Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion

Ulrike Löber[a,b,1], Matthew Hobbs[c,1], Anisha Dayaram[a], Kyriakos Tsangaras[d], Kiersten Jones[e], David E. Alquezar-Planas[a,c], Yasuko Ishida[f], Joanne Meers[e], Jens Mayer[g], Claudia Quedenau[h], Wei Chen[h,i], Rebecca N. Johnson[c], Peter Timms[j], Paul R. Young[e], Alfred L. Roca[f,2], and Alex D. Greenwood[a,k,2]

[a]Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany; [b]Berlin Center for Genomics in Biodiversity Research (BeGenDiv), 14195 Berlin, Germany; [c]Australian Museum Research Institute, Australian Museum, Sydney, NSW 2010, Australia; [d]Department of Translational Genetics, The Cyprus Institute of Neurology and Genetics, Nicosia 1683, Cyprus; [e]Australian Infectious Diseases Research Centre, The University of Queensland, St. Lucia, QLD 4067, Australia; [f]Department of Animal Sciences, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [g]Department of Human Genetics, Medical Faculty, University of Saarland, 66421 Homburg, Germany; [h]Max Delbruck Center, The Berlin Institute for Medical Systems Biology, Genomics, 13125 Berlin, Germany; [i]Department of Biology, Southern University of Science and Technology, Shenzhen, Guangdong, China 518055; [j]Faculty of Science, Health, Education & Engineering, University of the Sunshine Coast, Sippy Downs, QLD 4556, Australia; and [k]Department of Veterinary Medicine, Freie Universität Berlin, 14163 Berlin, Germany

Edited by Stephen P. Goff, Columbia University Medical Center, New York, NY, and approved July 2, 2018 (received for review May 4, 2018)

Endogenous retroviruses (ERVs) are proviral sequences that result from colonization of the host germ line by exogenous retroviruses. The majority of ERVs represent defective retroviral copies. However, for most ERVs, endogenization occurred millions of years ago, obscuring the stages by which ERVs become defective and the changes in both virus and host important to the process. The koala retrovirus, KoRV, only recently began invading the germ line of the koala (*Phascolarctos cinereus*), permitting analysis of retroviral endogenization on a prospective basis. Here, we report that recombination with host genomic elements disrupts retroviruses during the earliest stages of germ-line invasion. One type of recombinant, designated recKoRV1, was formed by recombination of KoRV with an older degraded retroelement. Many genomic copies of recKoRV1 were detected across koalas. The prevalence of recKoRV1 was higher in northern than in southern Australian koalas, as is the case for KoRV, with differences in recKoRV1 prevalence, but not KoRV prevalence, between inland and coastal New South Wales. At least 15 additional different recombination events between KoRV and the older endogenous retroelement generated distinct recKoRVs with different geographic distributions. All of the identified recombinant viruses appear to have arisen independently and have highly disrupted ORFs, which suggests that recombination with existing degraded endogenous retroelements may be a means by which replication-competent ERVs that enter the germ line are degraded.

koala retrovirus | recombination | retrovirus | endogenous retrovirus | genome evolution

In humans, about 8% of the genome consists of endogenous retrovirus (ERV)-like elements, comprising a larger proportion of the genome in humans and other species than protein coding regions within genes (1–3). The architecture of the human genome reflects a long evolutionary history of invasions of the germ line by infectious retroviruses (1, 2, 4–7). Phylogenetic analyses suggest that retroviruses have, from a deep evolutionary perspective, frequently jumped from one species to another and invaded the germ lines of new hosts (8–12). Almost all known ERVs completed invasion of their host germ lines millions of years ago, obscuring the early events critical to the invasion process. An exception is the koala retrovirus (KoRV).

KoRV is a full-length replication-competent endogenous retrovirus, the titer of which is correlated with chlamydiosis and hematopoietic neoplasia (13, 14). KoRV is thought to spread in koalas both horizontally by infection and vertically as an endogenous genomic element (15, 16). KoRV has a clinal geographic distribution among koalas; while 100% of northern Australian koala populations carry KoRV, the prevalence and copy number of KoRV is greatly reduced in southern Australia (14, 17, 18). Unlike other ERVs, KoRV is not present in the germ line of all members of the host species (19). Ancient DNA studies have shown that KoRV was ubiquitous in Queensland koalas in the late 19th century (16). Molecular dating places the initial entry of KoRV into the koala germ line within the past 50,000 y (16, 20). These studies strongly indicate that KoRV, unlike most known vertebrate ERVs, is in the early stages of the endogenization process in its koala host (17, 19).

Many ERVs in vertebrates are found at fixed positions in the genome across all members of the host species. By contrast, there is substantial variation in the host genomic integration sites for endogenous KoRV across koalas, with a very high degree of insertional polymorphism. Among modern and museum samples of koalas in Queensland, the population with the highest

Significance

Endogenous retroviruses (ERVs) are proviral sequences that result from host germ-line invasion by exogenous retroviruses. The majority of ERVs are degraded. Using the koala retrovirus (KoRV) as a model system, we demonstrate that recombination with an ancient koala retroelement disables KoRV, and that recombination occurs frequently and early in the invasion process. Recombinant KoRVs (recKoRVs) are then able to proliferate in the koala germ line. This may in part explain the generally degraded nature of ERVs in vertebrate genomes and suggests that degradation via recombination is one of the earliest processes shaping retroviral genomic invasions.

