

Online resource 6 for Transcriptomic and Genomic variants between Koala populations reveals underlying genetic components to disorders in a Bottlenecked Population.

R.E.Tarlinton^{1*}, J. Fabijan², F. Hemmatzedah², J. Meers³, H.Owen³, N. Sarker³, J.M.Seddon³, G.Simmons³, N. Speight², D.Trott², L. Woolford², R.D.Emes¹

*corresponding author rachael.tarlinton@nottingham.ac.uk

¹ School of Veterinary Medicine and Science, University of Nottingham, UK

² School of Animal and Veterinary Sciences, The University of Adelaide, Australia

³ School of Veterinary Science University of Queensland, Australia

Journal: Conservation Genetics

Analysis of the Type D env retroviral sequences detected in this study.

Alignment of the three differentially expressed syncytin like transcripts was performed in CLC sequence viewer version 11.01 (Qiagen Aarhus) and a consensus sequence generated. This 2295 Bp consensus sequence was identical with the full length envelope gene of the koala "ERV1" identified in (Yu et al., 2019). This retrolement is one of three full length retroviruses (in addition to KoRV) that is currently active within the koala genome (accumulating new somatic and germline insertions). Unlike KoRV however KoERV1, KoERVL1 and KoERVK14 share insertions in unrelated animals (15 copies in all 4 animals tested for KoERV1) and therefore appear to be older introductions to the koala genome than KoRV (present in an ancestral animal). KoERV1 is transcriptionally active in testes, brain and liver (though not to the same level as KoRV) and this study adds lymph node to that list. KoERV1 appears to be a fully functional retrovirus with intact ORF for the gag, pro-pol and env proteins, the LTRS of which appear to be intact and demonstrate minimal divergence (1bp only in the sequence in the Yu paper) (Figure 1). KoERVL1 and KoERVK14 do not appear to have intact ORF for all proteins and indeed are shorter than a typical functional retrovirus (approximately 6000 bp)

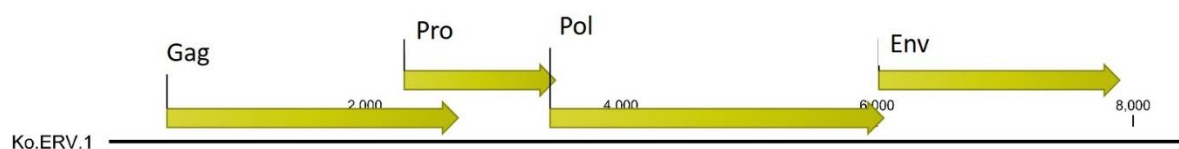


Figure 1 diagrammatic representation of the open reading frames of KoERV1

The reasons koalas have so many active retroviruses in their genome is not clear but marsupials are known to have a distinct lineage of AID/APOBEC genes and lack the APOBEC3 genes that are one of the primary restriction factors affecting species range of extant retroviruses in eutherian mammals (Ito et al., 2019). The human and murine variants of APOBEC3 proteins are known to restrict KoRV in cell culture (Nitta et al., 2015) and it is possible that marsupials have not yet evolved restriction factors able to counteract gammaretroviruses effectively.

A BLAST-P search with the consensus sequence from this study demonstrated matches against koala genome (Johnson et al., 2018) loci, with other matches including *Trichuris vulpea* type D endogenous retrovirus (Baillie and Wilkins, 2001) and a number of other retroviruses with D type envelope genes (Henzy and Johnson, 2013) including Baboon endogenous retrovirus and the domestic cat RD114 virus, Galidia (mongoose) ERV

