# Genome Reference Assembly for Bottlenecked Southern Australian Koalas

Adam Mark Blanchard [1,*], Richard David Emes[1], Alex David Greenwood [2], Nadine Holmes[3], Matthew William Loose [3], Gail Katherine McEwen[2], Joanne Meers[4], Natasha Speight[5], and Rachael Eugenie Tarlinton[1]

[1]School of Veterinary Medicine and Science, University of Nottingham, Leicestershire, United Kingdom
[2]Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany
[3]School of Life Sciences, University of Nottingham, United Kingdom
[4]School of Veterinary Science, University of Queensland, Australia
[5]School of Animal and Veterinary Sciences, University of Adelaide, Australia

*Corresponding author: E-mail: adam.blanchard@nottingham.ac.uk.

**Accepted:** 13 December 2022

## Abstract

Koala populations show marked differences in inbreeding levels and in the presence or absence of the endogenous Koala retrovirus (KoRV). These genetic differences among populations may lead to severe disease impacts threatening koala population viability. In addition, the recent colonization of the koala genome by KoRV provides a unique opportunity to study the process of retroviral adaptation to vertebrate genomes and the impact this has on speciation, genome structure, and function. The genome build described here is from an animal from the bottlenecked Southern population free of endogenous and exogenous KoRV. It provides a more contiguous genome build than the previous koala reference derived from an animal from a more outbred Northern population and is the first koala genome from a KoRV polymerase–free animal.

**Key words:** genomes, assembly, koalas.

## Significance

This high-quality genome build provides a baseline comparator for studies of koala genetics and retroviral integration. It is from a genetically distinct population than the current koala reference genome and does not contain intact endogenous Koala retrovirus.

## Introduction

Koalas are an iconic marsupial species classed as vulnerable on the IUCN red list. The species suffers a number of threats, including habitat loss and disease, with climate change–driven fire events, further decimating numbers in recent years (Charalambous and Narayan 2020). The disease threats to the population are complicated by stark differences in the disease patterns in different populations driven by underlying genetic differences (Sarker et al. 2020; Tarlinton et al. 2021). Wild koalas are confined to the Eastern Seaboard of Australia. There are five major genetic groups (Lott et al. 2022), but for the purposes of population management, two major genetic splits are recognized: Northern (New South Wales and Queensland) and Southern (Victoria and South Australia), the border between the states of New South Wales and Victoria forming a hard cutoff between the two populations (Neaves et al. 2016; Quigley et al. 2021).

Koalas in the southern states were essentially extinct by 1920 due to hunting pressure and were restocked across their southern range from a very small number of animals (possibly as few as 18) sourced from offshore island refugia (Martin et al. 1999). As a result of this, animals in the southern population have a markedly reduced genetic diversity compared with animals in the northern population (Neaves et al. 2016; Ruiz-Rodriguez et al. 2016; Johnson et al. 2018; Tarlinton et al. 2021). Our own work has demonstrated that many genes are homozygous in the southern animals (Tarlinton et al. 2021). Animals in the southern populations suffer from a number of diseases such as oxalate nephrosis and testicular aplasia that are not routinely seen in northern populations (Fabijan et al. 2020; Tarlinton et al. 2021) and are thought to have an underlying genetic basis (Cristescu et al. 2009; Speight et al. 2020).

The other major difference both disease- and genetics-wise between Northern and Southern animals is the presence of a functional recently endogenized retrovirus (Koala retrovirus or KoRV) in all Northern koalas but not in the Southern (Quigley et al. 2021; Blyton et al. 2022). Both Southern and Northern animals may have exogenous infectious KoRV, but the rate of KoRV-associated neoplasia is substantially lower in Southern koalas (Sarker et al. 2020; Joyce et al. 2021; Quigley et al. 2021). Although the definitive link is less clear than for neoplasia (McEwen et al. 2021), KoRV is also thought to cause underlying immunosuppression predisposing to chlamydia disease, which is also seen at a lower rate in Southern populations (Polkinghorne et al. 2013; Sarker et al. 2020). Endogenous retroviruses are present in all vertebrate genomes studied to date, and the entrance of these transposable elements into genomes is thought to be a major introduction of genetic diversity, potentially triggering speciation. However, most examples in genomes are ancient and are essentially represented by inactive viruses (Zheng et al. 2022). They are thought to be the remnants of past infectious viral integrations that have managed to enter germline cells and become fixed in a species. KoRV is part of a very small group of recently endogenized viruses, integrated sometime between 200 and 49,000 years ago (Ishida et al. 2015), and is unique in that parts of the species range do not yet have endogenous polymerase gene containing KoRV at all (Quigley et al. 2021).

