Koala Genome Supplementary Information

Deleted: 25

Deleted: 29

Deleted: 45

Supplementary Note

1 Koala Genome Sequencing and Assembly

1.1 Sample collection and DNA extraction

Tissue samples from three wild koalas (*Phascolarctos cinereus)* were collected into the gaseous phase of liquid nitrogen (-196°C) and then stored at -80°C. “Pacific Chocolate” (Australian Museum registration M.45022), a female from Port Macquarie in northeast New South Wales was sampled immediately after euthanasia by veterinary staff at the Port Macquarie Koala Hospital (27/06/2012), following unsuccessful treatment of severe chlamydiosis. Two koalas from southeast Queensland; a female, “Bilbo” (Australian Museum registration M.47724) from Upper Brookfield, and a male, “Birke", from Birkdale, were sampled by veterinary staff at Australia Zoo Wildlife Hospital following euthanasia due to severe chlamydiosis (20/08/2015) and severe injuries (26/8/2012) respectively.

Koala samples were obtained as part of veterinary care at the Port Macquarie Koala Hospital and Australia Zoo Wildlife Hospital, and from the Australian Museum Tissue Collection. Sample collection was performed in accordance with methods approved by the Australian Museum Animal Ethics Committee (Permit Numbers: 11–03, 15–05).

High Molecular Weight (HMW) DNA was extracted from heart tissue for “Pacific Chocolate” and kidney tissue for “Birke” using the DNeasy Blood and Tissue kit (Qiagen 69506), with RNase A (Qiagen 19101) added following digestion. HMW chromosomal DNA (20-150 kb) from “Bilbo” was extracted from spleen tissue using Genomic-Tip 100/G columns (Qiagen 10243) and DNA Buffer set (Qiagen 19060) with RNAse A(Qiagen 19101) treatment. The Qiagen Tissue Protocol was modified by replacing the mechanical tissue disruption with 12 hour incubation at 50 degrees and spooling the precipitated DNA. DNA was resuspended in 10Mm Tris-Cl buffer, ph 8.5.

1.2 Genome library preparation and sequencing

Female koala “Bilbo” reference genome sequencing using PacBio

HMW gDNA was assessed for quality at the Ramaciotti Centre for Genomics (RCG; University of New South Wales, Sydney, Australia) by Nanodrop 1000 Spectrophotometer (Thermo Fisher Scientific, USA), Qubit 2.0 Fluorimeter (Thermo Fisher Scientific, USA) and analyzed on an 0.75% KBB agarose gel using the Pippin Pulse™, pulse field electrophoresis gel system (Sage Science). 15 SMRTbell libraries were prepared (RCG) as per the PacBio 20-kb template preparation protocol, with an additional damage repair step performed after size selection. A minimum size cutoff of 15 or 20 kb was utilized in the size selection stage using the Sage Science BluePippin™ system. The SMRTbell libraries were quality checked using Qubit (Thermo Fisher, Cambridge, UK) and the Agilent BioAnalyzer system (Agilent, Stockport, UK).

The libraries were sequenced on the Pacific Biosciences RS II platform (Pacific Biosciences) employing P6 C4 chemistry with either 240 min or 360 min movie lengths. A total of 272 SMRT Cells were sequenced to give an estimated overall coverage of 57.3x based on a genome size of 3.5 Gbp.

Female koala “Bilbo” - Illumina HiSeq XTen sequencing for error correction

HMW gDNA was sequenced on an Illumina 150bpPE HiSeq X Ten sequencing run (Illumina) at the Kinghorn Centre for Clinical Genomics (The Garvan Institute of Medical Research, Sydney Australia). A TruSeq DNA PCR free library was constructed with a mean library insert size of 450 bp. 400,473,997 paired-end reads were generated yielding a minimum coverage of 34x.

Male koala “Birke” comparison genome - Illumina HiSeq2000

HMW gDNA was sheared to 300-400bp using the Covaris E220™ and a library was prepared using the TruSeq DNA library prep kit (Illumina) at RCG. The library was size selected using the Sage Science Pippin™ in the 400-500 bp range and sequenced on seven lanes of 100 bp paired-end sequencing on the Illumina HiSeq2000 (RCG), to produce an estimated 100x coverage on a 3.5 Gbp genome.

Female koala "Pacific Chocolate" comparison genome - DISCOVAR Bio-NANO

Amplification-free libraries were generated at the Earlham Institute (Norwich, UK) from 500 ng DNA sheared to an average 400 bp on a Covaris S2 (Covaris, Massachusetts, USA). A magnetic bead-based size selection was performed using 0.58x and 3x volume Hi Prep Beads (GC Biotech, Alphen aan den Rijn, The Netherlands). All subsequent enzymatic reactions were performed ‘on bead’. Fragmented DNA was blunt ended (end repair module, NEB, Hitchin), cleaned (1x Hi Prep bead supernatant clean-up), A-tailed (NEB) and TruSeq adapters (Illumina) were ligated with Blunt T/A ligase (NEB). Stop ligation buffer was added and two 0.7x volume Hi Prep Bead supernatant purifications undertaken. Final library molecules were eluted in 20 µl 10mM TRIS-HCl and quality checked using Qubit (Thermo Fisher, Cambridge, UK) and Agilent BioAnalyser (Agilent, Stockport, UK) and quantified using Kapa Biosystems Illumina quantification qPCR (Kapa Biosystems, London, UK).