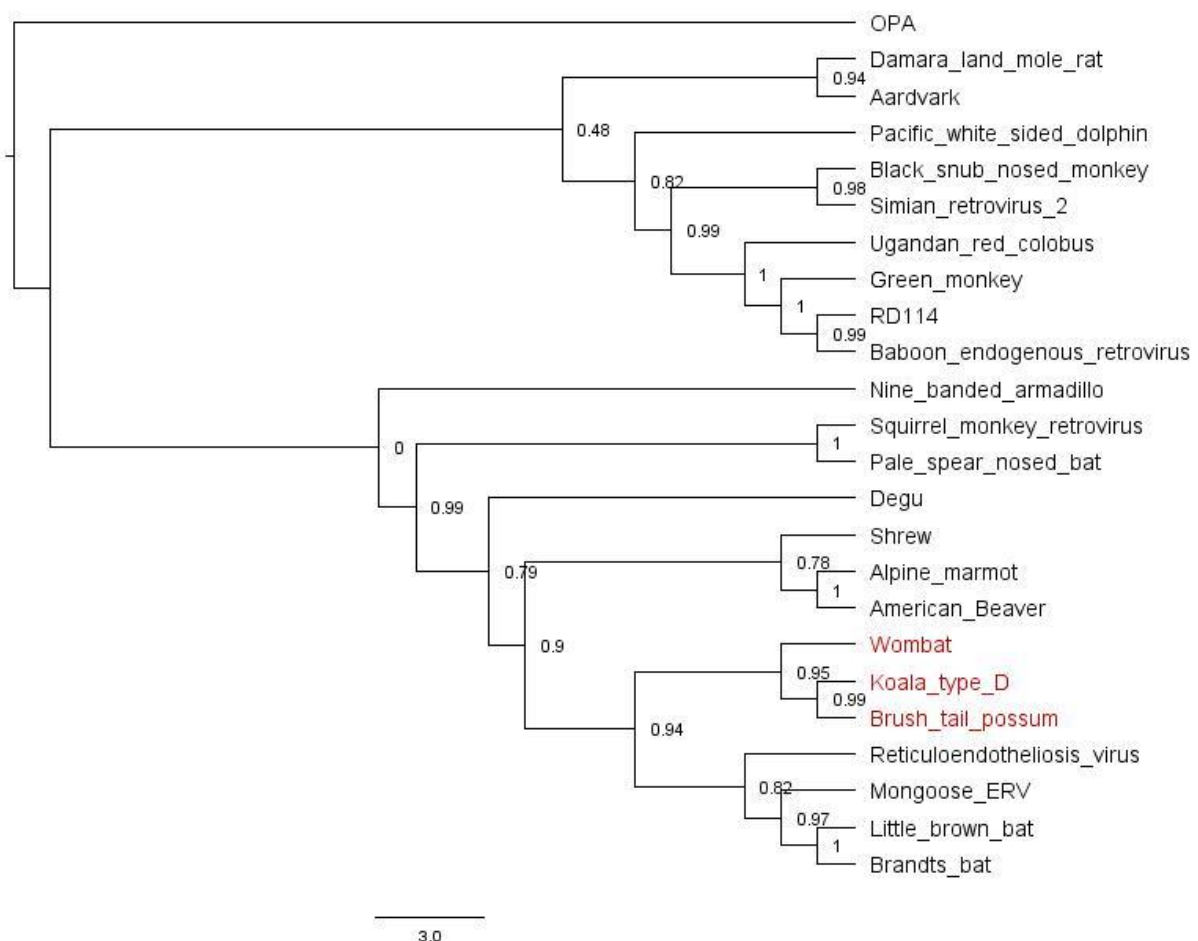
(Niewiadomska and Gifford, 2013) as well as a variety of genomic segments from a range of species.

Forty five copies of KoERV1 Env genes were retrieved from the koala genome and aligned using MUSCLE (Hirschberg kalign) in UGENE v 1.32 (Unipro), demonstrating that many of these loci have large indels (figure 2) adding further evidence that these are likely older retrotransposons