To complicate matters further, both Northern and Southern koalas have evidence of historical KoRV infection as defective recombinant sequences between KoRV and another older endogenous retroelement (Phascolarctos endogenous retrovirus or PhER), known as recKoRVs (Löber et al. 2018; Tarlinton et al. 2022). It is not entirely clear how endogenous and exogenous KoRV and recKoRV interact and whether they enhance or inhibit each other's replication and disease occurrence, but the scenario provides a unique opportunity to study the impacts of the entrance of a new class of retroelements into a mammalian genome in real time rather than by phylogenetic inference of this fundamental genomic process (Tarlinton et al. 2022).

There are two other published koala genomes (Johnson et al. 2018) derived from northern animals "Bilbo" and "Pacific Chocolate" (Johnson et al. 2018), alongside several additional transcriptome resources (Hobbs et al. 2014; Abts et al. 2015; Tarlinton et al. 2022). The most complete existing genome for Bilbo is assembled at a contig level (into 1,907 contigs with an N50 of 11.6 Mb). Here, we present a genome build of a Southern Australian animal "Wilpena" for use in comparative genomics of koala populations and studies of retroviral integration. This genome is more contiguous than the current reference sequence (1,265 contigs, N50 = 48.8 Mb) and from an animal known to be free of both endogenous and exogenous replication-competent KoRV (Tarlinton et al. 2022).

## Results and Discussion

Using 58 GB of ONT data (consisting of 2,572,260 reads with a mean read length of 24 kb and mean *Q* score of

**Table 1**
Summary of the Genome Assembly

| Genome | Wilpena<br>This Study | Bilbo<br>GCA_002099425 | Pacific Chocolate<br>GCA_900166895 |
|---|---|---|---|
| Assembly size | 3,234,982,288 bp | 3,192,581,492 bp | 3,358,707,742 bp |
| Number of contigs | 1,265 | 1,907 | 796,464 |
| Contigs ≥5,000 bp | 1,222 | 1,804 | 16,989 |
| Contigs ≥50,000 bp | 651 | 662 | 8,361 |
| Contigs N50 | 48,800,306 bp | 11,587,828 bp | 880,973 bp |
| Contigs N75 | 22,144,309 bp | 6,857,650 bp | 321,283 |
| Contigs L50 | 17 | 85 | 1,100 |
| Contigs L75 | 41 | 173 | 2,591 |
| Largest contig | 232,027,266 bp | 40,558,015 bp | 5,231,295 bp |
| GC content (%) | 39.09 | 39.05 | 39.03 |
| BUSCO completeness (%) | 92.9 | 94.0 | 90.0 |
| Genes | 27,669 | 32,109 | 33,654 |

13.7), 1,289 million (2 × 150 bp > *Q*30) Illumina reads were assembled into a draft genome using Flye. This resulted in an N50 of 48,782,874 bases and a length of 3,233,824,327 bp. A first-pass polish using Medaka and a final polish with Polca using the Illumina data resulted in a final high-quality genome assembly with an N50 of 48,800,306 bases, 1,265 contigs, and a total genome size of 3,234,982,288 bp (table 1).

The contigs were assessed for putative contamination using Conterminator (Steinegger and Salzberg 2020). From 1,265 contigs, 1,247 were assigned as koala and 18 were flagged as containing potential contamination. Of those 18, assignments were for North American opossum ($n = 2$), common brushtail ($n = 2$), gray short-tailed opossum ($n = 2$), common wombat ($n = 7$), and KoRV ($n = 2$). However, the same eight contigs were flagged multiple times with close marsupial relatives and so are unlikely to be true contamination. There were two contigs assigned as KoRV, these are nonfunctional partial recKoRV sequences (partial KoRV env and LTR) as reported previously in this animal (Tarlinton et al. 2022) and are not full-length endogenous or exogenous KoRV. The genome was soft-masked using REpeat Detector (RED) (supplementary table S1, Supplementary Material online) and genes predicted used braker2 along with publicly available Koala RNASeq data from multiple biological sites, predicting 52,384 putative genes. Functional annotation using the EggNOG mapper identified 27,669 genes with transcriptional support (supplementary table S2, Supplementary Material online).

## Conclusion

A highly contiguous reference genome, from a distinct southern population, is invaluable to understanding the challenges faced in conservation genetics for future breeding programs of Koalas. Not only will this enable more comprehensive comparative genomics to take place but it will also allow researchers to fully understand nonfunctional KoRV integration sites and whether they appear in similar regions of the genome to the northern population.