Author contributions: U.L., A.L.R., and A.D.G. designed research; A.D., K.T., K.J., D.E.A.-P., Y.I., and C.Q. performed research; J. Meers, W.C., R.N.J., and P.T. contributed new reagents/analytic tools; U.L., M.H., K.T., J. Mayer, P.R.Y., and A.D.G. analyzed data; and U.L., M.H., D.E.A.-P., A.L.R., and A.D.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra/; accession no. SRS2321692).

[1]U.L. and M.H. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: roca@illinois.edu or greenwood@izw-berlin.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1807598115/-/DCSupplemental.

abundance of KoRV, only a small proportion of KoRV integrations are shared among unrelated koalas (20–22). The lack of fixed KoRV proviral insertions among koalas, and the low proportion of shared integration sites among koalas that carry endogenous KoRV, provide further evidence for a recent invasion of the koala germ line by KoRV.

Most endogenous retroviruses in other species have highly disrupted proviral genomes, and those present in higher copy numbers across the germ line often show deletion of the proviral *env*, which codes for the viral envelope. ERVs that lack *env* have been found to be "superspreaders," i.e., elements that have reached a high copy number within the host germ line (23). Through recombination, exogenous retroviruses can exchange genetic information with ERVs in infected individuals. For example, murine leukemia virus (MLV) can recombine with endogenous MLVs to generate novel viruses, in effect remobilizing part of the ERV sequences (24). Recombination may also render ERVs defective, e.g., through disruption of ORFs or the generation of solo long terminal repeats (LTRs) (25). However, the role of recombination during the early stages of retroviral genomic invasion has not been directly examined. Here, we provide evidence that older endogenous retroelements recombine with and degrade invading retroviral genomes, even when the homology between them is limited. This occurs early during retroviral invasion and disrupts the invading retrovirus while simultaneously remobilizing an existing retroelement recombination partner within the host genome. By disrupting retroviruses invading the germ line, the process likely accelerates the retroviral transition from horizontal to vertical transmission, which is expected to benefit the host species.

Results

The Advantages of Long Read Sequence Technology for Retroviral Analysis. KoRV has generally been studied using 454 FLX or Illumina-based short read sequencing approaches (16, 22). These approaches have a number of limitations. First, identified polymorphisms in KoRV cannot generally be put in phase with other polymorphisms. Second, only small structural differences among KoRV sequences, such as short indels, can be detected. Large deletions and recombination events are missed given that reads are of short length. Third, the specific host integration site cannot be identified for polymorphisms or KoRV sequences because reads are not long enough to cover the DNA region from the integration site to the KoRV genes. Thus, KoRV variation has been studied in the aggregate by mapping reads to full-length proviruses. Using this method, little variation has been detected (16, 22). By contrast, the current study used PacBio technology, which produces long sequence reads. The koala genome was sequenced using this technology, and we here sequenced individual KoRV proviruses. This permitted us to identify structural variation across individual KoRV proviruses, link KoRV variants to genomic loci in the koala host, and determine the position and copy number for each type of variation detected among KoRV proviruses. A complex evolutionary history was revealed for KoRV.

recKoRV1, a Recombinant Koala Retrovirus. PacBio sequencing of one koala (Bilbo; Table 1 and ref. 26), the individual used to sequence and assemble the koala genome, has identified a proviral integrant called recombinant koala retrovirus 1 (recKoRV1), which includes the 5′ KoRV LTR followed by the *gag* leader region to position 1,177 (27). It also includes the KoRV region from position 7,619 of the *env* gene including the complete 3′ LTR. However, the sequence between these two fragments of KoRV is derived from another retroelement, designated the *Phascolarctos* endogenous retroelement (PhER) (Fig. 1) (27). PhER has partial homology to Repbase (28) but has no intact protein coding regions except potentially in the *env* region (27, 29). PhER has been found to be a transcriptionally expressed high copy number ERV (~30–40 full-length elements and hundreds of solo LTRs or fragmented copies).

Other Recombinants Between KoRV and PhER. In the current study, we examined an unrelated koala (Bilyarra; Table 1) to characterize recKoRV. We also identified KoRV-PhER recombination breakpoints and used them as queries to screen existing Illumina sequence datasets that had been previously generated but never examined for KoRV recombinants. Proviral integration sharing among koalas was examined on a per locus basis, while the presence or absence of specific recombination breakpoints was examined in the aggregate (Table 1 gives details regarding the koalas and datasets). Along with the two recombination sites of recKoRV1, an additional 15 KoRV-PhER recombinant sequences were identified (Fig. 1 and *SI Appendix*, Table S1). Although KoRV and PhER had dissimilar sequences, at five of the recombination breakpoints we identified microhomologies, short matching sequences that were shared at a breakpoint by both KoRV and PhER. These microhomologies may have enabled the recombination between the two elements at various breakpoints (*SI Appendix*, Table S1) (30–32). Of 17 recombination breakpoints identified, all but three were within 1,500 bp of the ends of the KoRV genome (Fig. 1). Most breakpoint sequences were determined using only short Illumina reads, and so it was not possible to determine the structure of recombinants or characterize the integration sites.

