored in black for each position.

See also Figure S5.

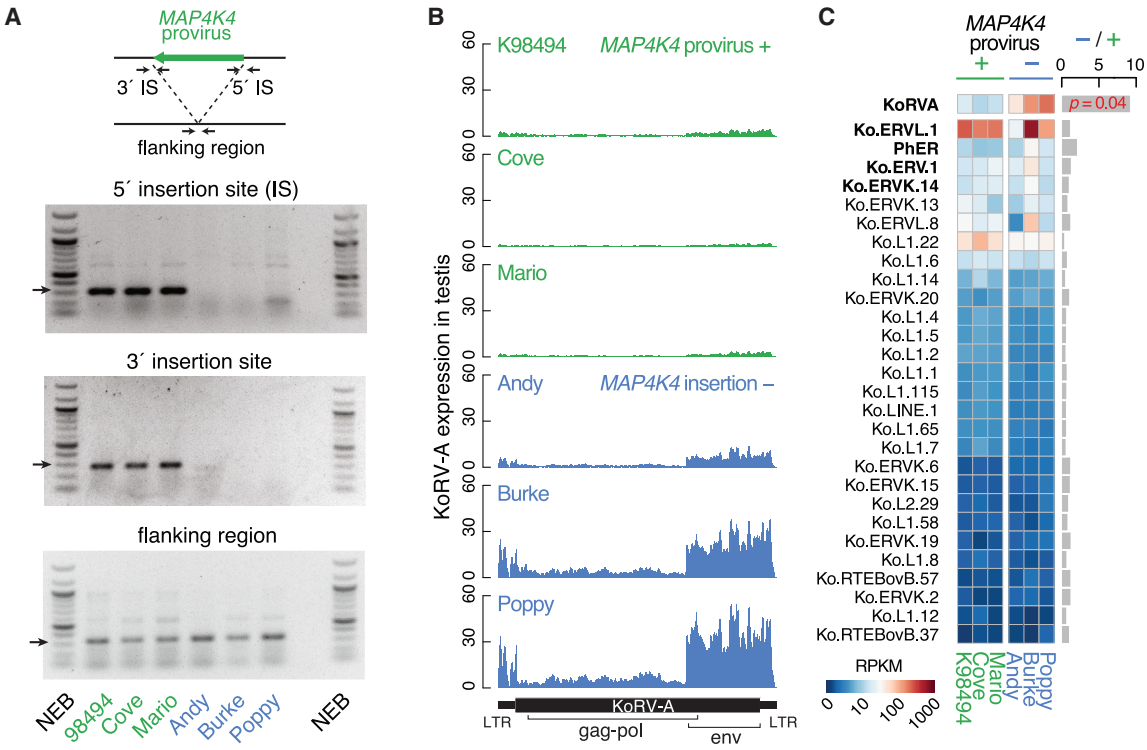



**Figure 5. Direct role of the *MAP4K4* provirus to KoRV-A silencing**

(A) PCR amplification of the 5′ insertion site (top), the 3′ insertion site (middle), and the flanking region (bottom) of the *MAP4K4* KoRV-A provirus in six additional koalas around the Brisbane area. 100 bp negative control (NEG) ladders are shown on the left and right of each panel. The sizes of amplified fragments are indicated by arrows. A schematic of the primers used for amplification is also presented.

(B) Coverage of RNA-seq reads from testes of six additional koala testes on the KoRV-A provirus. Koalas with the *MAP4K4* KoRV-A provirus are colored as green, with the other koalas colored as blue. Signal was normalized to RNA-seq library sequencing depth.

(C) Expression levels (in RPKM) of KoRV-A and other transposons in the testes of six additional koalas around Brisbane are shown in the heatmap. The barplot on the right indicates the fold change of transposon expression levels between the three koalas without the *MAP4K4* KoRV-A provirus and the three koalas with the *MAP4K4* KoRV-A provirus. Transposons with expression levels below 1 RPKM in all six koala testes are omitted.

gene are therefore processed into antisense piRNAs targeting KoRV-A.

We next sought to estimate the contribution of specific proviruses, including the *MAP4K4* provirus, to the antisense KoRV-A piRNA pool. For this analysis, we used long-read Nanopore DNA sequencing to define single nucleotide variants (SNVs) in individual KoRV-A proviruses in one Sunshine Coast and one Currumbin koala. Sequence variants were defined by deviations from a KoRV-A consensus derived from the Currumbin and Sunshine Coast animals (see STAR Methods). The Sunshine Coast animal carried 64 full-length proviruses. None completely matched the consensus, and no two were identical, with the majority showing 0.15%–0.2% variation (5–21 SNVs, median = 15). The *MAP4K4* provirus differs from the consensus at five sites, and all five polymorphisms were present in long RNAs and piRNAs (Figure S6). However, SNVs from numerous other proviruses were also present in long RNAs and piRNAs, indicating that numerous proviruses produce antisense piRNAs targeting KoRV-A. We therefore speculate that piRNAs derived from the *MAP4K4* provirus are loaded into a nuclear PIWI protein and that the resulting complexes guide transcriptional silencing and assembly of complexes that direct piRNA precursors production from dispersed elements (see below).

## DISCUSSION

Endogenous retroviruses represent approximately 8% of the human genome and have rewired critical regulatory circuits and been co-opted for essential functions but are also a source of genome instability linked to disease, including cancers.[1–10] While retroviral endogenization has had a significant impact on genome organization and evolution, germline infection is rare, and how these pathogens are tamed and co-opted for host functions is not well understood.[11–13] KoRV-A is spreading through wild koalas by horizontal transmission but is also transducing germ cells and incorporating into the inherited genome.[15,40] KoRV-A provirus copy number in the germ line is highest at the northern end of the range and lowest in the far southwestern region, where many animals do not carry full-length copies of the virus.[15–17] We have used the north-to-south sweep of KoRV-A to directly examine retroviral endogenization and the impact of this process on genome organization and function.

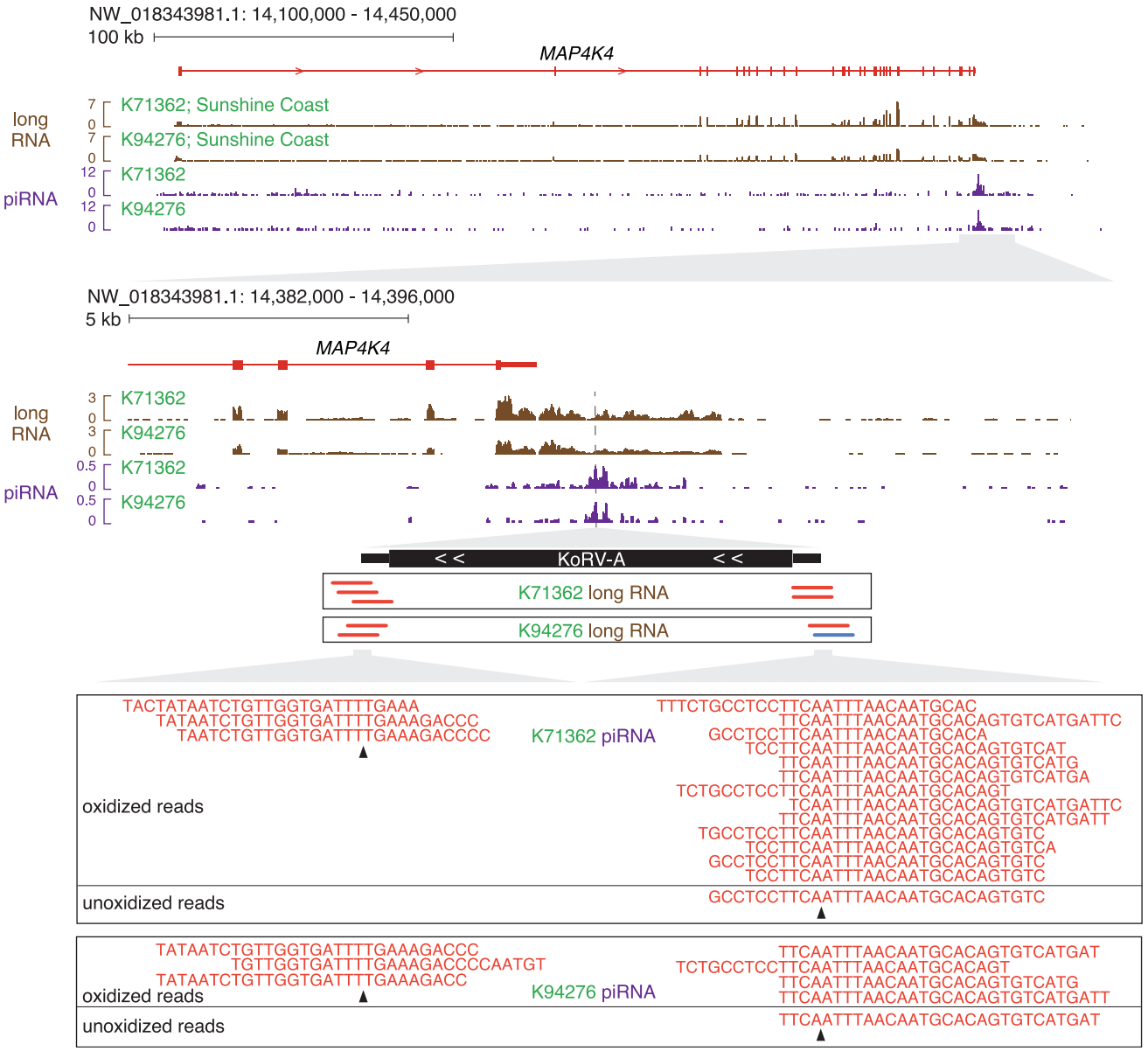



**Figure 6. piRNA production from the MAP4K4 KoRV-A provirus**

Top: genome browser snapshot for the region carrying the NW_018343981.1 KoRV-A provirus, showing RNA and piRNA expression. Expanded regions show the location of the provirus, which maps to readthrough transcripts from the gene, in a peak of piRNA expression.

Bottom: transcripts and piRNAs mapping to the junctions between the provirus and genome. Long RNAs and piRNAs map to both junctions.

See also Figure S6.

## An "innate" response to retroviral genome invasion

Antisense piRNAs have a conserved function in germline silencing of established endogenous retroviruses.[21,56,57] These small silencing RNAs, bound to PIWI proteins, base pair with targets and direct transcript cleavage and inhibitory modification of DNA and chromatin, leading to post-transcriptional and transcriptional silencing. Transcript cleavage by sense piRNAs produces precursors for antisense piRNAs, which are processed by the ping-pong and phased biogenesis pathways.[21,23,26,27] By contrast, our initial analysis of animals south of Brisbane found that KoRV-A is expressed at high levels in the germ line, but unspliced retroviral transcripts, which encode Gag, Pol, and the genome, are processed into sense-strand piRNAs.[34] We also found that selective piRNA production from unspliced retroviral transcripts is conserved from flies to placental mammals.[34] Sense piRNAs drive antisense piRNA production from complementary transcripts.[58] This initial response thus primes germ cells for antisense piRNA production and the transition to mature transcriptional silencing. We speculate that piRNA processing also reduces steady-state

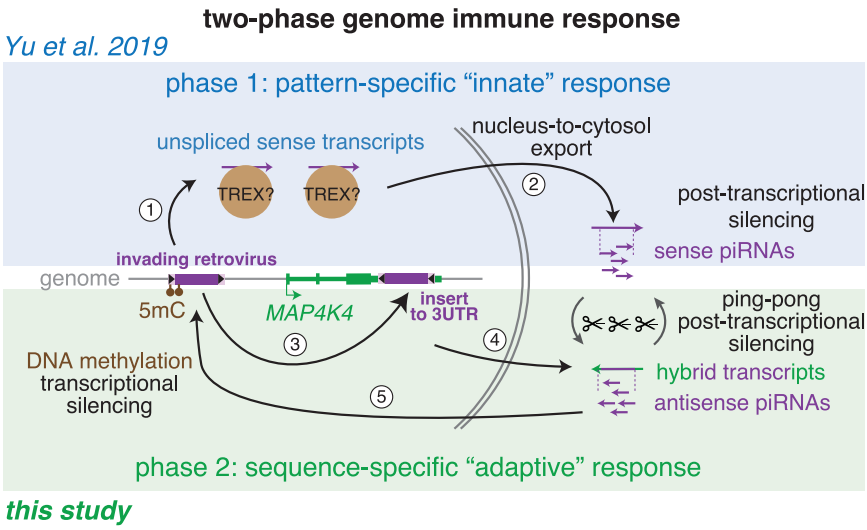



**Figure 7. Innate and adaptive piRNA response to endogenous retroviral invasion**

Diagram depicts a two-phased piRNA response to retroviral invasion. Upon a retrovirus entering the mammalian genome, its unspliced transcript serves as a conserved molecular pattern, initiating the pattern-specific "innate" piRNA response. This response leads to the cleavage of unspliced transcripts into sense-strand piRNAs. Subsequently, the sequence-specific "adaptive" piRNA response is triggered when the retrovirus inserts into the antisense strand of a protein-coding gene or cluster that produces piRNAs. Antisense piRNAs are then amplified by the ping-pong cycle with sense piRNAs, leading to post-transcriptional silencing of the retrovirus. Antisense piRNAs may also enter the nucleus and guide for transcriptional silencing of the provirus via DNA methylation.

levels of genomic transcripts, suppressing replication during the early stages of germline invasion.