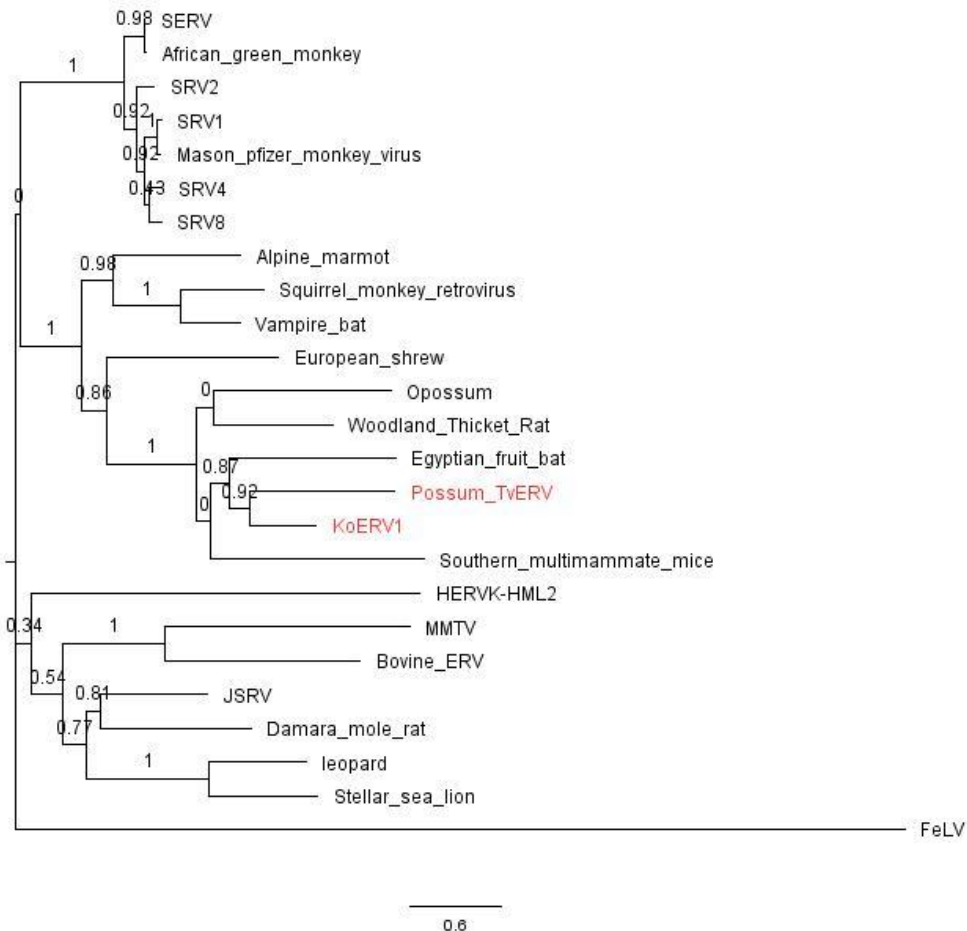


Supplementary figure 2: Graphical representation of sequence read depth in the 2500 Bp consensus sequence derived from 45 Koala type D retrovirus sequences,

Alignment of the full length 674 amino acid sequence from the consensus sequence with the corresponding portion of reference sequences of the known type D retrovirus env genes (Henzy and Johnson, 2013; Niewiadomska and Gifford, 2013) and sequence matches retrieved during the blast searches along with the sheep beta retrovirus ovine pulmonary adenoma carcinoma virus (OPA) as an outgroup was performed using Muscle (Edgar, 2004) in Seaview (Gouy et al., 2010) and maximum likelihood trees constructed using PHYML under GTR model with aLRT branch support also in Seaview (figure 3).



Supplementary Figure 3: PhyML Phylogenetic tree based on a 674 amino acid alignment of 23 type D envelope gene sequences and the sheep retrovirus OPA as an outgroup (the tree is rooted on OPA). Common names of species (for ERVs) and viruses are given. Branch support values greater than 0.5 are shown. Marsupial species are highlighted in red.



Supplementary Figure 4: PhyML Phylogenetic tree based on an amino acid alignment of 25 Beta retroviral gag sequences and the gamma retrovirus FeLV as an outgroup (the tree is rooted on FeLV). Common names of species (for ERVs) and viruses are given. Branch support values are shown. Marsupial species are highlighted in red.

Gag phylogenies (figure 4) of exogenous and full length endogenous betaretroviral gag genes (851 amino acids) with the corresponding portion of KoERV1 similarly demonstrate that the KoERV1 sequence is a betaretroviral gene most closely related to TvERV of brushtail possums.

The phylogenetic analysis demonstrates that this koala type D env sequence is part of an Australian marsupial specific clade including TvERV (common brush tail possum) and the common wombat (*Vombatus ursinus*). The common wombat is the koala's closest extant relative indicating that these sequences are likely derived from a common ancestor to both species. Adding further evidence that these viruses are likely older entrants to the genome than KoRV.