## Materials and Methods

### Sample Collection

DNA was derived from liver tissue from a 3-year-old female south Australian Koala, housed in a collection in the United Kingdom. The animal was originally derived from the Mt Lofty Ranges and Kangaroo Island populations in South Australia. Sample collection and nanopore sequencing from this animal were described in Tarlinton et al. (2022). Ethics approval for the use of postmortem material was granted by the University of Nottingham, School of Veterinary Medicine and Science, Committee for Animal Care and Research Ethics.

### Sample Preparation

DNA was extracted from frozen liver tissue using the QIAGEN Genomic-tip 100/G Kit and the QIAGEN Genomic Buffer Set (QIAGEN; 10243 and 19060). Frozen tissue was ground under liquid and 100 mg of frozen powder was added to 9.2 ml of buffer G2 containing 5 µl of RNase A (100 mg/ml) (QIAGEN; 19101), and the suspension was incubated at room temperature for 10 min. Proteinase K (100 µl) (QIAGEN; 19131) was added and the suspension was incubated at 50 °C for 1.5 h. The genomic-tip protocol was then followed, according to the QIAGEN Genomic DNA Handbook 06/2015.

### Genome Sequencing

Genomic DNA was needle-sheared 30 times with a 26 G needle (BD; 300300) and then treated with the Short Read Eliminator Kit (Circulomics; SS-100-101-01) to remove fragments <10 kb and progressively deplete fragments shorter than 25 kb. The processed DNA was used to generate a sequencing library using the Genomic DNA by Ligation PromethION Kit (Oxford Nanopore Technologies; SQK-LSK109). Library quantification was performed using the Qubit fluorometer and the Qubit dsDNA HS Assay Kit (ThermoFisher; Q32854), and 600 ng of the library was run over one PromethION flow cell (Oxford Nanopore Technologies; FLO-PRO002) on a PromethION Beta device. The same DNA preparation was subjected to Illumina NovaSeq 6000 paired-end 150 bp read sequencing (with an automated plant and whole-genome library preparation) by Novogene, Cambridge, United Kingdom.

### Read Processing

Illumina reads (both RNA and DNA) were trimmed to remove adaptors and reads with an overall quality of <*Q*30 using FastP v0.23.1 (Chen et al. 2018). The raw Nanopore data were base-called using Guppy v6.1.7 + 21b93d1a5 and the super-accurate mode (https://community.nanoporetech.com/downloads). Nanopore adaptors were removed using Porechop v0.2.4 (Wick et al. 2017) and reads shorter than 1,000 bp and with a quality of<*Q*10 were removed with NanoFilt v2.6.0 (De Coster et al. 2018).

### Assembly

The Nanopore reads were assembled using Flye v2.9.1 with the –nano-hq and –keep-haplotypes flags. The Flye draft assembly was first polished with Medaka v1.6.1 (https://github.com/nanoporetech/medaka) and then with the Illumina reads using POLCA (from MaSuRCA v4.0.9) (Zimin and Salzberg 2020). The resulting polished assembly was then gap-filled using Samba (from MaSuRCA v4.0.9) (Zimin and Salzberg 2022) before being assessed for completeness using BUSCO v5.4.2 (Manni et al. 2021) in genome mode with the Mammalian lineage database.

### Contamination Assessment

Each contig was assigned a taxonomic ID using BlastN (Altschul et al. 1990); this was parsed into a Conterminator (Steinegger and Salzberg 2020) to identify potential regions of contamination in the genome.

### Annotation

The final version of the genome was parsed through RED v1.16 (Girgis 2015) to soft-mask regions of repetitive elements. An index was created using HISAT2 v2.2.1 (Kim et al. 2019), and RNASeq data from accession: PRJNA230900 (Hobbs et al. 2017) were aligned producing SAM files. SAMTools v1.15 (Danecek et al. 2021) was used to convert SAM to BAM before being used in Braker2 v2.1.6 (Brůna et al. 2021) for genome annotation. Functional annotation was completed using EggNOG mapper v2.1 (Cantalapiedra et al. 2021).

## Supplementary material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

## Acknowledgments

Sample access was facilitated by the Longleat Safari Park. Funding was provided by the University of Nottingham, and A.D.G. and G.K.M. were supported by grant GR 3924/15-1 from the Deutsche Forschungsgemeinschaft (DFG).

## Author Contributions

R.E.T. oversaw project management, wrote funding proposals, collected postmortem samples, and wrote parts of the manuscript. N.H. performed the DNA extraction and nanopore sequencing. M.W.L. and A.M.B. performed bioinformatics analysis. A.M.B. wrote parts of the manuscript and performed data deposition in public repositories. A.D.G. and G.K.M. project-managed the Illumina sequencing. N.S. performed initial screening testing for KoRV on the animal. N.S. and J.M. provided a critical review of the manuscript. All authors read and reviewed the manuscript.