Innate immunity depends on recognition of sequence-independent pathogen-associated molecular patterns (PAMPs).[59–61] Intron retention is essential to retroviral replication, and we propose that the absence of splicing generates a sequence-independent PAMP that is recognized as non-self by a germline-specific genome immune system, directing genomic retroviral transcripts to the piRNA biogenesis machinery.[34] Transcripts that retain introns are generally degraded prior to nuclear export by host quality control systems, but retroviruses co-opt host factors, including UAP56, the THO complex, NXT1, and CRM1, to bypass these systems and drive nuclear export and replication.[36,37,62] Genetic and biochemical studies in *Drosophila* have identified germline-specific factors that drive transcription and suppress splicing of piRNA precursor, and UAP56, THO, NXT1, and CRM1 mediate nuclear export of these transcripts, allowing piRNA biogenesis.[63,64] Host factors co-opted by retroviruses to drive export of unspliced genomic transcripts in the soma thus might direct these same transcripts to the piRNA biogenesis system in the *Drosophila* germ line. We therefore propose that host factors bound to unspliced retroviral transcripts drive replication in the soma but generate a "non-self" signature in the germ line, triggering processing by the piRNA machinery.[34]

## The transition to transcriptional silencing

Here, we identify animals that have transitioned from this initial innate response of KoRV-A invasion to mature adaptive silencing (Figure 7). In these animals, KoRV-A expression is suppressed, sense and antisense KoRV-A piRNAs are produced at approximately equal levels, and proviral promoters are methylated (Figures 1, 2, and 3). In most systems, including koalas, large genomic clusters, enriched for transposon sequences, produce piRNAs targeting endogenous retroviruses and other mobile elements, suggesting that adaptive silencing is triggered by transposition into a cluster and sequence incorporation into piRNA precursors.[21,34,53,65] However, analysis of piRNA cluster evolution suggests that these specialized loci evolve from read-through transcription of protein-coding genes.[32] Consistent with this hypothesis, KoRV-A is not present in annotated piRNA clusters in koalas that silence the virus, but these animals share an antisense KoRV-A provirus downstream of the *MAP4K4* gene, and resulting readthrough transcripts are processed into antisense piRNAs (Figures 4, 5, and 6). Significantly, this provirus is sweeping through the koala population and shows signatures of positive selection (Figure 4). The *MAP4K4* provirus may therefore represent the earliest stage of piRNA cluster formation, which leads to adaptive KoRV-A transcriptional silencing.

While our data support a role for the *MAP4K4* provirus in initiating antisense piRNA-directed adaptive immunity, our polymorphism analysis indicates that many proviruses contribute to the piRNA pool (Figure S6). This is consistent with data in flies and mice, indicating that specific transposon families are targeted by piRNAs derived from clusters and multiple isolated transposon insertions.[47,48,53] In flies, antisense piRNAs have been proposed to direct assembly of the chromatin-bound machinery that drives piRNA precursor production.[66] piRNAs derived from the *MAP4K4* provirus, in complex with a PIWI protein, could therefore direct assembly of the piRNA precursor production machinery on additional matching proviruses, enhancing piRNA production and target silencing.

We speculate that the *MAP4K4* provirus provides a reproductive advantage that is driving a sweep through the northern population. Genome analysis of dame/sire/offspring trios indicates that few new germline proviruses are generated in each generation, suggesting that suppressing viral replication may provide a modest advantage.[40] However, animals that do not carry the *MAP4K4* provirus express the KoRV-A envelope at high levels (Figures 1C, 5B, and S2A), and retroviral envelope proteins are fusogenic and linked to diseases, including cancers.[67–69] Envelope overexpression could therefore compromise male fertility, leading to a reproductive advantage in animals that carry the *MAP4K4* provirus and suppress *ENV* expression.

McClintock found that genotoxic stress activates transposons and proposed that stress-induced mobilization of genetic elements could provide beneficial genetic diversity that contributes to adaptive evolution.[70] KoRV-A invasion of the koala genome is

a source of genotoxic stress stress and is generating significant genetic variation.[14,15,34,40,71] Based on the data presented here, we propose that the KoRV-A provirus in the *MAP4K4* UTR is an adaptive mutation that enhances host fitness by triggering piRNA-guided transcriptional silencing.

### Limitations of the study

The spread of KoRV-A through wild koalas not only made this work possible but also presented significant experimental limitations. Samples could only be obtained from animals that had to be euthanized for humanitarian reasons when resources for rapid flash freezing were available. This significantly limited the number and geographic location of animals analyzed. Animal age and disease status also varied and were generally unknown, increasing variability, and assays were limited to whole organ samples, preventing analysis of cell-type-specific expression of long and short RNAs. The repeated nature of KoRV-A proviruses also precluded analysis of the fraction of piRNAs derived from the *MAP4K4* provirus.

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. William E. Theurkauf (William.Theurkauf@umassmed.edu).

### Materials availability

No unique reagents were generated during the course of these studies.

### Data and code availability

All data used in this study have been deposited to GEO with accession numbers GEO: GSE276616 (RNA-seq), GEO: GSE276617 (small RNA-seq), and GEO: GSE276618 (Illumina and Nanopore DNA-seq). Data from our previous study[34] are deposited to GEO with accession number GEO: GSE128122. This paper does not report original code.

## ACKNOWLEDGMENTS

We would like to thank members of the Theurkauf, Weng, and Luban laboratories for their critical comments and discussion on the manuscript. We thank Aaron Netto for his assistance in collecting the koala tissue samples and the staff at the Currumbin Wildlife Hospital, the Australia Zoo Wildlife Hospital, and the RSPCA Brisbane for their assistance with sample collection. We also thank Zoe Huang for drawing the koala and the Brisbane River clipart in the graphical abstract. This work was supported in part by NIH grants HD049116, R37AI147868, and U54AI170856 and Mathers Charitable Foundation grant 2011-01111.

## AUTHOR CONTRIBUTIONS

W.E.T., Z.W., J.L., K.C., and T.Y. conceived the project. T.Y. performed the computational analyses. T.Y., B.X., and Z.H. developed the pipeline to analyze Nanopore DNA-seq data. M.A. performed the RNA and small RNA sequencing experiments. B.S.K. performed the DNA sequencing experiments. K.C. and M.B.J.B. provided the koala tissues. M.B.J.B. performed the dissection and freezing of the koala tissues. S.H. and M.B.J.B. performed PCR and RNA-seq for the six additional koalas around Brisbane. T.Y., Z.W., J.L., and W.E.T. wrote the manuscript. All authors proofread and approved the manuscript.

## DECLARATION OF INTERESTS

Z.W. is a co-founder of Rgenta Therapeutics, and she serves on its scientific advisory board.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - DNA and RNA isolation
  - Library preparation
  - PCR for MAP4K4 KoRV-A provirus
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Mapping statistics
  - Transposons and gene expression
  - Identification of transposon insertions using Illumina and BGI DNA-seq data
  - Constructing transposon insertion sequences via Nanopore DNA-seq data
  - Refining the KoRV-A consensus sequence in Sunshine Coast and Currumbin koalas
  - Calculation of DNA methylation levels via Nanopore DNA-seq data
  - Analysis of small RNA-seq data
  - Exon-exon junction discovery and splicing index calculation in transposons
  - piRNA processing efficiency calculation
  - Identification of transcriptions and piRNAs at KoRV-A insertion junction sites
  - Polymorphism analysis for KoRV-A using DNA-seq, RNA-seq, and small RNA-seq data
  - Estimating the number of full-length KoRV-A and recKoRV insertions
  - Validation of full-length KoRV-A proviruses estimation approach
  - Inferring population information of koalas from the Koala Genome Survey
  - Testing positive selection for the MAP4K4 KoRV-A provirus

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.02.006.

Received: October 4, 2024
Revised: January 3, 2025
Accepted: February 12, 2025
Published: March 7, 2025

## REFERENCES

1. Jern, P., and Coffin, J.M. (2008). Effects of Retroviruses on Host Genome Function. Annu. Rev. Genet. *42*, 709–732. https://doi.org/10.1146/annurev.genet.42.110807.091501.
2. Mao, J., Zhang, Q., and Cong, Y.-S. (2021). Human endogenous retroviruses in development and disease. Comput. Struct. Biotechnol. J. *19*, 5978–5986. https://doi.org/10.1016/j.csbj.2021.10.037.
3. Weiss, R.A. (2006). The discovery of endogenous retroviruses. Retrovirology *3*, 67. https://doi.org/10.1186/1742-4690-3-67.
4. Mitchell, R.S., Beitzel, B.F., Schroder, A.R.W., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. *2*, E234. https://doi.org/10.1371/journal.pbio.0020234.
5. Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E., et al. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science *302*, 415–419. https://doi.org/10.1126/science.1088547.
6. Dupressoir, A., Lavialle, C., and Heidmann, T. (2012). From ancestral infectious retroviruses to bona fide cellular genes: Role of the captured

*syncytins* in placentation. Placenta *33*, 663–671. https://doi.org/10.1016/j.placenta.2012.05.005.