It is possible that this sequence represents a marsupial version of the syncytin placental fusion proteins derived from retroviral envelope genes identified in multiple placental

species (Cornelis et al., 2017). Of note the retrovirus identified here is not of the same class as those identified in opossums and a broad range of other marsupials in (Cornelis et al., 2015).

The type D envelope genes are notorious for displaying a propensity for host swapping and gene swapping within host with pol and gag sequences from differing retroviruses in the gamma and beta retroviral groups (as is evident by the very varied range of species in this tree) (Henzy and Johnson, 2013; Niewiadomska and Gifford, 2013), though as more species are being sequenced it may become evident that there are clades of retrovirus associated with particular taxa as in this case.

Baillie, G.J., Wilkins, R.J., 2001. Endogenous Type D Retrovirus in a Marsupial, the Common Brushtail Possum, (*Trichosurus vulpeca*). *Journal of Virology* 75, 2499-2507.

Cornelis, G., Funk, M., Vernochet, C., Leal, F., Tarazona, O.A., Meurice, G., Heidmann, O., Dupressoir, A., Miralles, A., Ramirez-Pinilla, M.P., Heidmann, T., 2017. An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. *Proc Natl Acad Sci U S A* 114, E10991-e11000.

Cornelis, G., Vernochet, C., Carradec, Q., Souquere, S., Mulot, B., Catzeflis, F., Nilsson, M.A., Menzies, B.R., Renfree, M.B., Pierron, G., Zeller, U., Heidmann, O., Dupressoir, A., Heidmann, T., 2015. Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc Natl Acad Sci U S A* 112, E487-496.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27, 221-224.

Henzy, J.E., Johnson, W.E., 2013. Pushing the endogenous envelope. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 368, 20120506.

Ito, J., Gifford, R.J., Sato, K., 2019. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proc Natl Acad Sci U S A*.

Johnson, R.N., O'Meally, D., Chen, Z., Etherington, G.J., Ho, S.Y.W., Nash, W.J., Grueber, C.E., Cheng, Y., Whittington, C.M., Dennison, S., Peel, E., Haerty, W., O'Neill, R.J., Colgan, D., Russell, T.L., Alquezar-Planas, D.E., Attenbrow, V., Bragg, J.G., Brandies, P.A., Chong, A.Y., Deakin, J.E., Di Palma, F., Duda, Z., Eldridge, M.D.B., Ewart, K.M., Hogg, C.J., Frankham, G.J., Georges, A., Gillett, A.K., Govendir, M., Greenwood, A.D., Hayakawa, T., Helgen, K.M., Hobbs, M., Holleley, C.E., Heider, T.N., Jones, E.A., King, A., Madden, D., Graves, J.A.M., Morris, K.M., Neaves, L.E., Patel, H.R., Polkinghorne, A., Renfree, M.B., Robin, C., Salinas, R., Tsangaras, K., Waters, P.D., Waters, S.A., Wright, B., Wilkins, M.R., Timms, P., Belov, K., 2018. Adaptation and conservation insights from the koala genome. *Nature genetics* 50, 1102-1111.

Niewiadomska, A.M., Gifford, R.J., 2013. The extraordinary evolutionary history of the reticuloendotheliosis viruses. *PLoS biology* 11, e1001642.

Nitta, T., Ha, D., Galvez, F., Miyazawa, T., Fan, H., 2015. Human and murine APOBEC3s restrict replication of koala retrovirus by different mechanisms. *Retrovirology* 12, 68.

Yu, T., Koppetsch, B.S., Pagliarani, S., Johnston, S., Silverstein, N.J., Luban, J., Chappell, K., Weng, Z., Theurkauf, W.E., 2019. The piRNA Response to Retroviral Invasion of the Koala Genome. *Cell* 179, 632-643.e612.

List of sequence ID's (env)

North American Beaver *Castor canadensis* XP_020043268.1

Alpine Marmot *Marmota marmota marmota* XP_015346529.