The resulting library was diluted to 1 nM with elution buffer (Qiagen) and NaOH, and subsequently diluted to 10 pM in HT1 (Illumina). 135 µL diluted library pool, spiked with 1% PhiX Control v3 was loaded onto the Illumina cBot. The flow cell was clustered using HiSeq Rapid Paired-end Cluster Generation Kit v2, following the appropriate clustering recipe then the flow cell was loaded onto the Illumina HiSeq2500 instrument following manufacturer’s instructions. The sequencing chemistry used was TruSeq SBS Kit v2 using HiSeq Control Software 2.2.58 and RTA 1.18.64. The library was sequenced over five lanes for paired-end 250 bp sequences with a mean insert size of 395, producing >675 million paired-end reads for an estimated read coverage of 107x on a 3.5 Gbp genome. . Using DISCOVAR [1], five different de novo assemblies were created using 21x, 43x, 64x, 85x and 107x coverage and assembly stats were established for each assembly. Assembly stats (N90 - N10) for a range of coverages. Other than N90 (85x

best) and longest contig (43x best), 64x coverage provides the best overall contiguity (Supplementary Table 1).

Use of BioNano to extend scaffold sizes using the DISCOVAR assembly

BioNano was used to extend the scaffold sizes using the DISCOVAR assembly from “Pacific Chocolate”. A hybrid assembly was created, mapping the DISCOVAR assembly back to the restriction-site labelled BioNano molecules, resulting in increased N50 of 869 kb, almost three times the original assembly N50. TGAC’s Platforms and Pipelines group followed IrysPrep™ Animal Tissue Dounce protocol supplied by Bionano Genomics for extracting the DNA. 80mg of spleen frozen at -80°C was taken and ground in a 7ml dounce with 4 ml of MB buffer (10mM Tris, 10mM EDTA, 100mM NaCl, pH9.4), gently grinding 20 times. The dounce was placed on ice for 5 minutes to settle non-homogenized material. Every 500 µl removed had an added equal amount of 100% ETOH into an Eppendorf. This was mixed gently 5 times and incubated for 60 minutes on ice. The homogenate was spun down at 1,500xg for 7 minutes at 4°C, removing the supernatant and resuspension in 1 ml of MB buffer. The spin was repeated followed by removal of the supernatant, allowing the cells to come to room temperature. Cells were resuspended in around 50 µl of MB buffer using wide bore tip to a total volume of 66 µl. 40 µl of LMP agarose melted and at 43°C was added and mixed using a wide bore tip. 8 plugs of 90ul were cast using the Chef Mammalian Genomic DNA Plug Kit (Bio-Rad 170-3591). Once set at 4°C, 4 the plugs each were added to a lysis solution containing 200 µl proteinase K (QIAGEN 158920) and 2.5 ml of Bionano lysis buffer in two 50 ml conical tubes. These were put at 50°C for 2 hours on thermomixer, making a fresh proteinase K solution to incubate overnight. The 50 ml tubes were then removed from the thermomixer for 5 minutes before 50 µl RNAse A (Qiagen158924) was added and the tubes returned to the thermomixer for a further hour at 37°C.The plugs were then washed 7 times in Wash Buffer supplied in Chef kit and 7 times in 1xTE. One plug was removed and melted for 2 minutes at 70°C followed by 5 minutes at 43°C before adding 10ul of 0.2U/µl of GELase (Cambio Ltd G31200). After 45 minutes at 43°C the melted plug was dialyzed on a 0.1uM membrane (Millipore VCWP04700) sitting on 15 ml of 1xTE in a small petri dish. After 2 hours the sample was removed with a wide bore tip and mixed gently 5 times and left overnight at 4°C. A small amount was removed to QC on an Opgen Argus Q-Card and Qubit HS for the DNA concentration. 300ng of DNA was taken into the NLRS (Nick, Label, Repair and Stain) reaction using 1 µl Nt.BspQI (NEB R0644S). Following the NLRS a second QC step was performed using a Q-Card and Qubit HS before 16ul was loaded onto a single flow cell on a Bionano chip. The Chip loading was optimized and run for 30 cycles on the Bionano Irys using ICS1.6. The same chip was run a total of 13 times to generate 580 Gb of raw data with 260 Gb over 180 kb with a nick density of 10/100 kb. Images were converted to .bnx files using AutoDetect 2.1.0.6656 before analysis.

The PacBio-Bionano assembly was assembled using the Kansas State Irys-scaffolding pipeline, which uses custom scripts to prepare (see Code Availability and assemble the Bionano raw molecule maps, which it then uses to super-scaffold the reference FASTA genome. It did not produce a conflict file or SV information showing regions of the genomes where the Bionano molecule maps and the reference FASTA file do not concur.