7. Frank, J.A., and Feschotte, C. (2017). Co-option of endogenous viral sequences for host cell function. Curr. Opin. Virol. *25*, 81–89. https://doi.org/10.1016/j.coviro.2017.07.021.
8. Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087. https://doi.org/10.1126/science.aad5497.
9. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.
10. Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. *13*, 283–296. https://doi.org/10.1038/nrg3199.
11. Johnson, W.E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. Nat. Rev. Microbiol. *17*, 355–370. https://doi.org/10.1038/s41579-019-0189-2.
12. Katzourakis, A., and Gifford, R.J. (2010). Endogenous Viral Elements in Animal Genomes. PLoS Genet. *6*, e1001191. https://doi.org/10.1371/journal.pgen.1001191.
13. Koonin, E.V., Dolja, V.V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology *479*, 2–25. https://doi.org/10.1016/j.virol.2015.02.039.
14. Stoye, J.P. (2006). Koala retrovirus: a genome invasion in real time. Genome Biol. *7*, 241. https://doi.org/10.1186/gb-2006-7-11-241.
15. Tarlinton, R.E., Meers, J., and Young, P.R. (2006). Retroviral invasion of the koala genome. Nature *442*, 79–81. https://doi.org/10.1038/nature04841.
16. Tarlinton, R.E., Legione, A.R., Sarker, N., Fabijan, J., Meers, J., McMichael, L., Simmons, G., Owen, H., Seddon, J.M., Dick, G., et al. (2022). Differential and defective transcription of koala retrovirus indicates the complexity of host and virus evolution. J. Gen. Virol. *103*, 001749. https://doi.org/10.1099/jgv.0.001749.
17. Blyton, M.D.J., Young, P.R., Moore, B.D., and Chappell, K.J. (2022). Geographic patterns of koala retrovirus genetic diversity, endogenization, and subtype distributions. Proc. Natl. Acad. Sci. USA *119*, e2122680119. https://doi.org/10.1073/pnas.2122680119.
18. Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363–367. https://doi.org/10.1126/science.1130164.
19. Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744–747. https://doi.org/10.1126/science.1142612.
20. Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., and Nakano, T. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203–207. https://doi.org/10.1038/nature04916.
21. Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell *128*, 1089–1103. https://doi.org/10.1016/j.cell.2007.01.043.
22. Grivna, S.T., Pyhtila, B., and Lin, H. (2006). MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. Proc. Natl. Acad. Sci. USA *103*, 13415–13420. https://doi.org/10.1073/pnas.0605506103.
23. Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., and Hannon, G.J. (2008). A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. Mol. Cell *31*, 785–799. https://doi.org/10.1016/j.molcel.2008.09.003.
24. Saito, K., Nishida, K.M., Mori, T., Kawamura, Y., Miyoshi, K., Nagami, T., Siomi, H., and Siomi, M.C. (2006). Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev. *20*, 2214–2222. https://doi.org/10.1101/gad.1454806.
25. Vagin, V.V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. (2006). A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. Science *313*, 320–324. https://doi.org/10.1126/science.1129333.
26. Han, B.W., Wang, W., Li, C., Weng, Z., and Zamore, P.D. (2015). piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science *348*, 817–821. https://doi.org/10.1126/science.aaa1264.
27. Gainetdinov, I., Colpan, C., Arif, A., Cecchini, K., and Zamore, P.D. (2018). A Single Mechanism of Biogenesis, Initiated and Directed by PIWI Proteins, Explains piRNA Production in Most Animals. Mol. Cell *71*, 775–790.e5. https://doi.org/10.1016/j.molcel.2018.08.007.
28. Mohn, F., Handler, D., and Brennecke, J. (2015). piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science *348*, 812–817. https://doi.org/10.1126/science.aaa1039.
29. Homolka, D., Pandey, R.R., Goriaux, C., Brasset, E., Vaury, C., Sachidanandam, R., Fauvarque, M.-O., and Pillai, R.S. (2015). PIWI Slicing and RNA Elements in Precursors Instruct Directional Primary piRNA Biogenesis. Cell Rep. *12*, 418–428. https://doi.org/10.1016/j.celrep.2015.06.030.
30. Li, X.Z., Roy, C.K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B.W., Xu, J., Moore, M.J., Schimenti, J.C., Weng, Z., et al. (2013). An Ancient Transcription Factor Initiates the Burst of piRNA Production during Early Meiosis in Mouse Testes. Mol. Cell *50*, 67–81. https://doi.org/10.1016/j.molcel.2013.02.016.
31. van Lopik, J., Alizada, A., Trapotsi, M.-A., Hannon, G.J., Bornelöv, S., and Czech Nicholson, B. (2023). Unistrand piRNA clusters are an evolutionarily conserved mechanism to suppress endogenous retroviruses across the Drosophila genus. Nat. Commun. *14*, 7337. https://doi.org/10.1038/s41467-023-42787-1.
32. Konstantinidou, P., Loubalova, Z., Ahrend, F., Friman, A., Almeida, M.V., Poulet, A., Horvat, F., Wang, Y., Losert, W., Lorenzi, H., et al. (2024). A comparative roadmap of PIWI-interacting RNAs across seven species reveals insights into *de novo* piRNA-precursor formation in mammals. Cell Rep. *43*, 114777. https://doi.org/10.1016/j.celrep.2024.114777.
33. Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199–202. https://doi.org/10.1038/nature04917.
34. Yu, T., Koppetsch, B.S., Pagliarani, S., Johnston, S., Silverstein, N.J., Luban, J., Chappell, K., Weng, Z., and Theurkauf, W.E. (2019). The piRNA Response to Retroviral Invasion of the Koala Genome. Cell *179*, 632–643.e12. https://doi.org/10.1016/j.cell.2019.09.002.
35. Hanger, J.J., Bromham, L.D., McKee, J.J., O'Brien, T.M., and Robinson, W.F. (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. https://doi.org/10.1128/JVI.74.9.4264-4272.2000.
36. Sakuma, T., Davila, J.I., Malcolm, J.A., Kocher, J.-P.A., Tonne, J.M., and Ikeda, Y. (2014). Murine Leukemia Virus Uses NXF1 for Nuclear Export of Spliced and Unspliced Viral Transcripts. J. Virol. *88*, 4069–4082. https://doi.org/10.1128/JVI.03584-13.
37. Sakuma, T., Tonne, J.M., and Ikeda, Y. (2014). Murine Leukemia Virus Uses TREX Components for Efficient Nuclear Export of Unspliced Viral Transcripts. Viruses *6*, 1135–1148. https://doi.org/10.3390/v6031135.
38. McClintock, B. (1953). Induction of Instability at Selected Loci in Maize. Genetics *38*, 579–599. https://doi.org/10.1093/genetics/38.6.579.
39. Lillie, M., Pettersson, M., and Jern, P. (2024). Contrasting segregation patterns among endogenous retroviruses across the koala population. Commun. Biol. *7*, 350. https://doi.org/10.1038/s42003-024-06049-0.
40. McEwen, G.K., Alquezar-Planas, D.E., Dayaram, A., Gillett, A., Tarlinton, R., Mongan, N., Chappell, K.J., Henning, J., Tan, M., Timms, P., et al.

(2021). Retroviral integrations contribute to elevated host cancer rates during germline invasion. Nat. Commun. *12*, 1316. https://doi.org/10.1038/s41467-021-21612-7.

41. Löber, U., Hobbs, M., Dayaram, A., Tsangaras, K., Jones, K., Alquezar-Planas, D.E., Ishida, Y., Meers, J., Mayer, J., Quedenau, C., et al. (2018). Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. Proc. Natl. Acad. Sci. USA *115*, 8609–8614. https://doi.org/10.1073/pnas.1807598115.
42. Hogg, C.J., Silver, L., McLennan, E.A., and Belov, K. (2023). Koala Genome Survey: An Open Data Resource to Improve Conservation Planning. Genes *14*, 546. https://doi.org/10.3390/genes14030546.
43. Horwich, M.D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., and Zamore, P.D. (2007). The Drosophila RNA Methyltransferase, DmHen1, Modifies Germline piRNAs and Single-Stranded siRNAs in RISC. Curr. Biol. *17*, 1265–1272. https://doi.org/10.1016/j.cub.2007.06.030.
44. Kirino, Y., and Mourelatos, Z. (2007). Mouse Piwi-interacting RNAs are 2′-O-methylated at their 3′ termini. Nat. Struct. Mol. Biol. *14*, 347–348. https://doi.org/10.1038/nsmb1218.
45. Saito, K., Sakaguchi, Y., Suzuki, T., Suzuki, T., Siomi, H., and Siomi, M.C. (2007). Pimet, the Drosophila homolog of HEN1, mediates 2′-O-methylation of Piwi- interacting RNAs at their 3′ ends. Genes Dev. *21*, 1603–1608. https://doi.org/10.1101/gad.1563607.
46. Carmell, M.A., Girard, A., van de Kant, H.J.G., Bourc'his, D., Bestor, T.H., de Rooij, D.G., and Hannon, G.J. (2007). MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. Dev. Cell *12*, 503–514. https://doi.org/10.1016/j.devcel.2007.03.001.
47. Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T.W., et al. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev. *22*, 908–917. https://doi.org/10.1101/gad.1640708.
48. Shoji, M., Tanaka, T., Hosokawa, M., Reuter, M., Stark, A., Kato, Y., Kondoh, G., Okawa, K., Chujo, T., Suzuki, T., et al. (2009). The TDRD9-MIWI2 Complex Is Essential for piRNA-Mediated Retrotransposon Silencing in the Mouse Male Germline. Dev. Cell *17*, 775–787. https://doi.org/10.1016/j.devcel.2009.10.012.
49. Le Thomas, A., Rogers, A.K., Webster, A., Marinov, G.K., Liao, S.E., Perkins, E.M., Hur, J.K., Aravin, A.A., and Tóth, K.F. (2013). Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev. *27*, 390–399. https://doi.org/10.1101/gad.209841.112.
50. Sienski, G., Dönertas, D., and Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell *151*, 964–980. https://doi.org/10.1016/j.cell.2012.10.040.
51. Yu, T., Fan, K., Özata, D.M., Zhang, G., Fu, Y., Theurkauf, W.E., Zamore, P.D., and Weng, Z. (2021). Long first exons and epigenetic marks distinguish conserved pachytene piRNA clusters from other mammalian genes. Nat. Commun. *12*, 73. https://doi.org/10.1038/s41467-020-20345-3.
52. Duc, C., Yoth, M., Jensen, S., Mouniée, N., Bergman, C.M., Vaury, C., and Brasset, E. (2019). Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. Genome Biol. *20*, 127. https://doi.org/10.1186/s13059-019-1736-x.
53. Khurana, J.S., Wang, J., Xu, J., Koppetsch, B.S., Thomson, T.C., Nowosielska, A., Li, C., Zamore, P.D., Weng, Z., and Theurkauf, W.E. (2011). Adaptation to P element transposon invasion in Drosophila melanogaster. Cell *147*, 1551–1563. https://doi.org/10.1016/j.cell.2011.11.042.
54. Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., and Jensen, S. (2013). Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. Proc. Natl. Acad. Sci. USA *110*, 19842–19847. https://doi.org/10.1073/pnas.1313677110.
55. Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Mol. Biol. Evol. *31*, 1275–1291. https://doi.org/10.1093/molbev/msu077.
56. Parhad, S.S., and Theurkauf, W.E. (2019). Rapid evolution and conserved function of the piRNA pathway. Open Biol. *9*, 180181. https://doi.org/10.1098/rsob.180181.
57. Czech, B., and Hannon, G.J. (2016). One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem. Sci. *41*, 324–337. https://doi.org/10.1016/j.tibs.2015.12.008.
58. Li, C., Vagin, V.V., Lee, S., Xu, J., Ma, S., Xi, H., Seitz, H., Horwich, M.D., Syrzycka, M., Honda, B.M., et al. (2009). Collapse of Germline piRNAs in the Absence of Argonaute3 Reveals Somatic piRNAs in Flies. Cell *137*, 509–521. https://doi.org/10.1016/j.cell.2009.04.027.
59. Brubaker, S.W., Bonham, K.S., Zanoni, I., and Kagan, J.C. (2015). Innate Immune Pattern Recognition: A Cell Biological Perspective. Annu. Rev. Immunol. *33*, 257–290. https://doi.org/10.1146/annurev-immunol-032414-112240.
60. Carpenter, S., and O'Neill, L.A.J. (2024). From periphery to center stage: 50 years of advancements in innate immunity. Cell *187*, 2030–2051. https://doi.org/10.1016/j.cell.2024.03.036.
61. Paludan, S.R., Pradeu, T., Masters, S.L., and Mogensen, T.H. (2021). Constitutive immune mechanisms: mediators of host defence and immune regulation. Nat. Rev. Immunol. *21*, 137–150. https://doi.org/10.1038/s41577-020-0391-5.
62. Pessel-Vivares, L., Houzet, L., Lainé, S., and Mougel, M. (2015). Insights into the nuclear export of murine leukemia virus intron-containing RNA. RNA Biol. *12*, 942–949. https://doi.org/10.1080/15476286.2015.1065375.
63. Mohn, F., Sienski, G., Handler, D., and Brennecke, J. (2014). The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. Cell *157*, 1364–1379. https://doi.org/10.1016/j.cell.2014.04.031.
64. Zhang, Z., Wang, J., Schultz, N., Zhang, F., Parhad, S.S., Tu, S., Vreven, T., Zamore, P.D., Weng, Z., and Theurkauf, W.E. (2014). The HP1 homolog rhino anchors a nuclear complex that suppresses piRNA precursor splicing. Cell *157*, 1353–1363. https://doi.org/10.1016/j.cell.2014.04.030.
65. Hirano, T., Iwasaki, Y.W., Lin, Z.Y.-C., Imamura, M., Seki, N.M., Sasaki, E., Saito, K., Okano, H., Siomi, M.C., and Siomi, H. (2014). Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. RNA *20*, 1223–1237. https://doi.org/10.1261/rna.045310.114.
66. Akkouche, A., Mugat, B., Barckmann, B., Varela-Chavez, C., Li, B., Raffel, R., Pélisson, A., and Chambeyron, S. (2017). Piwi Is Required during Drosophila Embryogenesis to License Dual-Strand piRNA Clusters for Transposon Repression in Adult Ovaries. Mol. Cell *66*, 411–419.e4. https://doi.org/10.1016/j.molcel.2017.03.017.
67. Mothes, W., Boerger, A.L., Narayan, S., Cunningham, J.M., and Young, J.A. (2000). Retroviral entry mediated by receptor priming and low pH triggering of an envelope glycoprotein. Cell *103*, 679–689. https://doi.org/10.1016/s0092-8674(00)00170-7.
68. Sheehy, A.M., Gaddis, N.C., Choi, J.D., and Malim, M.H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature *418*, 646–650. https://doi.org/10.1038/nature00939.
69. Martin, G.S. (2004). The road to Src. Oncogene *23*, 7910–7917. https://doi.org/10.1038/sj.onc.1208077.
70. McClintock, B. (1984). The significance of responses of the genome to challenge. Science *226*, 792–801. https://doi.org/10.1126/science.15739260.
71. Sarker, N., Fabijan, J., Owen, H., Seddon, J., Simmons, G., Speight, N., Kaler, J., Woolford, L., Emes, R.D., Hemmatzadeh, F., et al. (2020). Koala retrovirus viral load and disease burden in distinct northern and southern koala populations. Sci. Rep. *10*, 263. https://doi.org/10.1038/s41598-019-56546-0.