Common shrew *Sorex araneus* XP_012788699.1

REV reticuloendotheliosis virus AGS83376.1
 Mongoose *Gallidia sp* ERV AGV92854.1
 Little brown bat *Myotis lucifigus* XP_023604428.1
 Brandt's bat *Myotis brandtii* XP_014406000.1
 Common Wombat *Vombatus ursinus* XP_027713288.1
 Brush tail possum *Trichosurus vulpeca* AAG28161.1
 Degu *Octodon degus* XP_023569241.1
 Squirrel Monkey Retrovirus (*Samiri sp.*) NP_041262.
 Pale spear nosed bat *Phyllostomus discolor* XP_028381007.1
 Pacific white sided dolphin *Lagenorhynchus obliquidens* XP_026956146.1
 Aardvark *Orycteropus afer afer* XP_007945967.1
 Damara land mole rat *Fukomys damarensis* XP_019064375.1
 Nine banded armadillo *Dasypus novemcinctus* XP_004482996.1
 Ugandan red colobus *Piliocolobus tephrosceles* XP_023051236.1
 Green Monkey *Chlorocebus sabaeus* BAP46325.1
 RD114 (domestic cat ERV, *Felis catus*) YP_001497149.1
 Baboon endogenous retrovirus (BAEV) YP_009109691.1
 Black snub nose monkey *Rhinopithecus bieti* XP_017706409.1
 Simian retrovirus 2 (SRV-2) Rhesus macaque *Macaca mulata* P51520.1

List of sequence ID's (gag)

HERVK HML2 *Homo sapiens* AF074086
 Brush tail possum *Trichosurus vulpeca* AAG28161.1
 Bovine *Bos Taurus* BosTau3_chr4_33488089_33501364
 MMTV *Mus musculus* NP_056882.1
 JSRV *Ovis aries* YP_238302.1
 SERV *Chlorocebus sabeus* AEJ22865.1
 SRV2 *Maccaca mulata* AAD43258.1
 SRV4 *Maccaca mulata* YP_003864100.1
 SRV1 *Maccaca mulata* P04025.2
 Mason Pfizer monkey retrovirus *Maccaca mulata* NP_056893.1
 Feline leukaemia virus *felis cattus* AAA93092.1
 Squirrel Monkey Retrovirus (*Samiri sp.*) NP_041259.1
 Woodland thicket rat *Grammomys surdaster* XP_028619792