1.3 Genome assembly and quality assessment

"Bilbo" Long-Read PacBio *de novo* assembly with FALCON (phaCin_unsw_v4.1)

An overlapping layout consensus assembly algorithm, FALCON (v 0.3.0) (https://github.com/PacificBiosciences/FALCON-integrate), was used to generate the draft genome using PacBio reads. Total genome coverage before assembly was estimated by total bases from reads divided by 3.5 Gbp genome size. The estimated total coverage is 57.3x. FALCON leverages error-corrected long seed reads to generate an overlapping layout consensus representation of the genome. Approximately 23x of long reads are required by FALCON as seed reads, and the rest are used for error correction. The seed read length of the reads at the 60% percentile is calculated as 10,889 bp. The FALCON assembly was run on Amazon Web Service Tokyo region using r3.8xlarge spot instances as compute node, number of instances varies from 12~20 depending on availability.

Assembly with FALCON (v.0.3.0) yielded a 3.42 Gb assembly (phaCin_unsw_v4.1) with an N50 contig size of 11.6 Mb (Supplementary Table 2). Approximately 30-fold coverage of Illumina short reads, were used to polish the assembly with Pilon [2]. The phaCin_unsw_v4.1 version includes the primary contigs (homozygous regions of the genome), comprising 1,906 contigs. The 5,525 alternate contigs assembled by FALCON represent heterozygous regions ranging in size from 345 bp – 280 kb, with an N50 of 48.8 kb. After filtering low-quality and duplicate reads, approximately 57.3-fold read coverage (of the estimated 3.5 Gb genome from imputed genome size).

"Pacific Chocolate" DISCOVAR-BioNano *de novo* assembly (phaCin_tgac_v2.0)

Five libraries of PCR-free 250 bp paired-end reads were produced on an Illumina HiSeq 2500. Each library gave approximately 20x coverage for the reported 3.5 Gb genome of koala. Using Discovar [1], five different *de novo* assemblies were created using 21x, 43x, 64x, 85x and 107x coverage and assembly statistics were established for each assembly (Supplementary Table 1). The best overall contiguity was obtained from the 64x coverage assembly and hence was used for all subsequent analyses. 64x coverage is close to Discovar's recommended coverage of "about 60x". Higher coverage does not necessarily give a better assembly as this introduces more errors which add complexity to the assembly graph and confounds the assembler. The assembly has an estimated genome size of 3.165 Gb (3.104 Gb using 10 kb+ contigs) and an N50 of 308 kb (see Supplementary Table 2). We calculated the number of shared and unique k-mers between the sequenced reads and resulting assembly using KAT comp [3].

BUSCO v2 was used to search for single copy orthologs in both the DISCOVAR-only (phaCin_tgac_1.0) assembly and also with the addition of BioNano data (phacCin_tgac_v2.0). In the DISCOVAR-only assembly 2,377 orthologs were found; 3,569 complete, 287 partial and 32 of these were found as duplicates. Adding BioNano data increased the number of complete orthologs by 26 and reduced the number of partial orthologs by 12. The FALCON de novo assembly of long reads (phaCin_unsw_4.1) recovered the higest number (3718, 95.1%) of complete or partial orthologs among the assemblies compared (Supplementary Table 3).

2 Koala Genome analysis

2.1 Scaffold construction and virtual assignment to chromosomes

The method of super-scaffolding is described by [4]. Briefly, tables of homologous genes were generated using the physical order of genes on the chromosomes of gray short-tailed opossum and tammar wallaby as references and koala phaCin_unsw_v4.1 (Bilbo) as target (Supplementary Table 4). Koala contigs that abutted with respect to gene order with their junction spanned by a contiguous series of consecutive genes in gray short-tailed opossum and tammar wallaby, were joined to form putative derived secondary contigs (Perl script modified from [4]). Orphaned terminal genes were skipped when joining contigs because where they linked in both directions between the terminal genes of two other scaffolds and the opposite terminal to the orphan (on the same contig) also supported joining, the chaining of contigs forked. Contigs with single genes were retained and secondary contigs generated by consensus were assembled into putative super-contigs (Perl script modified from [4]).

The koala chromosome complement differs from the ancestral-type 2n=14 karyotype observed in Southern hairy-nosed wombat (*Lasiorhinus latifrons*) by a simple fission of chromosome 2. We assigned super-contigs to koala chromosomes based on predictions from cross-species chromosome painting on the wombats [5], and other species with an ancestral-type karyotype [6]. We compared the sequence virtually assigned to koala chromosomes to the ancestral karyotype of Tasmanian devil [7] and gray short-tailed opossum [8], (Supplementary Table 4). Most chromosomes have a comparable amount of sequence assigned to them compared to Tasmanian devil and gray short-tailed opossum. Regions rearranged between marsupials will be under-represented on koala chromosomes because the super-contig approach focuses on regions with conserved synteny, explaining the ~400 Mb discrepancy between the total amount of sequence virtually assigned to koala chromosomes and 3.42 Gb genome assembly.