72. Yu, T., Huang, X., Dou, S., Tang, X., Luo, S., Theurkauf, W.E., Lu, J., and Weng, Z. (2021). A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies. Nucleic Acids Res. *49*, e44. https://doi.org/10.1093/nar/gkab010.

73. Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. *35*, 3100–3108. https://doi.org/10.1093/nar/gkm160.

74. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

75. Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinformatics *32*, 11. https://doi.org/10.1002/0471250953.bi1107s32.

76. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

77. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

78. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. https://doi.org/10.1093/bioinformatics/btu638.

79. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

80. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

81. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

82. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

83. Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. *40*, 37–52. https://doi.org/10.1093/nar/gkr688.

84. Fu, Y., Wu, P.-H., Beane, T., Zamore, P.D., and Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. BMC Genomics *19*, 531. https://doi.org/10.1186/s12864-018-4933-1.

85. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinforma. Oxf. Engl. *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

86. Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods *17*, 155–158. https://doi.org/10.1038/s41592-019-0669-3.

87. Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods *14*, 407–410. https://doi.org/10.1038/nmeth.4184.

88. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

89. Hofmeister, R.J., Ribeiro, D.M., Rubinacci, S., and Delaneau, O. (2023). Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat. Genet. *55*, 1243–1249. https://doi.org/10.1038/s41588-023-01415-w.

90. Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. Mol. Biol. Evol. *31*, 2824–2827. https://doi.org/10.1093/molbev/msu211.

91. Gainetdinov, I., Vega-Badillo, J., Cecchini, K., Bagci, A., Colpan, C., De, D., Bailey, S., Arif, A., Wu, P.-H., MacRae, I.J., et al. (2023). Relaxed targeting rules help PIWI proteins silence transposons. Nature *619*, 394–402. https://doi.org/10.1038/s41586-023-06257-4.

92. Zhang, Z., Theurkauf, W.E., Weng, Z., and Zamore, P.D. (2012). Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. Silence *3*, 9. https://doi.org/10.1186/1758-907X-3-9.

93. Wick, R.R., Judd, L.M., and Holt, K.E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. *20*, 129. https://doi.org/10.1186/s13059-019-1727-y.

94. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Chemicals, peptides, and recombinant proteins** | | |
| T4 RNA Ligase 2, truncated K227Q | NEB | Cat# M0351L |
| Sodium periodate | Millipore Sigma | Cat# 311448; CAS: 7790-28-5 |
| T4 RNA Ligase | ThermoFisher Scientific | Cat# AM2141 |
| Superscript III | ThermoFisher Scientific | Cat# 56575 |
| dNTP mix | NEB | Cat# N0447S |
| illustra PuRe Taq Ready-To-Go™PCR Beads | GE Healthcare | Cat# 27-9558-01 |
| DNase I | Qiagen | Cat# 1073395 |
| Hybridase™ Thermostable RNase H | Lucigen | Cat# H39500 |
| TURBO DNase | ThermoFisher Scientific | Cat# AM2238 |
| Random Primer | ThermoFisher Scientific | Cat# 58875 |
| RNaseOUT™ Recombinant Ribonuclease Inhibitor | ThermoFisher Scientific | Cat# 10777019 |
| Agencourt AMPure XP | Beckman Coulter Life Sciences | Cat# A63881 |
| RNase H | ThermoFisher Scientific | Cat# 18021-071 |
| DNA Polymerase I | NEB | Cat# M0209S |
| dUTP mix | ThermoFisher Scientific | Cat# R0133 |
| T4 DNA Polymerase | NEB | Cat# M0203L |
| Klenow DNA Polymerase | NEB | Cat# M0210S |
| T4 PNK | NEB | Cat# M0201L |
| Klenow 3' to 5' exo | NEB | Cat# M0212L |
| T4 DNA Ligase | Enzymatics | Cat# L603-HC-L |
| UDG | NEB | Cat# M0280S |
| Phusion Polymerase | NEB | Cat# M0530L |
| SequaGel UreaGel Concentrate | National Diagnostics | Cat# EC-830 |
| SequaGel UreaGel Diluent | National Diagnostics | Cat# EC-840 |
| SequaGel UreaGel Buffer | National Diagnostics | Cat# EC-835 |
| Certified Low Range Ultra Agarose | Bio-Rad | Cat# 1613107 |
| **Critical commercial assays** | | |
| ZR small-RNA™ PAGE Recovery Kit | Zymo Research | Cat# R1070 |
| QiaQuick Gel Extraction Kit | Qiagen | Cat# 28704 |
| mirVana™ miRNA Isolation Kit | ThermoFisher Scientific | Cat# AM1561 |
| Zymo RNA Clean and Concentrator™-5 | Zymo Research | Cat# R1015 |
| DNeasy Blood and Tissue Kit | Qiagen | Cat# 69504 |
| RNeasy Mini Kit | Qiagen | Cat# 74104 |
| NextSeq 500 Mid-Output v2 Kit (150 cycles) | Illumina | Cat# 20024907 |
| NextSeq 500 High-Output v2 Kit (75 cycles) | Illumina | Cat# 20024906 |
| NextSeq™ 2000 P3 XLEAP-SBS™ Reagent Kit (100 Cycles) | Illumina | Cat# 20100990 |
| **Deposited data** | | |
| Raw and analyzed data | This paper | GEO: GSE276616, GSE276618, GSE276618 |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| DNA-seq and RNA-seq from two male Currumbin koalas | Yu et al.[34] | GEO: GSE128122 |
| **Experimental models: Organisms/strains** | | |
| *Phascolarctos cinereus* (koala) | Currumbin Wildlife Hospital | K71362 |
| *Phascolarctos cinereus* | Currumbin Wildlife Hospital | K94276 |
| *Phascolarctos cinereus* | Currumbin Wildlife Hospital | K94283 |
| *Phascolarctos cinereus* | Currumbin Wildlife Hospital | K98224 |
| *Phascolarctos cinereus* | Currumbin Wildlife Hospital | K98314 |
| *Phascolarctos cinereus* | Currumbin Wildlife Hospital | K98214 |
| **Software and algorithms** | | |
| TEMP2 | Yu et al.[72] | https://github.com/w eng-lab/TEMP2; RRID: SCR_026074 |
| NAmmer | Lagesen et al.[73] | https://www.cbs.dtu.dk/services/RNAmme r/; RRID: SCR_017075 |
| Bowtie2 | Langmead and Salzberg[74] | https://bowtie-bio.sourceforge.net/bowtie2/index.shtml; RRID: SCR_016368 |
| Bowtie | Langmead[75] | https://bowtie-bio.sourceforge.net/index.shtml; RRID: SCR_005476 |
| STAR | Dobin et al.[76] | https://github.com/al exdobin/STAR; RRID: SCR_015899 |
| SAMtools | Li et al.[77] | https://samtools.sourceforge.net/; RRID: SCR_002105 |
| HTSeq | Anders et al.[78] | https://htseq.readthe docs.io/en/release_0.11.1/; RRID: SCR_005514 |
| Hisat2 | Kim et al.[79] | https://ccb.jhu.edu/s oftware/hisat2/index. shtml; RRID:SCR_015530 |
| Bedtools | Quinlan and Hall[80] | https://bedtools.readt hedocs.io/en/latest/; RRID: SCR_006646 |
| BWA | Li and Durbin[81] | http://bio-bwa.sourceforge.net/; RRID: SCR_010910 |
| cutadapt | Martin[82] | https://cutadapt.read thedocs.io/en/stable/; RRID: SCR_011841 |
| miRDeep2 | Friedländer et al.[83] | https://github.com/rajewsky-lab/mirdeep2; RRID: SCR_010829 |
| umitools | Fu et al.[84] | https://github.com/w eng-lab/umitools |
| Minimap2 | Li[85] | https://github.com/lh 3/minimap2; RRID: SCR_018550 |
| wtdbg2 | Ruan and Li[86] | https://github.com/ru anjue/wtdbg2; RRID: SCR_017225 |
| BLASTn | BLAST website | https://blast.ncbi.nlm.nih.gov/Blast.cgi; RRID: SCR_004870 |
| Nanopolish | Simpson et al.[87] | https://github.com/jts/nanopolish; RRID: SCR_016157 |
| bcftools | Danecek et al.[88] | https://github.com/sa mtools/bcftools; RRID: SCR_002105 |
| SHAPEIT5 | Hofmeister et al.[89] | https://github.com/od elaneau/shapeit5 |
| selscan2 | Szpiech and Hernandez.[90] | https://github.com/sz piech/selscan |

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Koala (*Phascolarctos cinereus*). Brain, liver, ovary, and testis tissues were isolated from eight wild koalas (6 males and 2 females; age unknown) that had been admitted to Currumbin Wildlife Hospital or Australia Zoo Wildlife Hospital for treatment and had to be euthanized for humanitarian reasons. More male samples were collected as this study focused on the genome response to retroviral invasion in male germline. Sample collection was performed under University of Queensland Animal Ethics Approval Certificate #ANRFA/SVS/335/17. Tissue was imported to the USA under US Fish and Wildlife Service permit #MA80344D-0.

## METHOD DETAILS

### DNA and RNA isolation

Total DNA was isolated from koala tissue samples using the DNeasy® Blood and Tissue Kit (QIAGEN). Total RNA was isolated from koala liver and testis tissues using the mirVana miRNA Isolation Kit (ThermoFisher Scientific).

### Library preparation

Small RNA sequencing libraries were prepared as outlined by the Zamore lab with some modifications.[91] For oxidized small RNA libraries, total RNA isolated using the mirVana miRNA Isolation kit was oxidized using 25 mM Sodium Periodate (NaIO4) in 30 mM borax, 30 mM boric acid, pH 8.6, for 30 min at 25 °C and consequently isolated using ethanol precipitation. Unoxidized and oxidized small RNAs were ligated to a 3' pre-adenylated adaptor at room temperature for 16 hours and purified from a 15% denaturing polyacrylamide-urea gel using the ZR small-RNA PAGE Recovery Kit (Zymo Research). 5' ligation was done at 25 °C for 2 hours and ligated products were isolated using ethanol. The ligated product was reverse transcribed, PCR amplified, and run on a 2% Certified Low Range Ultra Agarose (Bio-Rad) gel. After purification using the QIAquick® Gel Extraction Kit (QIAGEN), the small RNA libraries were single-end sequenced using the Illumina NextSeq 500 or 2000 system.

Strand-specific RNA sequencing libraries were prepared as previously described.[92] In brief, total RNA was isolated from frozen koala tissue samples using the mirVana miRNA isolation kit. rRNA was depleted using RNAseH and the DNA probes were digested using TurboDNase. RNAs longer than 200 nt were purified using the RNA Clean & Concentrator-5 kit (Zymo Research). These RNAs were fragmented and reverse transcribed. The second strand was synthesized by dUTP incorporation. After end repair and A-tailing, adaptor ligation and uracil-DNA glycosylase (UDG) treatment were performed. After PCR amplification, RNA libraries were paired-end sequenced using the Illumina NextSeq 500 system.