The number of Illumina sequences among koala datasets mapping to proviral recKoRV1 breakpoints far exceeded those mapping to any of the other recombination breakpoints identified (*SI*

Table 1. Koala samples and datasets utilized

Koala	Wild/zoo*	Sample sources	Sequence type	Database	Source
Bilyarra	SN241	Vienna Zoo (Tierpark Schönbrunn)	Long inverse PCR PacBio	SRS2321692	This paper
Bilbo	Wild	Australian Museum registration M.47724, Upper Brookfield Queensland	PacBio genome assembly	GCA_002099425.1	26, 27
Pacific Chocolate	Wild	Port Macquarie, New South Wales	Illumina sequences		26, 27
Birke	Wild	Australian Zoo Wildlife Hospital in Queensland	Illumina sequences		26, 27
One zoo and 6 museum specimens	SN265 and historical wild	Vienna Zoo (Tierpark Schönbrunn) and various museums	Hybridization capture Illumina sequences	KF786285 KF786284	22
Samples of 166 koalas	Wild	Collected across koala range in Australia	PCR and Sanger sequencing		17

*SN indicates the European koala studbook number for samples from zoological collections.

The 166 wild koalas in ref. 17 were sampled from across their geographic range. Pacific Chocolate was from New South Wales. All other samples were derived from the Queensland koala population, including all zoo koalas and museum specimens. Database refers to National Center for Biotechnology Information GenBank and Sequence Read Archive.

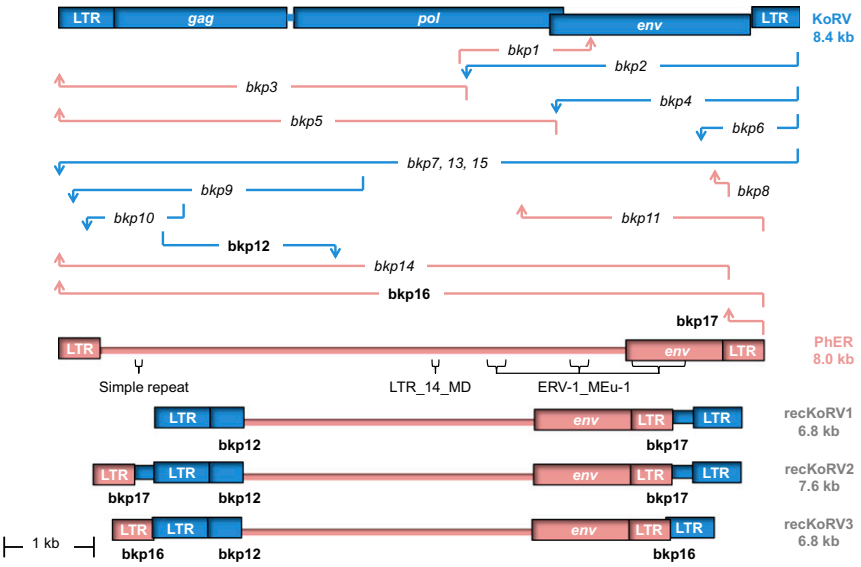

Fig. 1. Breakpoints in KoRV-PhER recombinants. The genomic structures of KoRV (blue) and PhER (pink) (27) are shown, including genes, LTRs, and Repbase repeat motifs identified in PhER. Locations of breakpoints (bkps) in 17 recombinant sequences (detailed in *SI Appendix*, Table S1) are represented by arrows, with pink upward-directed arrows used when PhER sequence is 5′ of the breakpoint, and blue downward directed arrows when KoRV sequence is 5′ of the breakpoint. For bkps within an LTR sequence, only one of the possible alignments is shown. Three recombinant sequences from long read (Pacbio) sequence datasets allowed assignment of breakpoints to recombinant elements recKoRV1, recKoRV2, and recKoRV3. Breakpoints identified only in short read (Illumina) sequence datasets are italicized.

Appendix, Fig. S1 and Table S1) (26, 27). We therefore focused on the evolutionary history of the recKoRV1 subtype of recombinants.

Absence of Reciprocal Recombinant recKoRVs. We screened for reciprocal recombination products relative to the structure of recKoRV1, i.e., containing PhER sequences flanking KoRV coding sequences, and found no evidence for them across the genome of Bilbo. This was not unexpected because viral integrases are generally LTR sequence specific and sequence alignment using blastn with tolerant/permissive parameter settings revealed no substantial sequence similarity in the LTR regions of KoRV and PhER. Because PhER does not code for an intact integrase, both KoRV and recKoRV1 would rely on KoRV integrase to insert into the genome. PhER-flanked reciprocal recombinants would likely lack the requisite LTR sequences to be recognized and integrated efficiently (33).

Comparison of LTRs and Integration Sites Among Koala Genomes. KoRV integration sites are highly insertionally polymorphic across unrelated koalas (19, 20). We examined KoRV and recKoRV1 integration sites in Bilyarra using a long read inverse PCR strategy and PacBio sequencing. This method allowed for identifying KoRV and recKoRV1 sequences and their integration sites in long single PacBio reads. Bilyarra exhibited a greater number of KoRV integration sites than found in the Bilbo reference genome (66 compared with 58). Among the KoRV integration sites (unique to each proviral locus) in Bilbo and Bilyarra, only two were shared (KoRV22 and KoRV35; *SI Appendix*, Fig. S1*B* and Table S2). In each of the two KoRV proviral loci shared by Bilbo and Bilyarra, deletions detected in the *env* gene would likely have precluded production of infectious virions. The other 120 integration sites were only detected in one of the two koalas. This suggests that individual KoRV integrants are found at low frequencies in their respective chromosomes and not generally shared by unrelated koalas. By contrast, many LTR sequences from Bilbo, Bilyarra, and other koalas were identical; they largely overlapped across a minimum spanning network, with few sequences unique to a specific koala (*SI Appendix*, Fig. S2). This indicates that KoRV proviruses at different loci have the same LTR sequence (20), as many LTR sequences, unlike integration sites, were shared among unrelated koalas.