2.2 Koala Centromeric and Telomeric Regions

To characterize the centromeric regions of the genome, chromatin immunoprecipitation (ChIP) was performed on koala endometrium tissue using the Invitrogen MAGnify Chromatin Immunoprecipitation System (Revision 6). Following tissue homogenization by mortar and pestle while submerged in liquid nitrogen, 10 ug of custom Meu anti-CenpA (rabbit) [9] or human anti-centromere protein IgG (Antibodies Incorporated) and 1 ug of negative control antibody (Invitrogen Rabbit IgG or Rb pAb to human IgG, Abcam ab2410, respectively) was employed in ChIP samples. Antibodies were coupled to beads for 6 hours at 4˚C. Chromatin was sheared using the Covaris S2 from a total of 100,000 cells per IP. Following overnight IP and DNA isolation, quantification was performed on all samples using QuBit. No detectable DNA was obtained from negative control antibody IPs. Libraries were constructed from 1 ng of DNA from CENP-A IP, CREST IP, using the Swift Biosciences Accel NGS 2S Plus DNA Library Kit, using the recommended nine amplification cycles. Illumina libraries were pooled and run on Illumina NextSeq 500 using the Mid Output Flow Cell Cartridge v2 (paired-end 300 cycles).

Reads were adapter clipped and trimmed with Trimmomatic 0.36 PE [10], and then mapped to the koala genome with Bowtie2 [11] using the "very sensitive, paired-end" parameters. Peaks were called by MACS2 (v 2.0.10.20131216) using the following parameters: –broad –keep-dup 2 –B –q 0.01 (Supplementary Tables 7-10). Surviving paired read files were converted to FASTA files, which were separated into smaller files of 1 million reads each. A single 1 million read FASTA file from each pair was repeat masked with *RepeatMasker* 4.0.3 [12] using the marsupial database and koala *de novo* database. Repeat class and type was summarized using buildSummary.pl in *RepeatMasker*. The number of reads for each repeat was normalized to the total number of repeats detected, to obtain a frequency of detection for each repeat type.

Repeat content of the centromeric regions was determined using RepBase annotated marsupial repeats and output from RepeatModeller analysis of koala. *RepeatMasker* was used to locate repeats. Candidate centromeric segments were identified using two sliding window analyses, with a window size of 200 kb and 20 kb and a step size of 100 kb and 10 kb, respectively. These regions were clustered using the heatmap.2 function in R (v 3.2.5) [13] package gplots (v 3.0.1), and a high density of ChIP-seq peaks from CENP-A and CREST were identified through manual curation of clusters of regions with similar peak densities. The repeats of these candidate regions were analyzed for biases between regions of interest and the remainder of the genome at the species, family and class level for: divergence from the model, total fraction of the bases in regions, frequency of repeat in regions and completeness of repeat. *RepeatMasker* output was converted for use with bedtools (v 2.25.0) [14]. Each candidate region was compared against the remaining candidates and the full genome (as the background) to separate candidate regions with centromeric characteristic repeats from background regions. Candidate regions were compared with background regions, using the Kolmogorov–Smirnov test (implemented using ks.test in R) [13] and Anderson-Darling test (implemented in kSamples package v 1.2-4 in R) for each metric to identify repeats with significant difference between the foreground and background regions. Using the average divergence from the repeat models and number of bases belonging to each repeat, the similarity of candidate region was visualized using multidimensional scaling (ggbiplot v0.55) and clustered using heatmaps (gplots v 3.0.1).

Centromeres are multi-megabase regions in eutherian genome assemblies with intractable higher order arrays of satellites (e.g. human and mouse) [15], but appear to be smaller in the tammar wallaby (*Notamacropus eugenii*) [9]. ChIP-seq using centromeric antibodies (CENP-A and CREST) [16] enabled identification of known and novel repeats within koala centromeric domains (Supplementary Table 6; also Supplementary Tables 7-10). Moreover, scanning the genome for CENP-A and CREST peak enrichment allowed for the identification of scaffolds containing putative centromeric regions (Supplementary Fig. 2). The centromeres of the koala appear to span smaller genomic regions; the total non-overlapping peaks called for both CENP-A and CREST ChIP encompass a total of 2.6 Mb of the koala haploid genome, equivalent to an average of 300 kb of centromeric material per chromosome. These peaks are likely to include CENP-A chromatin, and thus the functional centromere core, as well as pericentromeric DNA that binds to other centromere proteins, such as CENP-B. Like other species with small centromeres [9,15,17,18], koala centromeres lack higher order satellite arrays but are characterized by a distinguishing repertoire of repeats within CENP-A chromatin (Supplementary Tables 7-10). Previously annotated repeats found in ChIP peaks significantly differed from the whole

genome based on sequence divergence (unadjusted AD p-value <0.001) and included 5S, L2, L3, MIR3, RTE-1, UCON40, and LTR200 elements. Among the novel repeats, nine were found to have a significantly different pattern of divergence from the novel model within the ChIP peaks (unadjusted AD p-value <0.001). These include four known element classes (Mir3, LTR200, hATN1, and RTE1 elements), one novel element, and four composite elements that appear to be insertions of reverse transcribed element multimers. Similar composite elements have been recently described in gibbon centromeres, in which absence of higher order satellite arrays accompanied the evolution of novel composite elements with putative centromere function. Thus, the koala centromere composition supports mounting evidence that transposable elements represent a major, functional component of small centromeres when higher order satellite arrays are absent [9,18,19].