Standard PCR-free short-read DNA libraries for K71362 testis, K94276 testis, K94283 ovary, and K98214 ovary were constructed at BGI and sequenced using the BGISEQ-500 platform. Standard PCR-free short-read DNA libraries for K94283 brain and liver, K98214 ovary, K98224 brain, liver, and testis, and K98314 testis were constructed at the Broad and sequenced using the Illumina Novaseq 6000 platform. K98214 ovary was sequenced in both the BGI and Broad system and yielded 100% consistency in germline insertion identification of KoRV-A and other ERVs.

For Nanopore long-read DNA sequencing libraries. Genomic DNA was extracted from koala testes (K94276 and K98224) using the MagAttract HMW DNA kit from Qiagen, following the manufacturer's guidelines for isolating high-molecular-weight genomic DNA from frozen tissue. Short DNA fragments were subsequently removed with the Short Read Eliminator (SRE) kit from PacBio®, following the manufacturer's protocol. Library preparation was carried out using the Ligation Sequencing gDNA kit from Oxford Nanopore Technologies (ONT), and sequencing was performed to 27x (K94276) and 21x (K98224) genome coverage on the ONT PromethION platform at Cold Spring Harbor Laboratory Sequencing Technologies and Analysis Facility.

### PCR for MAP4K4 KoRV-A provirus

To further test for a link between the *MAP4K4* KoRV-A provirus and KoRV-A silencing in testis, we used PCR to assay for this provirus in six additional koalas from the Brisbane area. First, DNA was extracted from liver samples following the Digsol extraction protocol. We then used three PCR primer pairs to amplify the 5' insertion site, the 3' insertion site, and the flanking region of the *MAP4K4* KoRV-A provirus, respectively. The presence of the 5' insertion site was inferred from the amplification of a 314bp fragment using primers designed from the assemble of MAP4K4 provirus via Nanopore whole genome sequencing (forward: GATGTAAGAT CCAAGCTCTGG; reverse: CTGAGTCGCCCGAGTAC). Similarly, the presence of the 3' insertion site was inferred from the amplification of a 279 fragment (forward: GTGCATGACCACAGATATC; reverse: AGCCTGTTCAGTGAATGAA). The presence of the flanking region was inferred from the amplification of a 335 fragment (forward: GATGTAAGATCCAAGCTCTGG; reverse: GCAGAAGATGAG GCCTATGG). The annealing temperature of the amplification was 48 °C for the 5' insertion site and the flanking region, and 45 °C for the 3' insertion site. Specifically, the PCR reactions were performed using the following mix: H20 3.2 ul, 2x MyTaq reaction mix 5 ul, forward primer (10uM) 0.4, reverse primer (10uM) 0.4, and template 1.0. Finally, the PCR was programmed as follows: 95 °C 1 min, 30 cycles of 95 °C 15 sec, annealing temperature 15 sec, 72 °C 30sec, and final extension of 72 °C 10 min.

## QUANTIFICATION AND STATISTICAL ANALYSIS

Statistics were computed with R (version 4.4.2) using unpaired Student's t-test, Wilcoxon rank-sum test, and Z-test. Statistic methods, sample size and *p*-values can be found in figures, figure legends, or results. For box plots, boxes present median, lower quartile, and upper quartile, while whiskers donate maximum and minimum after removing outliers (more than 3/2 times of upper quartile or less than 3/2 times of lower quartile). Significance was defined as $p$-value < 0.05. *: $p < 0.05$, n.s.: not significant.

### Mapping statistics

All mapping statistics for DNA-seq, RNA-seq, and small RNA-seq data are available in Table S7.

### Transposons and gene expression

We first removed all RNA-seq reads that mapped to ribosomal RNAs (rRNAs) in the reference koala genome. We annotated the rRNAs using RNAmmer[73] (version 1.2) with default parameters and mapped RNA-seq reads to rRNAs using Bowtie2[74] (version 2.2.5) with default parameters. The remaining RNA-seq reads were mapped to the reference koala genome using STAR[76] (version 020201). PCR duplicates were eliminated based on UMI tags and alignment information using Umitools[84] (version 0.2.0). The resulting reads were then annotated to protein-coding genes, non-coding RNAs, and piRNA genes using HTSeq[78]

(version 0.9.1). Gene expression levels were quantified in reads per million unique mapped reads per kilobase (RPKM) using custom bash scripts.

Following rRNA removal, RNA-seq reads were aligned directly to transposon consensus sequences, as defined in our 2019 study, using Hisat2[79] (version 2.1.0) with parameters "-k 100 –no-mixed –no-discordant". PCR duplicates were removed using Umitools, and reads mapping to multiple loci were counted as one divided by the number of times they map to the genome. Transposon expression levels were determined using Bedtools[80] (version 2.27.1) and normalized to RPKM based on the total reads uniquely mapped to the koala genome."

### Identification of transposon insertions using Illumina and BGI DNA-seq data

We used TEMP2[72] (version 0.1.6) to identify transposon insertions in Illumina and BGI DNA-seq data. Firstly, TEMP2-insertion was used with default parameters to identify new transposon insertions corresponding to the reference genome. Secondly, TEMP2-absence with default parameters was used to confirm the absence of transposon insertions annotated in the reference genome within the dataset. The newly identified transposon insertions, along with those present in the reference genome, constituted the comprehensive set of transposon insertions in a specific koala tissue. To ensure high confidence, transposon insertions with frequencies no greater than 0.3 across all koala samples were filtered out, and the remaining insertions were considered to be present in the germline genome.

### Constructing transposon insertion sequences via Nanopore DNA-seq data

To leverage Nanopore long-read DNA-seq data for constructing transposon insertion sequences, we first employed Guppy-basecaller[93] (version 5.0.16) with the high accuracy configuration to convert the Nanopore raw signals into raw reads. Subsequently, the raw reads were aligned to the koala genome using Minimap2[85] (version 2.17) with the parameters "-x map-ont". Soft clipped reads (those with clipped 5' or 3' ends) and spanning reads (those containing internal deletions) were collected. A soft clipped read or spanning read was designated as supporting a transposon insertion if the soft clipped or deleted portion could be mapped back to transposon consensus sequences using Minimap2 with parameters "-x map-ont". Adjacent reads supporting transposon insertions were clustered together and used for assembling the insertion sequence via wtdbg2[86] (version 2.5) with default parameters. An insertion with a sequence originating from both KoRV-A and PhER was classified as a recKoRV insertion, and the structure of recKoRV was defined accordingly for Currumbin and Sunshine Coast koalas. Hence, insertions with sequences solely from KoRV-A were deemed as full-length KoRV-A insertions. Finally, BLASTn[94] (version 2.15.0) was used to compare the KoRV-A insertion sequences with KoRV-A consensus sequences, enabling the identification of nucleotide polymorphisms for each KoRV-A insertion.

### Refining the KoRV-A consensus sequence in Sunshine Coast and Currumbin koalas

To refine the KoRV-A consensus sequence in Sunshine Coast and Currumbin koalas, we first retrieved the original KoRV-A consensus sequence (accession number: NC_039228.1) from the NCBI GeneBank database.[35] We then compared the assembled KoRV-A sequences from each insertion to the original KoRV-A consensus sequence. Any nucleotide with an alternative allele frequency greater than 0.7 was considered for refinement. The refined KoRV-A sequence was then designated as the consensus sequence in Sunshine Coast and Currumbin koalas, and subsequently utilized for polymorphism analysis.

### Calculation of DNA methylation levels via Nanopore DNA-seq data

To calculate DNA methylation levels using Nanopore DNA-seq data, we used Nanopolish[87] (version 0.13.3) with default parameters. Nanopolish output was further fed into its module calculate_methylation_frequency.py to calculate the frequency of DNA methylation levels with parameter "-c 2". This procedure was used to identify DNA methylation levels in both the koala genome and transposon consensus sequences. The DNA methylation levels within transposon consensus sequences are indicative of the average DNA methylation levels for each transposon. To ascertain DNA methylation levels for a specific transposon insertion, we first collected Nanopore DNA-seq reads that supported the transposon insertion, specifically soft-clipped and spanning reads alignable to transposon consensus sequences. These reads were then used to calculate DNA methylation levels for the inserted transposon. Similarly, Nanopore DNA-seq reads supporting the reference allele without the insertion, with overhangs of at least 100 bp at the insertion site, were collected and used to compute DNA methylation levels for the reference allele.

We further found there is a 1.07-fold difference of the median DNA methylation levels in intergenic regions in the Currumbin and the Sunshine Coast koala testis. Therefore we used a factor 1.07 for calculating the DNA methylation levels in the Currumbin koala; any values larger than 100% after multiplying 1.07 are clipped to 1.

### Analysis of small RNA-seq data

Because piRNAs are 2'–O-methylated at their 3' termini, which renders them resistant to oxidation, we performed oxidized small RNA transcriptome sequencing to enrich piRNAs. We first removed the adapter sequence (TGGAATTCTCGGGTGCCAAGGAACTCCA GTCACCGATGTATCTCGT) from all small RNA-seq reads using cutadapt[82] (version 1.15). Reads shorter than 18nt or longer than 32nt were filtered out prior to alignment. Subsequently, reads mapping to rRNAs, miRNAs, snoRNAs, snRNAs, and tRNAs were excluded. rRNA and miRNA annotations were obtained from our previous study using RNAmmer and miRDeep2, respectively.[34,73,83]

The remaining small RNA-seq reads were independently mapped to the reference koala genome, transposon consensus sequences, and spike-in sequences using Bowtie[75] (version 1.1.0) with parameters "-v 1 -a –best –strata". PCR duplicates were removed using Umitools[84] with default parameters.

We quantified piRNA abundance at piRNA clusters and individual transposon insertions in the genome using small RNA-seq samples. Abundance was normalized by sequencing depth, defined as the total number of uniquely genome-mapped reads after the removal of rRNAs, miRNAs, snoRNAs, snRNAs, and tRNAs, as this normalization method is more stable than normalizing to spike-ins. Nucleotide content and ping-pong amplification were analyzed for reads mapping to the genome, transposons, and piRNA clusters. For ping-pong amplification analysis, 5' to 5' overlaps between all pairs of piRNAs mapping to opposite genomic strands were computed. The Z-score for the 10-nt overlap was calculated using the 1-9 nt and 11-30 nt overlaps as the background. No difference was identified in KoRV-A piRNA ping-pong Z-score between Sunshine Coast koala testis and Currumbin koala testis. To further confirm our results, we performed unoxidized small RNA transcriptome sequencing and did the same analysis. Notably, the oxidized and unoxidized libraries show consistent results (Figure S3).

### Exon-exon junction discovery and splicing index calculation in transposons

To discover exon-exon junctions and calculate splicing indexes in transposons, we followed our previously described methods with a modification specific to the KoRV-A exon-exon junction.[34]

Briefly, RNA-seq data that could not be mapped to rRNAs were utilized to identify high-confidence exon-exon junctions in transposons, and splicing indexes were calculated on the identified exon-exon junctions. Firstly, RNA-seq reads fully mappable to the reference koala genome (using Bowtie2[74] in very sensitive end-to-end mode with soft clipping disabled) were considered primary transcripts and discarded. Subsequently, the remaining RNA-seq reads were mapped to transposon consensus sequences using Hisat2[79] with parameters "–no-mixed –no-discordant" to detect exon-exon junctions. For splicing index calculation, long RNA reads devoid of rRNAs and small RNA reads lacking rRNAs/miRNAs/tRNAs/snRNAs/snoRNAs were used. The calculation proceeded as follows: Reads were mapped to transposon consensus sequences using Hisat2 ("–no-mixed –no-discordant") with the detected junctions. Hisat2 output comprised reads mapping to splice sites (unspliced reads) and those mapping to exon-exon junctions. The junction-mapping reads were then aligned back to the reference genome (Bowtie2 in very sensitive end-to-end mode with soft clipping disabled), and reads that mapped were discarded. Long RNA reads spanning at least 7 bps of the exon-exon junctions and piRNA reads spanning at least 3 bps of the exon-exon junctions were designated as spliced reads. The splicing indexes for each exon-exon junction detected by long RNA and piRNA were computed as the ratio of spliced reads to unspliced reads (plus one to avoid division by zero).