Egyptian fruit bat *Rousettus aegyptiacus* XP_015980625

Opossum *Monodelphis domestica* XP_016279143

European shrew *Sorex araneus* XP_012790545

Vampire bat *Desmodus rotundus* YP_009140128

SRV-8 Simian retrovirus 8 YP_009305199

African green monkey *Chlorocebus sabaeus* BBC20776

European marmot *Marmota marmota marmota* XP_015359287

Southern multimammate mice *Mastomys coucha* XP_031215958

Leopard *Panthera pardus* XP_019270316

Damara mole rat *Fukomys damarensis* XP_010624009

Stellar sea lion *Eumetopias jubatus* XP_027952711

Nucleotide sequence syncytin like transcript XM-020999245.1

ATGAGGTCTGCCCTTGGTTTATGGTTACCGGCCAGATATTCATCCGGGAAAAGTGGCAGAAGAA
AATGACGATCCTCGTAATTTTCTCCTGTTCCGTGGGAGCCACGGTGGTTTTCAATCTCCTGACC
CAGTAGCGCTCAGCCGTGCCCTCTTTGGAAGCCCTTGTGACTGCGGCGGGGGGCACGTTCACTGT
CCGTCCCGCCACCTACACGCAAGAAGTGGACTGCGGGGAAAAGACAGCCTACCTCGCCTACCGT
CATTCTATCACGGGCGTCTCCAAACCCAAGTGGGAATGCGTAAGGAAGCCCCGAGTTATCCCTC
CGAATGGTGGCCAGCCCCGGGCCTTGCCCCAGTGATTGCAACTACCTTGAGGCCTTGCACTCCAC
ATGCTATAGTTCTACCCAGCAGTGCAGTGGCAGCTCAGGAACAGTCTATCTTACTACAATACAGG
AAAGAGAATATGGTGGCTCCACAGGAGGGGACTGGGCGCATATAACTAGCCAAGGCCTACGTG
GAAGTACCAATAAGTATGCTCAGGCTTCTGTGCTAAAGAAAATATTGGAAAAAATGTCTGTTGG
CCCCTTCATGCACCTGTCCATATCTCTGATGGGGGTGGCCCTACAGACCAAATTAGGGAACATGA
GGTTAGGGAACGAATAAATGAAATAATTAAGAGTATGTATCCCCCATCCATTACCATCCACTTG
CACTGCCCAAACCCCGGGGAACGGACCTAGATGTCCAAACTTCAGAAATCCTTGAAGCAACACTT
AAGGCCCTTAATTACTCTAACCCTGATTTGGCTGCCGATTGCTGGCTGTGTATGGCCCTTGGCAC
ACCTATGCCCATAGCTTTGATGTCCAAAAATGGCTCCACATCCACAAATTGCACACTTAGTCCACC
TTTTAAGGTACAACCTGTTGGTCTCCGCCCTCTCCTTGTATCCAAGCCCCATTTTCAGAAATCCAC
TGCTGATATTGATGTTGTTTTGCTACCTTTGCTAATTGCTCTGAAGTCAAAAACTATACCTCACA
GGAACCTCTCTGTCCCCTGCCGGGACAACCTTTTCGCCTGTGGGGGAAATATGGCCTACTCTGCT
CTCCCCCAGAATTGGACAGGACTCTGCACACAAATTTCAATCCTCCCTGATATCGACATCATTCC
GGGAGATGAGCCTGTCCCACTGCCAGTTTTGATTATATTACTGGAAGACCCAAGAGGGCTGTT
GCATTTATCCCCCTCCTTGTTGGCCTTGGGATTACAGGGGCACTCGCTACCGGCTCTGCCGGCAT
AGGAGTTGCTGTTGATTCATATACAAAATATCCCATCAATTTCTTAATGATGTGCAAACCTCTC
GGGGACGATAAATGACCTTCAAGACCAAATTGACTCCCTGGCGGGAGTCGTTCTTCAGAATAGG
AGGGGCCTAGATTTACTCACGGCAGAACAAAGGGGGAATTTGCTTAGCCCTACAAGAAAAATGTT
GCTTCTATGCAAATAAATCTGGGATAGTGCGGGATAAAATCAAAAACTGCAAGAGGATCTGATT
GAAAGACGCAGAAAACCTTTTGAATCCCTCTGGAGTGGCTTGAATGGAGTCTCCCATACCT
CCTCCCTCTTCTTGACCTTTAATTGGTCTCCTCCTACTTTTTCTTTTGGACCTTGGGCCTTTACC
AGACTAACCTCTTTCATTAAGACGCAGATTGAGGCTTCTCTTAAAAACACCGTCAATATCCATTAC
CACCAACTTGCTTCGTCAGAAGATCCTACTGATACCATGGAAGAGGGTCTCCAGTTCTCGAAGCT
CGTGCATCCGGAATCTCGCTGCTCGCGGTGGTGGAGAAATATTAAGGATAAAAAGTGATCTGGC
GGCTGTGAGAGAGCCCATGACGGGAATATTGAGGTCCCATGTTTCAAAAAAGATCAGCGCGGCT
GCAGTTGCTTCCCATGACGGGAATAGAGTGAGGCCATGATGGGTTACATCGTGGGGTCCCCTC
GCTTAGCCTAAGACAGGGACCACTGCCACTAATGCGCTTTGAACATGGAGGCCGACCTCATAGC

CTAAGACAGGCCTCTCACCTGCTATTTTAAATACAAAATGGGGGACATGCTGGAGGCCAGCACCT
GGTAGTGGGCCCCTAACCCCGTGAGATTACCACACCCCAAGCAGCTTGTTCCGGGACAAACCGT
ACCATAACGTGTTCTGTCAGGGCTAACCGCAAGATGTCGCCACACCGCCCCTGGAAAGTCCCG
GAGGCCCATTTTCCCCTTTGCCCTCCCGCTTTGCCCTATATAAGCTTGTAACGCTCAAAAAT
AAAATTAGACCTTGCCTACCATC