A scan of the genome for the canonical telomere repeat (TTAGGG)n revealed only a single scaffold enrichment for this repeat. Approximately 3.5 kb of one end of phaCin_unsw_v4.1.fa.scaf00239 consists of a long stretch of tandemly arrayed (TTAGGG) units. Unfortunately, this scaffold was not assigned to a chromosome, rendering validation via terminal location impossible. The lack of canonical telomeric repeat arrays is either an indication of incomplete assembly across repeat-dense regions, typical of mammalian genome assembly attempts, or that the koala telomeric regions carry unique and unknown sequence repeats.

ChIP-seq using the anti-CENP-A antibody derived from tammar wallaby showed a similar composition of repeats to the ChIP-seq data for human anti-centromere proteins (i.e. CREST sera) (Supplementary Fig. 2). While some enrichment for specific repeats was observed for the CENP-A and CREST IP DNA pools (namely LINEs, SINEs and LTRs), no single distinguishing element could be identified that delineated centromeric domains in the koala. To further assess sequence specificity or centromere-specific divergence among classes of repeats, further analyses among all repeats annotated were performed.

The CENP-A peaks were combined and 5 kb of flanking sequence was added to each peak. When clustering regions as either ChIP region or whole genome, sequence divergence from the repbase model appeared to be more accurate when grouping the regions compared to the length of an individual repeat normalized to the region of DNA that it was found within. (Supplementary Fig. 2, Supplementary Tables 7-10). Further, the repeats in ChIP defined regions appear to have a lower percent sequence divergence from the repeat model when compared to the remaining genome (Supplementary Fig. 2). Among the known repeats, ChIP regions significantly differed from the whole genome based on sequence divergence from the model (unadjusted AD p-value <0.001): 5S, L2, L3, MIR3, RTE-1, UCON40, and LTR200 elements. Among the novel repeats, nine repeats were found to have a significantly different pattern of divergence from the novel model within the ChIP peaks (unadjusted AD p-value <0.001). These include four known element classes (Mir3, LTR200, hATN1, and RTE1 elements), one novel element, and four composite elements that appear to be insertions of reverse transcribed element multimers. The four composite elements are combinations of MAR1-MIR3, L2b, HAL1, and ERVs, with signatures of reverse transcribed polyA tails separating each subcomponent within a single repeat unit.

A scan of genome scaffolds for enriched CENP-A and CREST peaks along a 200 kb sliding window indicated only a small number of scaffolds contained a high density of CENP-A or

CREST signal (scaf00036, scaf00385, scaf00439, scaf00170, scaf00255, scaf00337, and scaf00433). A scan of genome scaffolds for peak enrichment using the previously identified novel repeats that defined the combined CENP-A and CREST peaks identified only one region that contained a repeat that was statistically different from the whole genome (COMPOSITE C, unadjusted AD p-value 8.92 E-05) at scaffold00433:0-200000.

COMPOSITE A (rnd-5_family-3614):
MAR1c_Mdo_L2b_L2b_(T)n_L2b_(T)n_MIR3_MarsA_MAR1c_Mdo_(T)n_MAR1c_Mdo_(T)n

COMPOSITE B (rnd-5_family-2308): (T)n_(T)n_ERV41_MD_I-int_ERV41_MD_I-int_(ATTGATT)n_(T)n

COMPOSITE C (rnd-3_family-8):
HAL1_Opos1_HAL1_Opos1_HAL1_Opos1_HAL1_Opos1_WSINE1a_HAL1_Opos1_(TATA)n_A-rich_(ATTTTT)n_HAL1_Opos1

COMPOSITE D (rnd-5_family-361):
L2B_ME_(T)n_L2B_ME_(T)n_L2B_ME_(T)n_L2B_ME_(T)n_L2-2_ME_(T)n_L2B_ME_(T)n_L2B_ME_(T)n

2.3 Analysis of repeat structures

Repeat content was called using *RepeatMasker* with combined RepBase libraries (v 2015-08-07) and RepeatModeller calls generated from the genome assemblies. The resulting calls were then filtered using custom python scripts (see Code Availability) to remove short fragments, and combine tandem or overlapping repeat calls. Repeats that were embedded within larger fragments were flagged as “nested” within the final output. Endogenous retroviruses (ERV) sequences were called separately using RetroTector [20] (Quality score cutoff = 500), which uses motif based methods rather than sequence similarity to identify ERVs and related retrotransposons. These sequences were then compared to the same repeat libraries above using BLASTN to determine the most closely related repeat sequence, and merged with the filtered repeat calls based on position. Overlapping repeats were again checked, and embedded repeats flagged as “nested”.

Finally, the merged repeat calls were compared with gene annotations to remove spurious matches to tRNA sites and quantify the numbers and locations of repeats (between genes, within introns, or within UTRs in the 5` and 3` untranslated regions). Repeats that were located within coding exons were removed and total genome content of repeats was calculated.

2.4 X chromosome inactivation

X chromosome inactivation (XCI) is an epigenetic process that transcriptionally silences one X chromosome in the somatic cells of females. In eutherian and marsupial mammals this silencing is mediated by independently evolved long non-coding RNAs – *XIST* [21] and *Rsx* [22], respectively.

Like *XIST*, *Rsx* is large (estimated ~27 kb in gray short-tailed opossum), repeat-rich and predicted to form stem-loop structures. The large repetitive regions of *Rsx* have resulted in its incomplete assembly in currently available marsupial genomes [8,23].