We observed that recKoRV transcripts overlap with the 5' splice site of the KoRV-A exon-exon junction. Therefore, the unspliced reads at the 5' splice site could originate from either the unspliced KoRV-A transcript or the recKoRV transcript. Consequently, the splicing index for the KoRV-A exon-exon junction was defined as the ratio of spliced reads to unspliced reads at the 3' junction site (plus one to avoid division by zero). In line with our previous results, we found the KoRV-A transcript splicing index is higher than the KoRV-A piRNA splicing index in the two additional Currumbin koalas (K98224 and K98314).[34] Similarly, the pattern is also observed in the two male Sunshine Coast koalas, suggesting unspliced KoRV-A transcripts are preferentially processed into sense piRNAs in both koalas north and south of the Brisbane River (Figure S7).

### piRNA processing efficiency calculation

We calculated piRNA processing efficiency as the ratio of piRNA to long RNA abundance for exons and the intron of KoRV-A. Specifically,[34] processing efficiency was determined by dividing piRNA reads per million mapped reads by long RNA reads per million mapped reads. However, accurately mapping long RNA reads to the edges of KoRV-A using Hisat2 proved challenging, whereas piRNAs, due to their short length, could be more mappable. This discrepancy could artificially inflate processing efficiency. Hence, we excluded edge sequences (100 bps) of transposons from the analysis of piRNA processing efficiency. Similar to our previous findings, we observed higher KoRV-A piRNA processing efficiency in the intron region than in the exon regions in both Sunshine Coast and Currumbin koalas, suggesting unspliced KoRV-A transcripts are more likely to be recognized and processed into sense piRNAs (Figure S7).[34]

### Identification of transcriptions and piRNAs at KoRV-A insertion junction sites

To verify the transcription and piRNA processing of specific KoRV-A insertions, we aimed to identify transcripts and piRNAs spanning their insertion junction sites. Firstly, we constructed junction sequences for KoRV-A insertions by concatenating the 5' and 3' ends of KoRV-A insertion sequences (assembled by Nanopore long-read sequencing) with their flanking region sequences. Subsequently, we aligned small RNA and RNA reads to the koala reference genome and removed those that were alignable. The remaining reads were then aligned back to the junction sequences. We implemented strict filtering to eliminate potential false positives resulting from the contamination of other small RNAs and transcripts. Only small RNAs with lengths ranging from 24 to 32 nucleotides and uniquely aligned with either a first nucleotide T or a tenth nucleotide A were considered confident piRNAs. Similarly, only RNA reads with lengths exceeding 65 nucleotides and aligning to the flanking sequences were designated as valid transcripts. Finally, piRNAs with overhangs of at least three base pairs on the insertion sequence and at least five base pairs on the flanking sequence were identified as deriving from specific KoRV-A insertions.

Likewise, RNA reads with overhangs of at least ten base pairs at junction sites were classified as transcripts originating from specific KoRV-A insertions.

### Polymorphism analysis for KoRV-A using DNA-seq, RNA-seq, and small RNA-seq data

To investigate whether a single KoRV-A insertion dominates or multiple KoRV-A insertions contribute to the transcription and the production of sense and antisense piRNAs, we conducted polymorphism analysis by comparing allele frequencies obtained from DNA-seq, RNA-seq, and small RNA-seq data. Initially, we mapped DNA-seq reads, RNA-seq reads, and 24-32nt small RNA-seq reads to the transposon consensus sequences using BWA mem,[81] Hisat2,[79] and Bowtie,[75] respectively. Subsequently, we utilized Samtools mpileup[77] (version 1.13) to summarize allele frequencies for each polymorphic locus in transposons. Considering that KoRV-A can recombine with PhER to form recKoRV, we partitioned the KoRV-A sequence into two parts: 1) sequence shared by full-length KoRV-A and recKoRV, at 5' and 3' ends; 2) sequence specific to full-length KoRV-A, at the internal region. Finally, allele frequencies of RNA, sense and antisense piRNAs were compared to DNA for all polymorphic loci to infer which KoRV-A insertions are the source of transcription and piRNA production.

### Estimating the number of full-length KoRV-A and recKoRV insertions

Full-length KoRV-A and recKoRV insertions are indistinguishable by short-read DNA-seq due to the limitation of read length. Therefore, we sought a detour to estimate the number of full-length KoRV-A and recKoRV insertions utilizing the structure of recKoRV identified by Nanopore DNA-seq in Sunshine Coast and Currumbin koalas. We first mapped Illumina/BGI DNA-seq reads to the KoRV-A consensus sequence directly using BWA mem[81] with default parameters and calculated the coverage of DNA-seq reads across KoRV-A. The average coverage of DNA-seq reads at 3000-6000 of KoRV-A ($cov_{internal}$), which is specific to full-length KoRV-A, and the average coverage of DNA-seq reads at 300-1028 and 7619-8300 of KoRV-A ($cov_{ends}$), which is shared between full-length KoRV-A and recKoRV, are calculated. 1028 is the leftmost junction site and 7619 is the rightmost junction site for all recKoRV variants in KoRV-A. Then for each koala tissue, the numbers of full-length KoRV-A ($num_{full\text{-}length\text{-}KoRV\text{-}A}$) and recKoRV ($num_{recKoRV}$) insertions are defined by the number of KoRV-A insertions ($num_{KoRV\text{-}A}$):

$$num_{full-length-KoRV-A} = num_{KoRV-A} \times cov_{internal} \div (cov_{internal} + cov_{ends})$$

$$num_{recKoRV} = num_{KoRV-A} \times cov_{ends} \div (cov_{internal} + cov_{ends})$$

### Validation of full-length KoRV-A proviruses estimation approach

We validated our approach for estimating the number of full-length KoRV-A proviruses using data from two koalas (K94276 and K98224), for which both short-read and long-read DNA sequencing data are available. For K94276, among the 79 identified KoRV-A proviruses, our estimation indicated that 61 were full-length, closely aligning with the 63 full-length proviruses observed from long-read DNA sequencing. For K98224, the estimation perfectly matched the observation, with 76 out of 83 KoRV-A proviruses identified as full-length. These results demonstrate that our estimation method reliably predicts the number of full-length KoRV-A proviruses.

### Inferring population information of koalas from the Koala Genome Survey

DNA-seq data was collected from the Koala Genome Survey for 430 koalas across Australia.[42] We further excluded 12 koalas that are captive with unknown origins and 27 koalas without sufficient coverage ($\leq$20x), yielding 391 koalas from 57 populations in total. We used TEMP2[72] (version 0.1.6) to identify transposon insertions present in the germline genome of these koalas. Utilizing the insertions of KoRV-A, Ko.ERV.1, Ko.ERVL.1, Ko.ERVK.14, and PhER, we performed hierarchical clustering with the Ward method to infer the evolutionary relationships for the 391 koalas from the Koala Genome Survey and the 12 koalas from our studies (data not shown). Koalas from different populations were clearly clustered and two main clusters were formed for the koalas from the Queensland state and very north New South Wale state, one includes our 3 Sunshine Coast koalas (cluster1) and the other one includes our 5 Currumbin koalas (cluster2). We further confirmed the origins of koalas in the two clusters and found Brisbane River is a geographical barrier for them. Thus we classified those populations mainly located in cluster1 as the north of the Brisbane River population and those populations mainly located in cluster2 as the south of the Brisbane River population. In total, 83 koalas were classified as north of the Brisbane River, and 101 koalas were classified as south of the Brisbane River.

### Testing positive selection for the MAP4K4 KoRV-A provirus

To assess whether the MAP4K4 provirus is under positive selection, we employed the haplotype-based statistics $nS_L$.[55] Single nucleotide variations (SNVs) at contig NW_018343981.1 were first called for our koalas and those from the Koala Genome Survey using bcftools (version 1.19) with default parameters.[88] The genotypes of the *MAP4K4* provirus were integrated into the called SNVs to generate a final genotype file. This file was then phased using SHAPEIT5 (version 5.1.1) with the "phase_common_static" command and default settings.[89] The phased genotypes were subsequently analyzed using selscan2 (version 2.0.3) with the "–nsl" parameter

to calculate $nS_L$ scores for the MAP4K4 provirus and other polymorphic sites."[90] The $nS_L$ statistic evaluates the ratio of haplotype homozygosity for minor and major alleles, focusing on northern Brisbane koalas. Extreme $nS_L$ values (> 99th or < 1st percentile) indicate selection, with high positive values suggesting selection on the minor allele and negative values indicating selection on the major allele. The *MAP4K4* provirus, a minor and derived allele, exhibited a top extreme $nS_L$ value, suggesting that it may be under positive selection.

# Supplemental figures

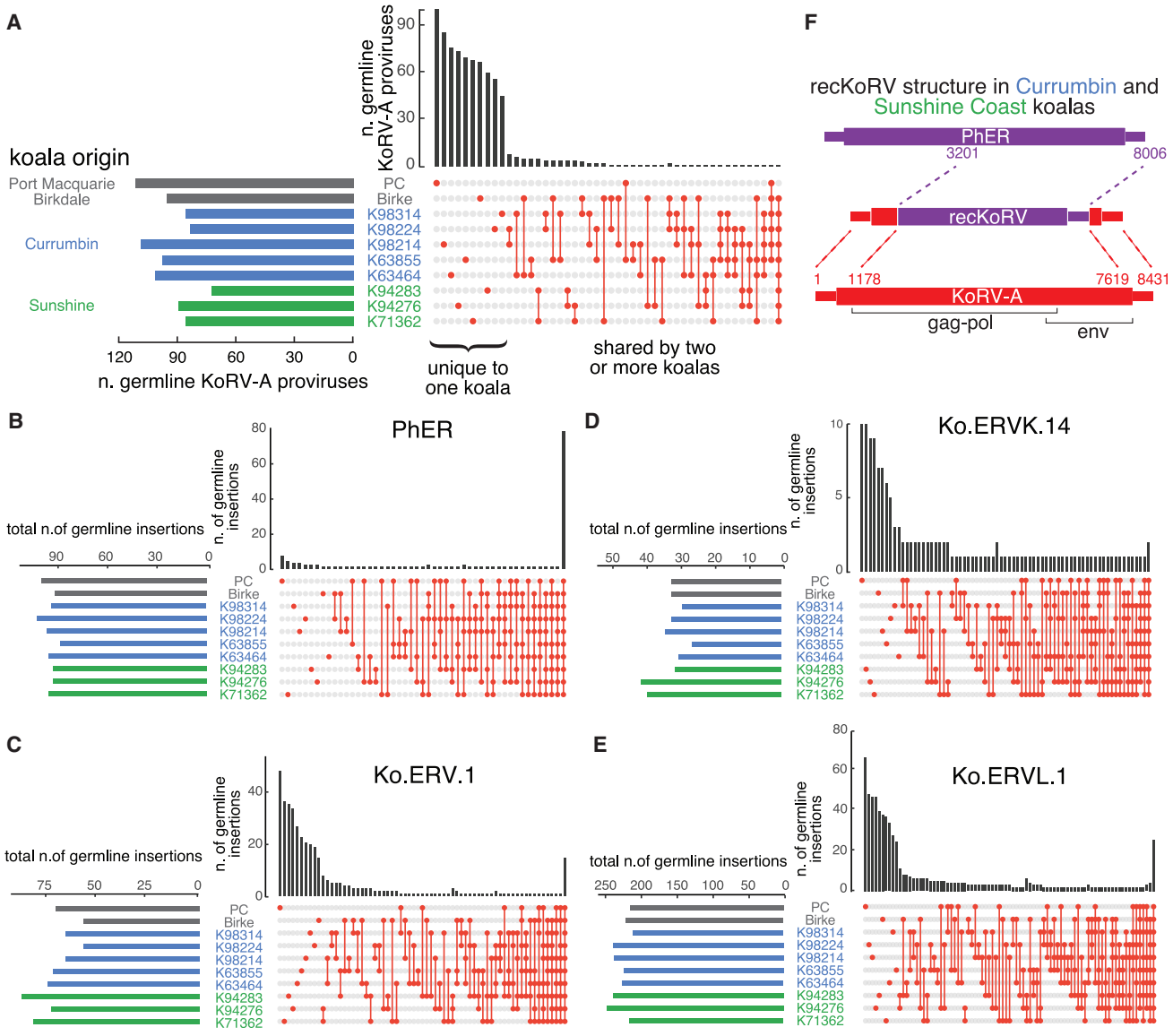
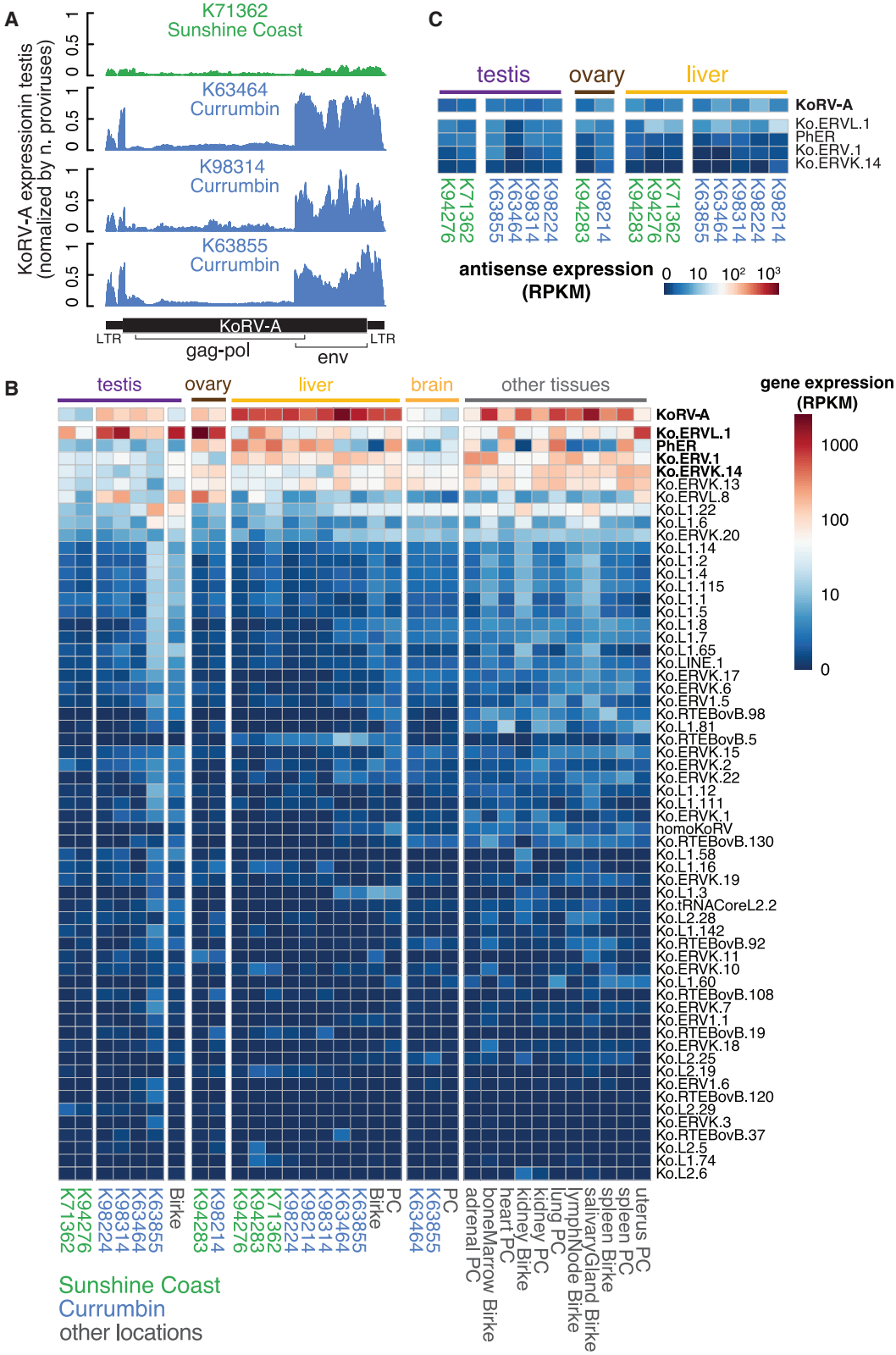