Twenty-four recKoRV integrations were identified in Bilyarra, of which 14 were characterized as recKoRV1. In Bilbo 12 recKoRV1 sites were identified (identified through PacBio sequence reads that included both the integration sites and one or both of the recKoRV1 breakpoints). None of the recKoRV1 integration sites was shared between Bilbo and Bilyarra (*SI Appendix*, Fig. S1*B*). This may indicate that recKoRV1 integrations have not had sufficient time to become broadly distributed among koalas. The absence of shared loci carrying recKoRV1 between Bilbo and Bilyarra would suggest that recKoRV1 has been able to retrotranspose to different loci in the koala genome and/or that the same recombination event has occurred between KoRV and PhER on more than one occasion. The recKoRV LTRs varied across recKoRV1 loci, were often identical to KoRV LTRs, and included four of the five most common KoRV LTRs (*SI Appendix*, Fig. S2). This suggests that the same breakpoints between KoRV and PhER have been used in independent recombination events to generate recKoRV1s multiple times, because random mutations from a single ancestral recKoRV1 LTR would not exactly match those that happen to distinguish the most common KoRV LTR sequences. Target site duplications of 4–10 bp were detected at the integration sites of KoRV and recKoRV1 in the large majority of cases (*SI Appendix*, Table S2), suggesting that the integrations involved a retrovirus-typical reverse transcription as opposed to meiotic recombination.

The 17 Identified Recombination Breakpoints, Including recKoRV1 Breakpoints, Have Dissimilar Distributions Among Koala Populations. Unlike most ERVs, KoRV greatly varies in prevalence across its host populations. While all Queensland koalas are positive for KoRV with high copy numbers in their genomes, southern Australian koalas have a much lower prevalence and copy number, with KoRV completely absent from some individuals (34). Sequences for all 17 recombination breakpoints identified between KoRV and PhER were used to query the koala reference genome and Illumina sequence datasets. Of the 17, 11 were identified in the genome of Pacific Chocolate, a koala from New South Wales, but were absent from the genome of Birkie (Table 1) from Queensland. The other six recombination breakpoints, including the breakpoints of recKoRV1, were identified in Birkie but absent from Pacific Chocolate. The lack of overlap may suggest that independent recombination events between KoRV and PhER have occurred in koalas from the two Australian regions. Screening of sequence datasets that had been generated after hybridization capture of KoRV identified the two recKoRV1 recombination breakpoints in all other koalas examined, including both museum and modern samples (*SI Appendix*, Fig. S1 and Table S1). Five of the 11 breakpoints in Pacific

EVOLUTION

Chocolate from New South Wales were specific to that individual. The remaining six breakpoints were detected only sporadically among existing Illumina datasets, with the exception of breakpoint 10, which was found in most museum and a zoo koala but not in Birkie.

An Extended Analysis of the Geographic Distribution of recKoRV1. To more precisely characterize the geographic distribution of recKoRV1 among koalas, the presence or absence of the 3′ recKoRV1 recombination breakpoint was examined using PCR. To span the recombination breakpoint, the 5′ PCR primer matched the upstream PhER sequence and the 3′ primer matched the downstream KoRV LTR sequence. We screened for the 3′ recKoRV1 recombination breakpoint in 166 koalas from 11 populations across Australia that had previously been screened for KoRV prevalence (17). KoRV and the recKoRV1 3′ breakpoint were both present across all koalas in Queensland and inland New South Wales (Fig. 2) with the notable exception of St. Bees Island (Fig. 2, population B) in which the recKoRV1 3′ breakpoint was only detected in 4 of 15 koalas, although KoRV was ubiquitous among St. Bees koalas. The coastal population of Port Stephens in New South Wales (Fig. 2 population G) was 100% KoRV positive but devoid of recKoRV1. This is consistent with the absence of recKoRV1 recombination breakpoints in the genome of Pacific Chocolate (from nearby Port Macquarie, New South Wales). Further south in Victoria, both Mornington Peninsula and Gippsland were negative for the recKoRV1 breakpoint and either positive for KoRV (Gippsland; Fig. 2, population K) or negative for both KoRV and recKoRV1 (Mornington Peninsula; Fig. 2, population J).

In South Australia, 8 of 16 koalas were positive for KoRV in Adelaide Hills (Fig. 2, population H), and 15 of 32 were KoRV positive in Kangaroo Island (Fig. 2, population I). No koalas were recKoRV1 positive in the Kangaroo Island population (founded by koala translocation in the 19th century). In the Adelaide Hills, four of eight KoRV-positive koalas were also positive for recKoRV1. Two of the eight KoRV-negative koalas were recKoRV1 positive in Adelaide Hills, the only koalas in the dataset to show this pattern.

Discussion

Degradation of ERV genomes, and loss of *env* in particular, may benefit the host by preventing the production of virulent retroviruses that can spread horizontally (23). Our findings suggest that the recombination mediated degradation of retroviruses, which has been postulated for many human and other vertebrate ERVs, and the genomic proliferation of recombinants both occur at the earliest stages of retroviral germ-line invasion (35–38). This is supported by the presence of recKoRV1s in koalas across almost all of Australia in both modern and historical samples, and their high copy number in the koala genomes examined (Figs. 1 and 2). KoRV is thought to have invaded the koala germ line relatively recently, within the last 50,000 y (20). Thus, within this time frame the recKoRVs were generated and recKoRV1s arrived at the widespread distribution revealed here.