Small tandem repeats were discovered in koala *RSX* sequence using the Tandem Repeat Finder program [24], using +2, -3, and -7 as scores for match, mismatch and gap opening respectively. Alignments of consensus repeat units with the *RSX* sequence were processed to obtain nucleotide frequency at each position. Sequence logos were generated with the SeqLogo package [25], in BioConductor. Secondary structure of the conserved sequence between gray short-tailed opossum and koala *RSX* repeat unit was determined using RNAfold [26] (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi).

Here, long PacBio reads permitted complete assembly of koala *Rsx*. Koala RSX was located on scaffold196, on which gene order was completely conserved with gray short-tailed opossum. This scaffold contains many genes located on the X chromosome in humans and maps to the X chromosome in gray short-tailed opossum; the marsupial X is completely conserved between all marsupial species examined by gene mapping and/or chromosome painting [27]. Thus it is likely to be borne on the X in koala, consistent with a function in X chromosome inactivation. It was predicted to be 30,170 bp long with 4 exons, within a genomic context conserved with gray short-tailed opossum *Rsx* (Supplementary Fig. 4). As expected koala *Rsx* was expressed in all female tissues, but no male tissues [28].

Koala *Rsx* has a complex internal repeat structure. We detected at least three large repeat arrays (Supplementary Fig. 4), which were also detected in gray short-tailed opossum *Rsx* (Supplementary Note Fig. 1). Within the ~12 kb 5' repeat array there was a 33mer with 342 copies. The consensus sequence of this 33mer shared 70.5% sequence identity with the 34mer previously identified in the gray short-tailed opossum *Rsx* 5' repeat array (Supplementary Fig. 4) [22]. The complete conservation of palindromes within the repeat suggests that they are important for function – indeed this region of both the koala and gray short-tailed opossum repeat are predicted to form almost identical stem-loops (Supplementary Fig. 4).

3 Koala Genome Annotation

BUSCO analysis on the draft assembly, was run against the mammalian ortholog database with the --long parameter on all genomes under comparison. This initial analysis showed the assembly only reached about 60% of genome completeness, suggesting a high number of indels in the draft genome. The genome polishing tool, Pilon [2], was employed to improve draft assembly from FALCON. About 30x of 150 bp paired-end Illumina X Ten short reads from "Bilbo" was used as an input for this polishing process, which was run on a compute cluster provided by Intersect Australia Limited.

Annotations were generated using the automated genome annotation pipeline MAKER [29,30]. We masked repeats in the assembly by providing MAKER with a koala-specific repeat library generated with RepeatModeler [31], against which *RepeatMasker* (v 4.0.3) [12] queried genomic contigs. Gene annotations were made using a protein database combining the Uniprot/Swiss-Prot [32], protein database and all sequences for human (*Homo sapiens*), gray short-tailed

opossum (*Monodelphis domestica;* henceforth opossum), Tasmanian devil (*Sarcophilus harrisii*) and tammar wallaby (*Notamacropus eugenii*; henceforth wallaby) from the NCBI protein database [33], and a curated set of marsupial and monotreme immune genes [34]. We downloaded all published koala mRNAseq reads from SRA (PRJNA230900, PRJNA327021) and reassembled *de novo* male, female and mammary transcriptomes using the default parameters of Trinity v 2.3.2 [35]. Each assembly was filtered such that contigs accounting for 90% of mapped reads were passed to MAKER as homologous transcript evidence. *Ab initio* gene predictions were made using the programs SNAP [36], Genemark [37], and Augustus [38]. Three iterative runs of MAKER were used to produce the final gene set.

The "Pacific chocolate" assembly was annotated using a modified version of the Oyster River Protocol [39] as follows: Transfuse (https://github.com/cboursnell/transfuse) was used to create a merged set of transcripts from Trinity (v 2.0.6) assembled RNAseq data from adrenal, kidney, spleen, brain, liver, uterus, heart and lung. Transfuse uses TransRate [40] to assess the quality of *de novo* transcripts and merge multiple transcriptome assemblies into a single high quality transcriptome. In addition, we used the dammit! software (https://github.com/camillescott/dammit), which uses open-source, publicly-available databases (Pfam-A, Rfam, OrthoDB, BUSCO and uniref90) to accurately annotate *de novo* assembled transcripts. We then used GMAP for mapping and aligning the dammit-annotated transcripts to the assembly. MAKER was also used to annotate the "Pacific Chocolate" assembly as described above for "Bilbo" assembly.

3.1 Gene familes

Gene families were called using NCBI Blast (2.3.0) OrthoMCL (2.0.9, [41]). The protein sequences of genes belonging to orthogroups identified by OrthoMCL were aligned using MAFFT (7.2.71, [42]) and the gene tree was inferred using TreeBest (1.9.2, [43]) providing a species tree to guide the phylogenetic reconstruction. Custom scripts were applied to identify families with expansion within the koala, Diprotodontia, Australidelphia and marsupial lineages. Custom scripts are available at GitHub https://github.com/DrRebeccaJ/KoalaGenome.