**Figure S1. KoRV-A and other ERV proviruses in koalas: Unique and shared, related to Figure 1**

(A) The left bar plot illustrates the total number of KoRV-A proviruses in the germline genome of each koala, including two previously published koala genomes, Birke from Birkdale and PC from Port Macquarie. The right UpSet plot displays the number of KoRV-A proviruses in all possible koala combinations, unique to one koala or shared by multiple koalas. Most KoRV-A insertions are koala-specific, confirming the recent KoRV-A invasion.

(B–E) These panels mirror the analysis presented in (A), focusing on four other ERVs: (B) PhER, (C) Ko.ERV.1, (D) Ko.ERVK.14, and (E) Ko.ERVL.1. The analysis reveals that most PhER insertions ($n$ = 73, depicted in the far-right bar) are ubiquitous across the studied koalas. However, the presence of koala-specific PhER insertions points to active transposition events. Similar patterns of both shared and unique insertions are observed for the other ERVs, suggesting ongoing evolutionary dynamics. Due to space limitations in (E), koala combinations with just one Ko.ERVL.1 insertion are omitted, although all combinations of insertions are included in (A)–(C).

(F) The recombination structure of recKoRV detected by Nanopore long-read genome sequencing in Currumbin and Sunshine Coast koalas




(legend on next page)

**Figure S2. Expression levels of KoRV-A and other transposons in koala tissues, related to Figure 1**

(A) Similar to Figure 1C, RNA-seq reads mapped to the KoRV-A provirus in testes of the other Sunshine Coast koala K71362 and three additional Currumbin koalas K98314, K63464, and K63855. Signal was normalized to RNA-seq library sequencing depth and the number of KoRV-A proviruses per koala.

(B) Expression levels of transposons in testis (purple), ovary (brown), liver (yellow), brain (orange), and other somatic tissues (gray). Transposons with expression levels below 1 RPKM in all tissue samples are omitted.

(C) Same as Figure 1D, but for antisense transcript levels of KoRV-A and other ERVs.

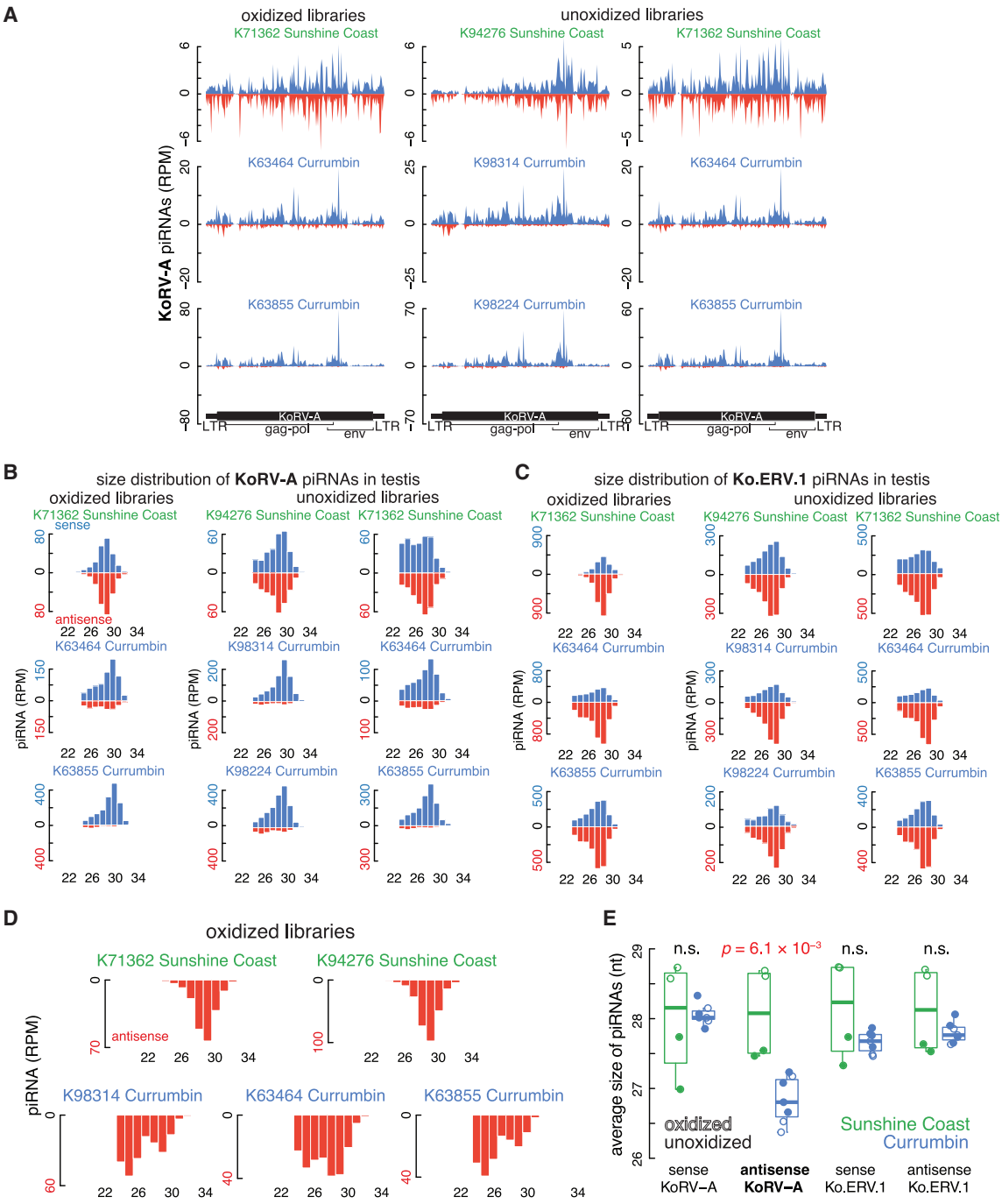



**Figure S3. Characteristics of KoRV-A and ERV piRNAs in koala testis, related to Figure 2**

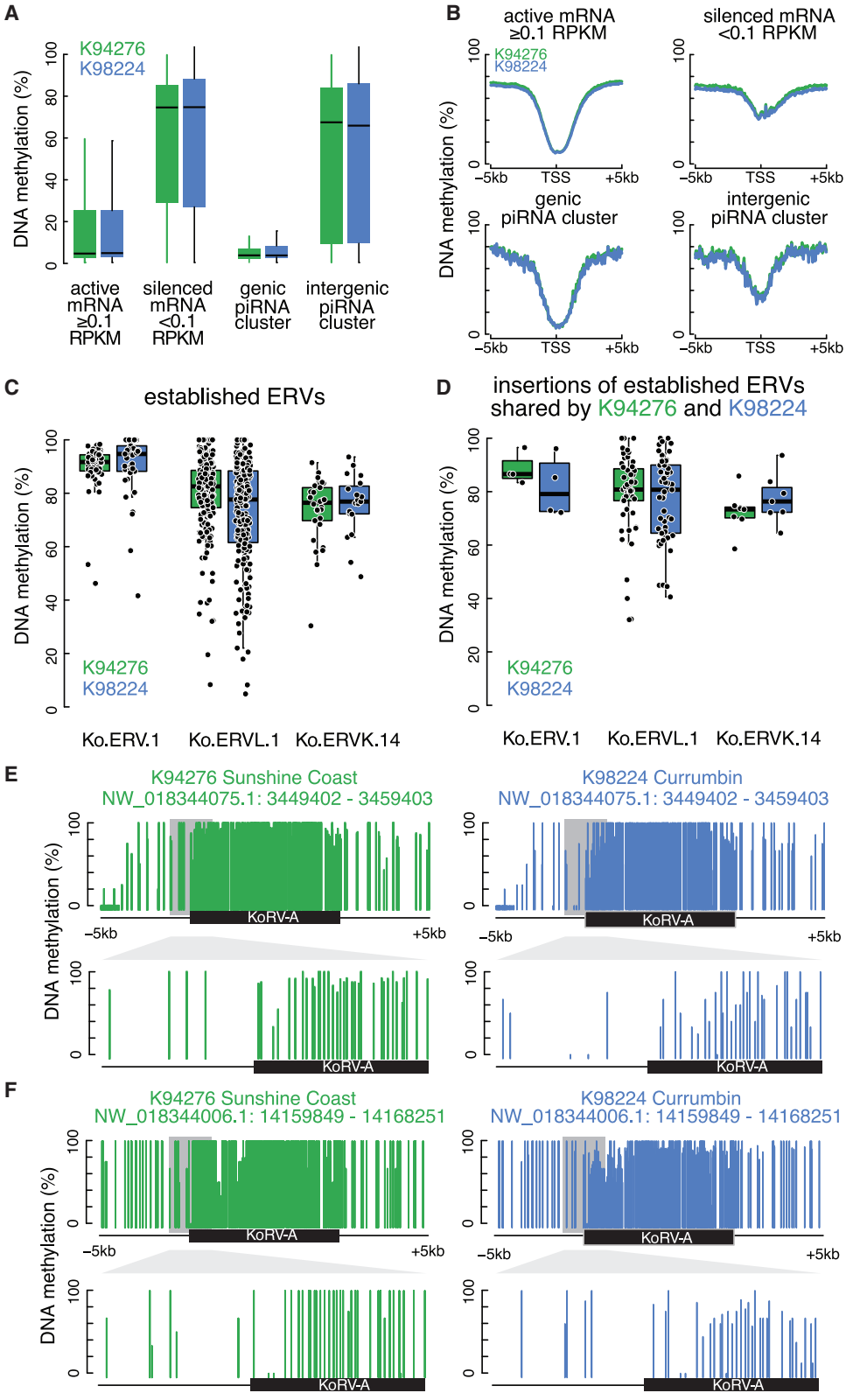
(A) Similar to Figure 2A, coverage of sense and antisense piRNAs at KoRV-A from oxidized libraries in testes of another Sunshine Coast koala and two additional Currumbin koalas, and from unoxidized libraries in testes of two Sunshine Coast koalas and four Currumbin koalas.