Seventeen recombination breakpoints were detected between KoRV and PhER. Recombination occurred in some cases at microhomologies, short sequences common to the two retroelements that likely enabled recombination at many of the breakpoints, including those of recKoRV1. Transcripts of PhER have been detected in the koala transcriptome, suggesting that PhER could be copackaged with KoRV in the same virion (29), enabling recombination. KoRV integrants may also have recombined with retrotranscribed PhER during meiosis. In both KoRV and recKoRV1, target site duplications were generally detected in the host genome flanking the 5′ and 3′ ends of the provirus, indicating that recombination between retroelements at different loci had not affected these integrants. Such recombinants, which can delete large regions of the genome, may have been removed by selection. Once a recKoRV is established in the germ line, it can spread vertically (and geographically) across koala populations.

A high degree of population structuring was detected among the different recombination breakpoints between KoRV and PhER. In particular, the recKoRV1 3′ breakpoint was completely absent from some populations in New South Wales (Fig. 2), while the genomes of two koalas, one from Queensland and one from coastal New South Wales, differed dramatically in their complement of recombination breakpoints (*SI Appendix*, Fig. S1). Genetic differentiation between Queensland and New South Wales koalas has been reported in previous studies (39, 40),

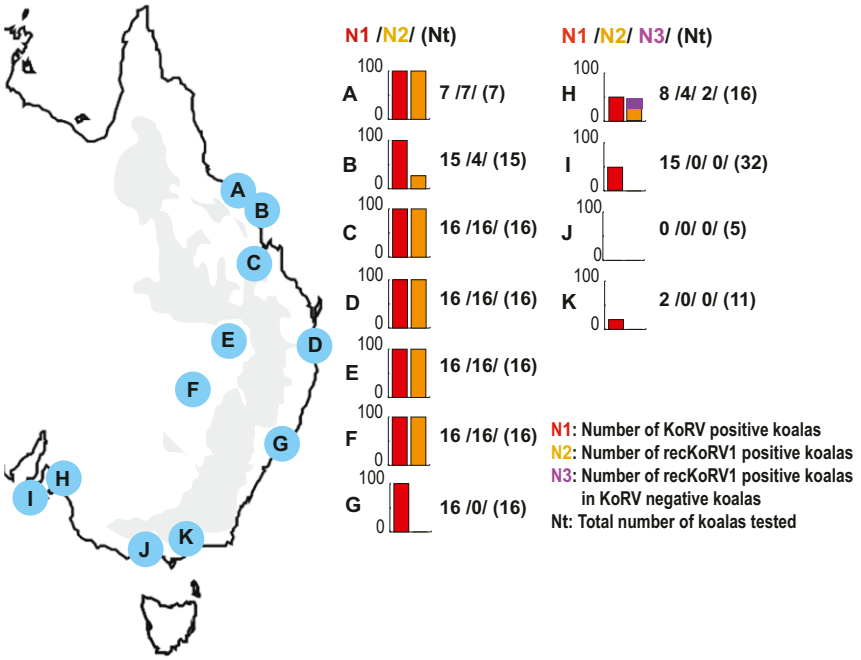

Fig. 2. Prevalence of recKoRV1 in KoRV-positive and KoRV-negative koalas across Australia. The proportion of recKoRV1-positive koalas in both KoRV-positive and KoRV-negative koalas was determined by PCR assay. The percent of KoRV-positive and KoRV-negative koalas with or without recKoRV1 is shown for each population in the bar charts. The numbers to the right of each chart indicate the number of koalas in each respective category (N1, N2, and N3). Nt (in parentheses) refers to the total number of koalas tested at each locality. Red bars on the graphs indicate the percent of koalas that were KoRV positive, orange indicates the percent recKoRV1 positive, and purple indicates the percent of koalas recKoRV1 positive but KoRV negative. The Great Dividing Range is indicated on the map in gray. The localities sampled were as follows: A, Hamilton Island, Queensland (QLD); B, St Bees Island, QLD; C, Central QLD; D, Currumbin Wildlife Sanctuary, QLD; E, South-West QLD, F, West Pilliga, New South Wales (NSW); G, Port Stephens, NSW; H, Adelaide Hills, South Australia (SA); I, Kangaroo Island, SA; J, Mornington Peninsula, Victoria (VIC); K, Gippsland, VIC.

suggesting restricted gene flow between koalas in the two states, perhaps in part due to the Great Dividing Range (Fig. 2). The barrier to gene flow cannot be complete because KoRV is present at high frequency in all of New South Wales and was thus likely transferred from koalas in Queensland at some point. Additionally, koala populations do not show high degrees of genetic structure compared with other marsupials, although recent barriers to gene flow may exist particularly in New South Wales (41, 42). We cannot rule out the possibility that regional differences in PhER expression may affect the genesis or distribution of recKoRVs by altering the amount or type of PhER template available for recombination. However, it is also possible that PhER and KoRV may be expressed and recombine in any population where both are present. This is supported by the analysis of KoRV-PhER recombination breakpoints in the genomes of a koala from Queensland and a koala from New South Wales. The two carried completely distinct sets of recombination breakpoints (*SI Appendix*, Fig. S1), suggesting that recombinants between KoRV and PhER formed independently in the two populations.