3.2 Characterization and expression of cytochrome P450 genes

CYP2 gene characterization was carried out as per methodology presented in section 3.1 above. In addition, through manual curation and analysis of available expression data we further refined the *CYP2C* gene models in koala, identifying three previously unannotated loci (Locus_88, Locus_7815 and Locus_9223). Expression of these 31 *CYP2* genes was carried out using the 15 koala transcriptomes (Supplementary Fig. 6).

3.3 Sequence evolution of cytochrome P450 genes

We tested for evidence of selection across cytochrome P450 genes using a multispecies alignment (n = 152 sequences) in HyPhy [44], hosted by the datamonkey webserver [45]. The multispecies alignment was reduced to only those sites with data in koala and at least one other

species (n = 498 codons; Supplementary Data 1). The substitution model was selected using the model selection tool CodonTest [46]. The genetic algorithm recombination detection method [47] found evidence of recombination and topological incongruence with a breakpoint at nucleotide position 761; data were thus analyzed in two partitions. Sites under positive selection across the tree were identified, and normalized dN-dS for each codon evaluated, using the Single Likelihood Ancestral Counting (SLAC) method [48]. The Mixed Effects Model of Episodic diversifying selection (MEME) tested for episodic selection, and identified codons potentially under positive selection in only a part of the tree, while under purifying selection elsewhere [49]. A significance threshold of α = 0.1 was used, as both SLAC and MEME are conservative, relative to related methods.

To test whether koala-specific lineages showed greater evidence of selection at particular codons, relative to non-koala lineages (31 koala tips and 121 non-koala tips) Empirical Bayes Factor (EBF) across each node and branch of the tree, at each codon were examined. When comparing koala to non-koala lineages the natural log was taken of all EBF values > 0. Codons were considered to show a unique pattern of selection (at α = 0.05) when the mean EBF for koala-specific tips fell outside the 95% CI for non-koala tips. Differences among codons on koala-specific sequences were examined by plotting the EBF of statistically significant codons for each gene, as well as mean values and standard error. The latter analysis was conducted on the natural scale, and therefore included all data (N = 70 sites).

To determine whether different koala cytochrome P450 genes showed, on average, different levels of episodic selection we examined the mean and 95% CI of evidence for episodic selection, across codons, per gene. We took advantage of 15 publicly available RNA-Seq libraries (PRJNA230900) to assess the expression of the cytochrome P450 genes across 11 different tissues (heart, liver, testis, uterus, kidney, spleen, brain, lung, adrenal gland, lymph, salivary gland). The reads were trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), and mapped to the koala transcriptome using Kallisto (PMID: 27043002). The expression values were normalized across libraries using Sleuth (http://biorxiv.org/content/early/2016/06/10/058164).

3.4 Sequence evolution of vomeronasal receptor genes

We identified all orthogroups (Supplementary Data 2) that were expanded in koala (Supplementary Fig. 5). An expansion was defined as at least 5 genes per group in koala and 2 or fewer genes in the other species in the orthogroups analysis (Tasmanian devil, gray short-tailed opossum, tammar wallaby, human, mouse, dog, platypus, chicken). Orthogroups were then characterized into gene families and assigned putative functions on the basis of sequence similarity. Two CYP450 orthogroups were identified in this manner (discussed further in 3.3) as well as one orthogroup of vomeronasal 1 receptor type 1 genes. This vomereonasal 1 receptor group contains an expansion of 6 koala genes, compared to one in both Tasmanian devil and gray short-tailed opossum, and none in wallaby/human/mouse/dog/platypus/chicken. We performed Clustal-codon alignment of these eight sequences (total 2,757 bp, 919 codons; Supplementary Note Fig. 2).

Tests for sequence evolution proceeded as for the CYP450 analysis. GARD analysis did not reach final convergence due to computational limitations, potentially due to either a large number of breakpoints in the alignment or heterotachy. Sequence analysis results were broadly similar whether or not the GARD results were accounted for. Mean dN/dS (SLAC method) = 0.563, 22 sites show significant evidence of negative selection at α = 0.1, no sites show pervasive positive selection. Evaluation of episodic positive selection with MEME revealed 31 sites with evidence of episodic selection at α = 0.1.

3.5 Analysis of Olfactory Receptor (OR) genes

All scripts used for these analyses are available from https://github.com/patelhardip/orgfinder.

We discovered 1,169 olfactory receptor (OR) genes in the koala genome assembly, of which 580 (49.6%) are full-length OR genes without any in-frame interruption (also labelled as functional OR genes in the literature as opposed to pseudogenes or truncated genes) (Supplementary Note Fig. 3). The total OR repertoire in koala is smaller compared to gray short-tailed opossum (1,431 genes), tammar wallaby (1,660 genes) and the Tasmanian devil (1,279 genes). Amongst marsupials, gray short-tailed opossum and Tasmanian devil genomes contain 1,106 (77.3%) and 913 (71.4%) functional OR genes; more than that of the koala. However, the tammar wallaby genome has least number of functional OR genes and proportionally it accounts for only 33.0% of total OR genes in marsupial. This may be attributed to the poor quality of the tammar wallaby genome assembly though.

OR genes can be classified in to two classes; class I (families 51-56) and class II (families 1-13) [50]. Despite specialization in diet and nocturnal life style, there is little difference in the OR repertoire of koala compared to other marsupials (Supplementary Note Fig. 3). For example, we discovered OR genes from both classes and all families in the koala genome. Perhaps, OR genes don't have a large role to play in discerning food or lifestyle choices in marsupials.