(B) Similar to Figure 2B, bar plots represent the abundance and size distribution of KoRV-A piRNAs from oxidized libraries in testes of another Sunshine Coast koala and two additional Currumbin koalas, and from unoxidized libraries in testes of two Sunshine Coast koalas and four Currumbin koalas.

(C) Same as (B), but for Ko.ERV.1.

(D) Bar plots depict the abundance and size distribution of antisense KoRV-A piRNAs from oxidized and unoxidized libraries in koala testes. The plot is autoscaled.

(E) Boxplots show the average length of sense or antisense piRNAs for KoRV-A or Ko.ERV.1 in koala testes. Length from oxidized libraries is marked on hollow dots, while that from unoxidized libraries is marked in filled dots.




*(legend on next page)*

**Figure S4. DNA methylation levels in the two koala testes, related to Figure 3**

(A) DNA methylation levels at promoters of active protein-coding genes, silenced protein-coding genes, and intergenic and genic piRNA clusters.

(B) Meta plots depict DNA methylation levels at promoters of active protein-coding genes, silenced protein-coding genes, and intergenic and genic piRNA clusters. The results are the same as (A).

(C) Boxplots demonstrate DNA methylation levels in the promoters of Ko.ERV.1, Ko.ERVL.1, and Ko.ERVK.14 insertions. Many of these ERV insertion promoters are both DNA hypo-methylated in the testes of Sunshine Coast and Currumbin koalas.

(D) Same as (C), but only showing those insertions shared between the Sunshine Coast koala and the Currumbin koala. No significant difference in DNA methylation levels is observed between Sunshine Coast and Currumbin koalas.

(E) An example illustrates the DNA methylation level at one of the shared full-length KoRV-A proviruses in Sunshine Coast koala testis and Currumbin koala testis. Promoters are marked in gray rectangles and zoomed in at the bottom panel.

(F) The same as (A), but for the other shared full-length KoRV-A insertions in Sunshine Coast koala testis and Currumbin koala testis.

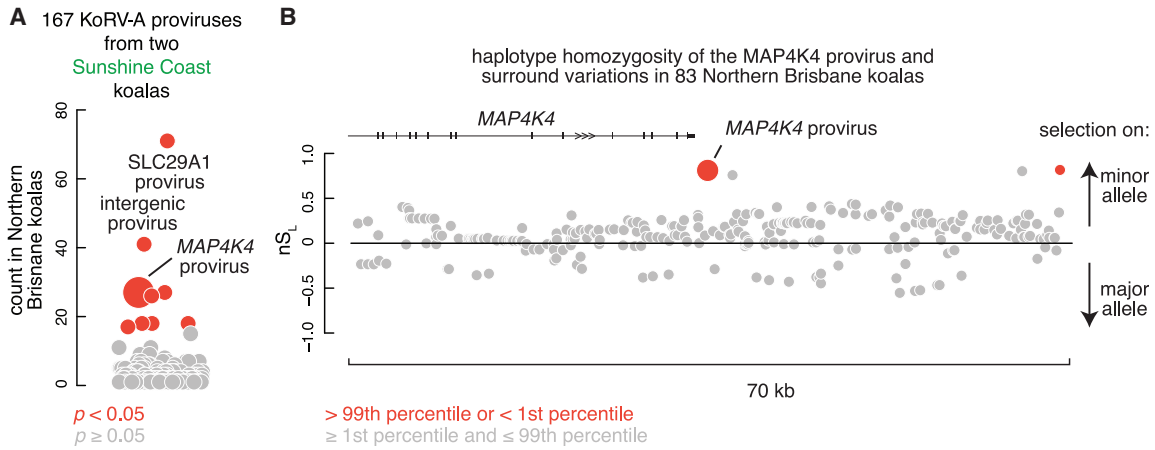



**Figure S5. The *MAP4K4* KoRV-A proviruses are likely under positive selection, related to Figure 4**

(A) A plot depicts the number of northern Brisbane River koalas sharing the 167 KoRV-A proviruses identified in the two male Sunshine Coast koalas. Proviruses significantly shared by the Brisbane River koalas (Z test $p$ value < 0.05) are marked in red.

(B) $nS_L$ for polymorphic sites surrounding the *MAP4K4* KoRV-A provirus at a 70 kbp window. Positive values represent selection on minor alleles and negative values on major alleles. Outliers (>99th or <1st percentile) are colored in red. The *MAP4K4* provirus is marked in the center.

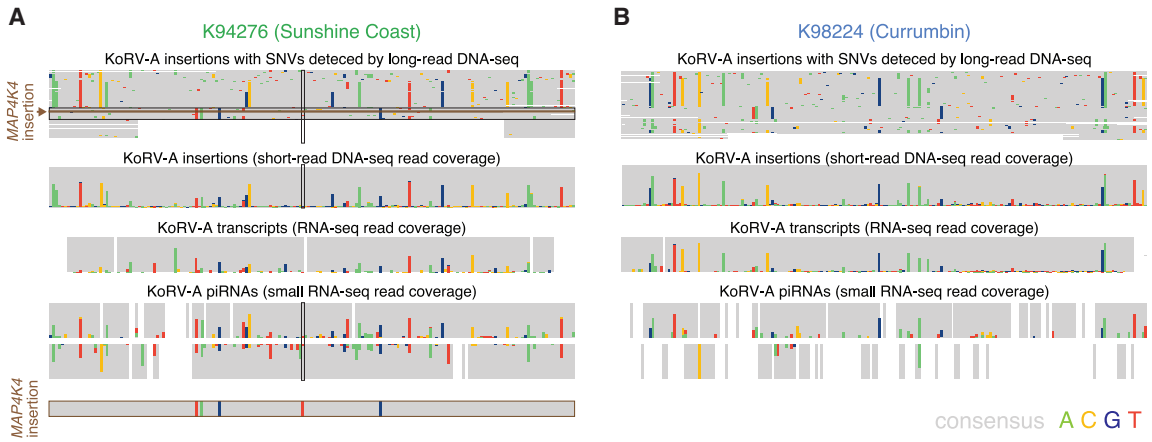



**Figure S6. Polymorphism analysis of KoRV-A in the two koalas, related to Figure 6**

(A and B) Comparison of the genome, transcripts, and piRNAs of KoRV-A polymorphism sites in one Sunshine Coast koala (A) and one Currumbin koala (B). The top panel displays all polymorphic alleles of KoRV-A insertions assembled by Nanopore DNA-seq. The subsequent panels illustrate the polymorphism of the genome, sense transcripts, and sense and antisense piRNAs, respectively. Alternative alleles are denoted in different colors (N > T as red, N > G as blue, N > A as green, and N > C as yellow), indicating polymorphic variations. Both transcripts and piRNAs present various alleles as the *MAP4K4* provirus and other proviruses, suggesting that the *MAP4K4* provirus and additional proviruses contribute to the transcript and piRNA pool of KoRV-A.

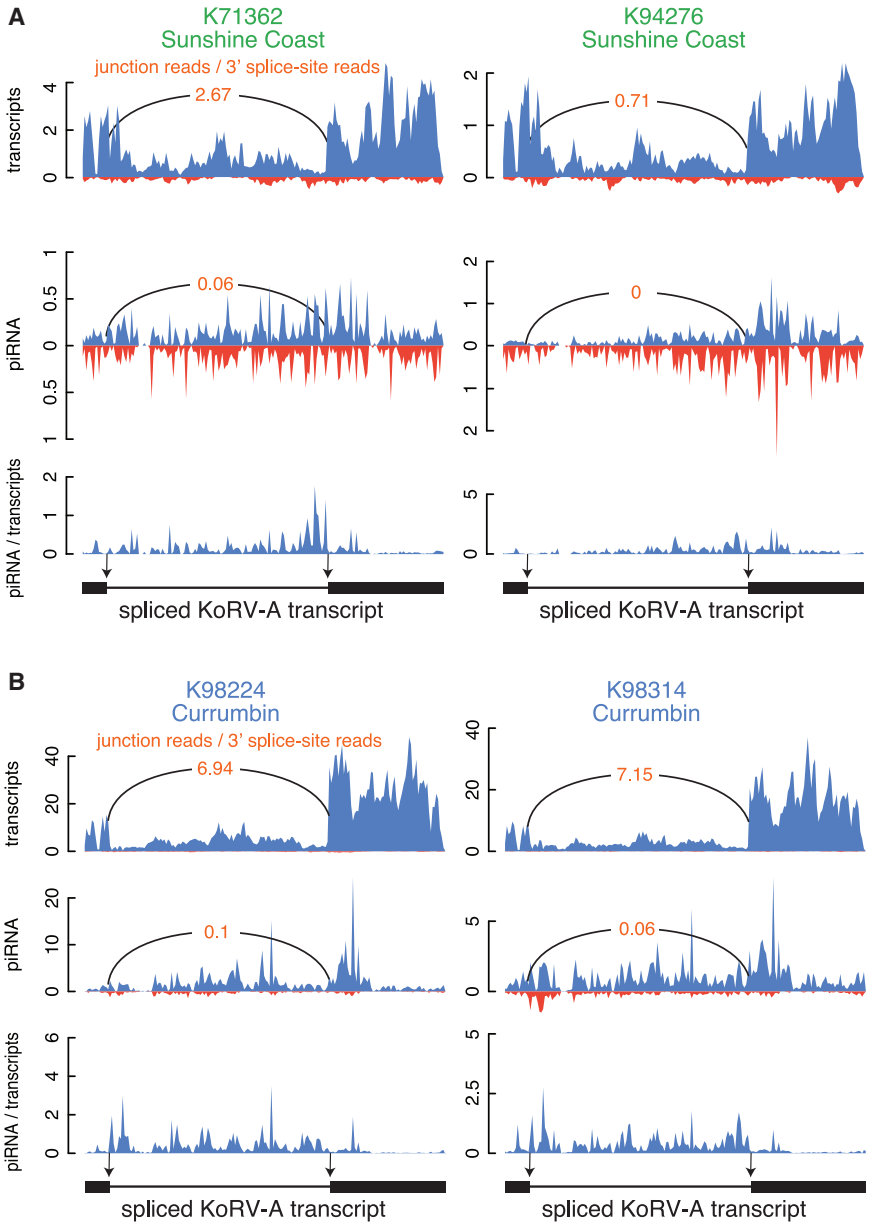



**Figure S7. RNA and piRNA profiles of KoRV-A in the testes of Currumbin and Sunshine Coast koalas, related to STAR Methods**

(A) Transcription (top), piRNA abundance (middle), and piRNA processing efficiency (bottom; defined by piRNA/transcript) for KoRV-A in two male Sunshine Coast koalas are depicted. The splicing indexes for the KoRV-A splice junction detected by RNA-seq and small RNA-seq are shown.

(B) Similar analysis is presented for two male Currumbin koalas sequenced in this study.