Several populations showed atypical patterns in the distribution of recKoRV1. In St. Bees Island off Queensland, only 4 of 15 koalas were recKoRV1 positive, but all were KoRV positive (Fig. 2). This contrasts with mainland Queensland for which all koalas tested ($n = 48$) were positive for both recKoRV1 and KoRV. The St. Bees Island population was founded by translocation of 12–17 koalas from mainland Queensland in the 1930s (43). The founding population of St. Bees was small, likely with insertional polymorphisms in each recKoRV1 locus. After the population expanded, the koalas would reflect random combinations of the small numbers of founder chromosomes. It may be that loci carrying recKoRV1 were randomly lost through genetic drift, although it is also possible that selection may have played a role.

In the Adelaide Hills of South Australia, several KoRV-negative individuals proved to be recKoRV1 positive (Fig. 2). KoRV copy number has been shown to decrease dramatically in southern Australia based on qPCR targeting the *pol* gene, and KoRVs rarely exist in both chromosomes in a given koala individual even where copy numbers are high (17, 20). The recKoRV1-positive individuals lacking KoRV likely reflect Mendelian segregation of integrants in a population where both KoRV and recKoRV1s are present at low copy numbers and at low frequencies at their respective loci, so that only a limited proportion of individuals carry either or both.

KoRV would suffer loss of virulence after recombination with PhER because none of the recombinants are predicted to code for an intact virus. Existing genomic elements like PhER would proliferate by having parts of their sequences incorporated into recKoRVs. While recKoRVs could still potentially exert deleterious effects on the host, e.g., by retrotransposition into new genomic locations, other potentially deleterious effects of the provirus would be reduced relative to intact KoRV, notably the ability of these elements to produce infectious retrovirus. The switch to a proviral form that is disrupted by recombination may be one aspect of the transition from horizontal to vertical transmission among ERVs. Over time, this would be expected to result in an increase in recKoRV abundance at the expense of virulent KoRV proviruses, potentially reducing the impact of the latter. The pressure to make this transition may be higher in long-lived species that are more likely to be affected by ERVs with oncogenic potential (44). During the transition period when infectious KoRV and recombinants coexist, KoRV particles may de novo generate and horizontally transmit recKoRVs (*SI Appendix*, Fig. S2) and in this manner coinfect host cells, although superinfection resistance would likely limit novel infection-mediated proliferation of both KoRV and recKoRVs (45).

In detecting large numbers of recombinants between KoRV and PhER, we establish that recombination with existing retroelements may be one way in which the ability of retroviruses invading the germ line to faithfully replicate is disrupted, by removing their ability to encode active viruses associated with disease. This would not be the only mechanism by which a host species controls an invading ERV, since other factors are likely to play a role, such as methylation or antiretroviral proteins or disruptive within-KoRV recombination (as was evident for KoRV22 and KoRV35, the only shared KoRV integrants identifed) (46). Nor would recKoRV lack potentially deleterious aspects of a provirus, as activation or disruption of genes at or near insertion sites may still occur. However, the deleterious effects of recKoRVs are not likely to be as great as those of KoRVs, and recKoRVs may thus be less subject to purifying selection than replication competent KoRVs, allowing recKoRVs to persist in the host germ line. Several lines of evidence suggest that production of recKoRVs may reflect a general means of accommodation between ERV and host. The recKoRV proviruses would have a reduced ability to proliferate relative to intact KoRV. The process of recKoRV formation has occurred frequently and independently, given the many recKoRVs identified and geographic differences in the occurrence of breakpoints. The degraded nature of recKoRV1 is also consistent with inferences drawn from more ancient ERVs in vertebrate genomes, notably the concept of genomic superspreaders, which suggests that retroviruses that lose the *env* gene will be more successful at propagating in host genomes than intact ERVs (23). It is also consistent with the exchange of sequences between divergent retroviruses, which has been inferred for ERVs in various host species (35–38, 47). For example, some human ERVs are believed to have recombined before their proliferation (38). This suggests that recombination-based degradation has occurred during invasions of vertebrate germ lines by different groups of retroviruses. Our study demonstrates empirically that the generation of such recombinants occurs during the early stages of genomic invasion by ERVs of a host germ line.

Materials and Methods

Koala Samples, PCR, and Sequencing. Four sources of genomic data were employed in the current study (additional details on the samples and datasets are provided in Table 1 and *SI Appendix*). We used Illumina-based genome sequences (unassembled) from two koalas, Pacific Chocolate and Birkie (Table 1) (26). This dataset was screened for KoRV and PhER breakpoints. Additionally, existing Illumina datasets were reexamined for KoRV and PhER breakpoints (Table 1 and *SI Appendix, SI Materials and Methods*) (22). Two PacBio-based datasets were used to investigate KoRV and PhER breakpoints, and to identify and characterize proviral integration sites. The first of these was the assembled genome of Bilbo (26), and the second consisted of the integration site-enriched PacBio sequences from Pci-SN241 (Bilyarra, described below) (Table 1). DNA extraction and analyses of the data are described in detail in the *SI Appendix*. Finally, 166 wild koala DNA samples were collected by J. Meers, and P.Y. and their associates (17). The DNA from the 166 samples was extracted using the Blood & Tissue DNA Extraction Kit (Qiagen) or was provided by collaborators. The DNA was amplified for the recoKoRV 3′ breakpoint (with the koala *actin* gene used as a positive control for DNA quality). The amplified recKoRV PCR products were Sanger sequenced to establish their identity as the recKoRV1 3′ breakpoint. The *SI Appendix* provides details on the PCR and sequencing.