## Data Availability

All raw sequence data are available on the NCBI SRA under the accession number SAMN30742200. The final genome build (*Phascolarctos cinereus* K01) is available under the NCBI accession number JAOEJA000000000.

## Literature Cited

Abts KC, Ivy JA, DeWoody JA. 2015. Immunomics of the koala (*Phascolarctos cinereus*). Immunogenetics 67:305–321.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Blyton MDJ, Young PR, Moore BD, Chappell KJ. 2022. Geographic patterns of koala retrovirus genetic diversity, endogenization, and subtype distributions. Proc Natl Acad Sci U S A. 119:e2122680119.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 3:lqaa108.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 38:5825–5829.

Charalambous R, Narayan E. 2020. A 29-year retrospective analysis of koala rescues in New South Wales, Australia. PLoS One 15: e0239182.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890.

Cristescu RH, et al. 2009. Inbreeding and testicular abnormalities in a bottlenecked population of koalas (*Phascolarctos cinereus*). Wildl Res. 36:299–308.

Danecek P, et al. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10:giab008.

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34:2666–2669.

Fabijan J, et al. 2020. Pathological findings in koala retrovirus-positive koalas (*Phascolarctos cinereus*) from Northern and Southern Australia. J Comp Pathol. 176:50–66.

Girgis HZ. 2015. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinform. 16:227.

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.

Hobbs M, et al. 2017. Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. Sci Rep. 7:15838.

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.

Johnson RN, et al. 2018. Adaptation and conservation insights from the koala genome. Nat Genet. 50:1102–1111.

Joyce BA, Blyton MDJ, Johnston SD, Young PR, Chappell KJ. 2021. Koala retrovirus genetic diversity and transmission dynamics within captive koala populations. Proc Natl Acad Sci U S A. 118:e2024021118.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37:907–915.

Löber U, et al. 2018. Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germline invasion. Proc Natl Acad Sci U S A. 115:8609–8614.

Lott MJ, et al. 2022. Future-proofing the koala: synergising genomic and environmental data for effective species management. Mol Ecol. 31:3035–3055.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 38:4647–4654.

Martin R, Handasyde KA, Lee AK. 1999. The koala: natural history, conservation and management. 2nd ed. Sydney (Australia): UNSW Press.

McEwen GK, et al. 2021. Retroviral integrations contribute to elevated host cancer rates during germline invasion. Nat Commun. 12:1316.

Neaves LE, et al. 2016. Phylogeography of the koala, (*Phascolarctos cinereus*), and harmonising data to inform conservation banks. PLoS One 11:e0162207.

Polkinghorne A, Hanger J, Timms P. 2013. Recent advances in understanding the biology, epidemiology and control of chlamydial infections in koalas. Vet Microbiol. 165:214–223.
Quigley BL, Wedrowicz F, Hogan F, Timms P. 2021. Phylogenetic and geographical analysis of a retrovirus during the early stages of endogenous adaptation and exogenous spread in a new host. Mol Ecol. 30:2626–2640.
Ruiz-Rodriguez CT, et al. 2016. Koalas (*Phascolarctos cinereus*) from Queensland are genetically distinct from 2 populations in Victoria. J Hered. 107:573–580.
Sarker N, et al. 2020. Koala retrovirus viral load and disease burden in distinct northern and southern koala populations. Sci Rep. 10: 263.
Speight N, Bacci B, Stent A, Whiteley P. 2020. Histological survey for oxalate nephrosis in Victorian koalas (*Phascolarctos cinereus*). Aust Vet J. 98:467–470.Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol 21:115.
Tarlinton RE, et al. 2021. Transcriptomic and genomic variants between koala populations reveals underlying genetic components to disorders in a bottlenecked population. Conserv Genet. 22:329–340.
Tarlinton RE, et al. 2022. Differential and defective transcription of koala retrovirus indicates the complexity of host and virus evolution. J Gen Virol. 103:001749.
Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genomics 3(10).
Zheng J, Wei Y, Han G-Z. 2022. The diversity and evolution of retroviruses: perspectives from viral "fossils". Virol Sin. 37:11–18.
Zimin AV, Salzberg SL. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol. 16:e1007981.
Zimin AV, Salzberg SL. 2022. The SAMBA tool uses long reads to improve the contiguity of genome assemblies. PLoS Comput Biol. 18:e1009860.

**Associate editor**: Bonnie Fraser