Inverse PCR and PacBio Sequencing of recKoRV. DNA from Bilyarra was fragmented and circularized in four steps (detailed in *SI Appendix*). These were as follows: (*i*) DNA fragmentation, (*ii*) fragment end repair and circularization, (*iii*) KoRV LTR-based LTR long inverse PCR, and (*iv*) PacBio sequencing of the products. PCR products were submitted for PacBio library construction and sequencing to the Max Delbrück Center, Berlin, and standard PacBio RSII sequencing was performed (details provided in *SI Appendix*). Bioinformatics analyses were conducted both to identify Bilyarra's integration sites and to identify whether the integrants were KoRV or recKoRV (*SI Appendix*). Bilyarra PacBio sequences were aligned to a custom KoRV database. Blastn (NCBI Nucleotide-Nucleotide BLAST 2.2.29+) was used with default options to generate the alignments (48). Reads that did not include regions homologous to KoRV were considered KoRV-negative reads. KoRV-positive reads were aligned to the full-length PhER sequence using blastn with default parameters. PhER-negative reads were considered KoRV sequences and not recKoRV sequences. PhER-positive reads were initially considered to be sequences potentially containing recKoRV1. The 5′

breakpoint was generally more poorly covered by PacBio sequences because the breakpoint is several kilobases from the start of the proviral genome, compared with the 3′ breakpoint that is 200 bp from the end of the 3′ LTR (*SI Appendix*, Fig. S3). Description of the procedures for individual locus confirmation by PCR and Sanger sequencing is in *SI Appendix*.

To isolate the host genomic sequences flanking integration sites for KoRV and recKoRV, the KoRV containing reads were aligned using blastn to the KoRV-A or KoRV-B reference sequences (AB721500.1; KC779547). Regions homologous to the reference sequences were removed. The isolated host genomic sequences flanking integrations sites were clustered using Tribe-MCL (I = 1.4) (49), a Markov cluster-based approach, processing distance-based information of a blastn matrix for all KoRV containing reads (49). The recKoRV1-containing reads were aligned using blastn to the KoRV-A and KoRV-B reference sequences, as well as to PhER, all known recKoRV breakpoints and the consensus sequence of recKoRV1. Regions homologous to any of the reference sequences were removed. The isolated flanking regions were clustered using Tribe-MCL (I = 4). A consensus sequence for every cluster was created by constructing a multiple sequence alignment using MAFFT (v7.305b) (50) and computing a consensus sequence using the Perl module BioPerl:SimpleAlign (30% identity, gap removal) (51). Further curation is detailed in *SI Appendix*.

Network Analysis of LTRs. LTRs from all insertion sites of Bilbo and Bilyarra were aligned with the LTR sequences of the KoRV proviruses examined in ref. 20, along with KoRV-A (AB721500.1) and KoRV-B (KC779547.1). The iPCR primer gaps were removed from all sequences. Multiple sequence alignment was performed using MAFFT L-INS-i (50). The alignment was cropped to the most conserved regions (>89% identity) on both ends, realigned, and manually curated. A haplotype network was constructed using the R (52) package Pegas (53) with the distance model "indelblock," performing an iterative refinement for the smallest sum of distances.

ACKNOWLEDGMENTS. We thank Baptiste Mulot (Zoo Beauval) and Hanna Vielgrader (Tiergarten Schönbrunn) for assistance with zoo koala information throughout the project and thank the sample providers listed in *SI Appendix*. R.N.J. thanks the Australian Museum Foundation, BioPlatforms Australia, and New South Wales Environmental Trust for support. Y.I., A.L.R., and A.D.G. were supported by Grant R01GM092706 from the National Institute of General Medical Sciences (NIGMS). D.E.A.P. and A.D.G. were supported by the Morris Animal Foundation, Grant D14ZO-94. D.E.A.P. was supported by a postdoctoral fellowship of the Deutscher Akademischer Austauschdienst, Grant 2014 57129705. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the National Institutes of Health.

1. Pontius JU, et al.; Agencourt Sequencing Team; NISC Comparative Sequencing Program (2007) Initial sequence and comparative analysis of the cat genome. *Genome Res* 17:1675–1689.
2. Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921, and erratum (2001) 412:565.
3. Weiss RA, Stoye JP (2013) Virology. Our viral inheritance. *Science* 340:820–821.
4. Bromham L (2002) The human zoo: Endogenous retroviruses in the human genome. *Trends Ecol Evol* 17:91–97.
5. Blikstad V, Benachenhou F, Sperber GO, Blomberg J (2008) Evolution of human endogenous retroviral sequences: A conceptual account. *Cell Mol Life Sci* 65:3348–3365.
6. Suntsova M, et al. (2015) Molecular functions of human endogenous retroviruses in health and disease. *Cell Mol Life Sci* 72:3653–3675.
7. Buzdin AA, Prassolov V, Garazha AV (2017) Friends-enemies: Endogenous retroviruses are major transcriptional regulators of human DNA. *Front Chem* 5:35.
8. Hayward A, Grabherr M, Jern P (2013) Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proc Natl Acad Sci USA* 110:20146–20151.
9. Fiebig U, Hartmann MG, Bannert N, Kurth R, Denner J (2006) Transspecies transmission of the endogenous koala retrovirus. *J Virol* 80:5651–5654.
10. Denner J (2007) Transspecies transmissions of retroviruses: New cases. *Virology* 369: 229–233.
11. Holmes EC (2011) The evolution of endogenous viral elements. *Cell Host Microbe* 10: 368–377.
12. Escalera-Zamudio M, Greenwood AD (2016) On the classification and evolution of endogenous retrovirus: Human endogenous retroviruses may not be 'human' after all. *APMIS* 124:44–51.
13. Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF (2000) The nucleotide sequence of koala (*Phascolarctos cinereus*) retrovirus: A novel type C endogenous virus related to Gibbon ape leukemia virus. *J Virol* 74:4264–4272.
14. Tarlinton R, Meers J, Hanger J, Young P (2005) Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. *J Gen Virol* 86:783–787.
15. Oliveira NM, Satija H, Kouwenhoven IA, Eiden MV (2007) Changes in viral protein function that accompany retroviral endogenization. *Proc Natl Acad Sci USA* 104: 17506–17511.
16. Avila-Arcos MC, et al. (2013) One hundred twenty years of koala retrovirus evolution determined from museum skins. *Mol Biol Evol* 30:299–304, and erratum (2013) 30:1237.
17. Simmons GS, et al. (2012) Prevalence of koala retrovirus in geographically diverse populations in Australia. *Aust Vet J* 90:404–409.
18. Stoye JP (2006) Koala retrovirus: A genome invasion in real time. *Genome Biol* 7:241.
19. Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. *Nature* 442:79–81.
20. Ishida Y, Zhao K, Greenwood AD, Roca AL (2015) Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. *Mol Biol Evol* 32:109–120.
21. Cui P, et al. (2016) Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without an assembled reference genome. *PeerJ* 4:e1847.
22. Tsangaras K, et al. (2014) Hybridization capture reveals evolution and conservation across the entire koala retrovirus genome. *PLoS One* 9:e95633.
23. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R (2012) Env-less endogenous retroviruses are genomic superspreaders. *Proc Natl Acad Sci USA* 109: 7385–7390.
24. Evans LH, et al. (2009) Mobilization of endogenous retroviruses in mice after infection with an exogenous retrovirus. *J Virol* 83:2429–2435.
25. Belshaw R, et al. (2007) Rate of recombinational deletion among human endogenous retroviruses. *J Virol* 81:9437–9442.
26. Johnson RN, et al. (2018) Adaptation and conservation insights from the koala genome. *Nat Genet*, 10.1038/s41588-018-0153-5.
27. Hobbs M, et al. (2017) Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. *Sci Rep* 7:15838.
28. Jurka J (2000) Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet* 16:418–420.
29. Hobbs M, et al. (2014) A transcriptome resource for the koala (*Phascolarctos cinereus*): Insights into koala retrovirus transcription and sequence diversity. *BMC Genomics* 15:786.
30. Glover L, Jun J, Horn D (2011) Microhomology-mediated deletion and gene conversion in African trypanosomes. *Nucleic Acids Res* 39:1372–1380.
31. Verdin H, et al. (2013) Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. *PLoS Genet* 9:e1003358.
32. Vissers LE, et al. (2009) Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. *Hum Mol Genet* 18:3579–3593.
33. Chen A, Weber IT, Harrison RW, Leis J (2006) Identification of amino acids in HIV-1 and avian sarcoma virus integrase subsites required for specific recognition of the long terminal repeat ends. *J Biol Chem* 281:4173–4182.
34. Martin R, Handasyde KA (1999) *The Koala: Natural History, Conservation and Management* (UNSW Press, Sydney).
35. Flockerzi A, Burkhardt S, Schempp W, Meese E, Mayer J (2005) Human endogenous retrovirus HERV-K14 families: Status, variants, evolution, and mobilization of other cellular sequences. *J Virol* 79:2941–2949.
36. Hancks DC, Kazazian HH, Jr (2010) SVA retrotransposons: Evolution and genetic instability. *Semin Cancer Biol* 20:234–245.
37. Hughes JF, Coffin JM (2005) Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics* 171:1183–1194.
38. Vargiu L, et al. (2016) Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* 13:7.
39. Houlden BA, et al. (1999) Phylogeographic differentiation in the mitochondrial control region in the koala, *Phascolarctos cinereus* (Goldfuss 1817). *Mol Ecol* 8:999–1011.
40. Houlden BA, England PR, Taylor AC, Greville WD, Sherwin WB (1996) Low genetic variability of the koala *Phascolarctos cinereus* in south-eastern Australia following a severe population bottleneck. *Mol Ecol* 5:269–281.
41. Dennison S, et al. (2016) Population genetics of the koala (*Phascolarctos cinereus*) in north-eastern New South Wales and south-eastern Queensland. *Aust J Zool* 64:402–412.
42. Neaves LE, et al. (2016) Phylogeography of the koala, (*Phascolarctos cinereus*), and harmonising data to inform conservation. *PLoS One* 11:e0162207.
43. Lee KE, et al. (2012) Genetic diversity in natural and introduced island populations of koalas in Queensland. *Aust J Zool* 60:303–310.
44. Katzourakis A, et al. (2014) Larger mammalian body size leads to lower retroviral activity. *PLoS Pathog* 10:e1004214.
45. Nethe M, Berkhout B, van der Kuyl AC (2005) Retroviral superinfection resistance. *Retrovirology* 2:52.
46. Goodier JL (2016) Restricting retrotransposons: A review. *Mob DNA* 7:16.
47. Escalera-Zamudio M, et al. (2015) A novel endogenous betaretrovirus in the common vampire bat (*Desmodus rotundus*) suggests multiple independent infection and cross-species transmission events. *J Virol* 89:5180–5184.
48. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
49. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584.
50. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30:772–780.
51. Stajich JE, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12:1611–1618.
52. Team RDC (2011) R: A Language and Environment for Statistical Computing (The R Foundation for Statistical Computing, Vienna), Version 3.4.2.
53. Paradis E (2010) pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics* 26:419–420.