We downloaded 107,803 olfactory receptor CDS sequences in FASTA format from the NCBI Nucleotide Database by using '"olfactory receptor"[All Fields] AND 'Chordata[Organism] AND (biomol_mrna[PROP] AND refseq[filter])' as the search term on 22nd May 2015 along with corresponding "Feature Table". These sequences were filtered to obtain full-length OR sequences with uninterrupted open reading frame using extractCDS.pl script. Following filter steps were performed sequentially:

1) 4,862 CDS without start codon at the beginning and stop codon at the end of the translation were removed.

2) 26,092 sequences without "olfactory receptor â€œ term in the FASTA header description were removed.

3) 3,991 sequences containing non-ACGT characters were removed.

4) 1,672 sequences with non-amino acid characters were removed.

5) 4,635 sequences shorter than 300aa length were removed.

8) 5,679 sequences longer than 330aa were removed.

All-vsl-all BLASTP (version 2.2.30+, -evalue 1e-5, [51]) was performed using 60,872 remaining OR sequences were retained for further analysis as a result of the filter steps above. BLASTP alignments were processed to remove alignments to the self, alignments with percent identity <=50 and >=95 resulting in 27,858,657 pair-wise alignments of 60,842 OR amino acid sequences. Pair-wise sparse matrix of alignments between two OR amino acid sequences was created by using negative log10(e-value) as the distance measure. This sparse matrix was processed using MCL algorithm [52] to find clusters of similar OR genes based on the e-value of their pair-wise alignments. We identified 102 clusters containing 2 to 4921 sequences within any given cluster. We selected two amino acid sequences (chosen randomly when more than 2 sequences present in a cluster) from each cluster and performed multiple sequence alignment by using the Muscle software (v3.8.31, default parameters, [53]). Subsequently, Muscle was used to iteratively generate profile alignment of all 102 multiple sequence alignments. Amino-acid sequence alignment of 204 sequences were back translated to nucleotide alignments and HMM profile was build using the HMMER suit (version 3.1b1 May 2013, [54]). This HMM profile was subsequently used to scan reference genome sequences for identification of putative OR genes as described below.

We chose genomes of 14 tetrapod species along with the Koala genome to identify and compare OR gene family repertoire. We used nhmmer program [55] using the HMM profile for OR genes to identify candidate OR region in each of the genome assembly. Candidate regions of OR HMM hits along with 500 nt flanking sequences were isolated and these sequences were aligned back to the 204 reference OR sequences using fasty tool from FASTASEARCH suit (version 36.8.8, [56]) to obtain conceptual translations. All fasty alignments were processed to identify open reading frames. Susequently, false positive sequences were removed by aligning all discovered candidate OR gene sequences to HMM database containing OR profile and other rhodopsin gene family profiles. If the best hit of a sequence was not the OR HMM then it was discarded as false positive and removed from further analysis.

3.6 Analysis of Aquaporins

Aquaporins are the proteins responsible for the transport of water and other small molecules across cell membranes and have numerous roles in organisms' water balance [57]. The koala aquaporins are preferentially expressed in various tissues in accordance with their known or presumptive functions in other mammals [57]. For example, expression levels of AQP2 and AQP6 suggest that they have major roles in water conservation in the kidney. The koala genome contains one gene for each of the aquaporin types in other marsupials and eutherian mammals, except that only one gene is known for AQP12 (as for other marsupials) rather than the two in eutherian mammals and that there are two genes for AQP5 and AQP7 (Supplementary Table 14).

AQP5 has been suggested to be a central component in the ability of other mammals to sense water concentration [58-60]. It is highly expressed in the oral cavity, particularly in the tongue, vomeronasal organ and salivary gland. The synteny of one AQP5 locus (PCI4_00018408-RA) with AQP2 and AQP6 found in other marsupials is preserved in koala. There is no recognized

homologue of the second locus (PCI4_00002921-RA) in the three other marsupials with genomic sequence data but it is highly similar to the homologues of PCI4_00018408-RA, having 90.59% identity to the a fragment of 797 bases of the AQP5 gene of *Sarcophilus harrisii*, compared to 94.35% identity of the PCI4_00018408-RA gene with this fragment. Both AQP5 genes are expressed principally in the salivary gland and lung, with PCI4_00002921-RA being expressed at about 10% of the level of the PCI4_00018409-RA gene. Elucidating the functioning of AQP5 tissue specificity might significantly advance the behavioral and physiological genomics of koala.

3.7 Taste receptor genes

The G protein-coupled receptor (GPCR) families of TAS1Rs (taste receptors type 1, the umami and sweet taste receptors) and TAS2Rs (taste receptors type 2, the bitter taste receptors) are expressed on the mammalian oral cavity and recognize taste compounds in diet. The *TAS1R* family consists of only 3 genes, encoding heteromeric complexes of TAS1R1/TAS1R3 (functioning as the umami taste receptors for amino acid and nucleotide perception) and TAS1R2/TAS1R3 (the sweet taste receptor for sugar perception). On the other hand, the *TAS2R* family consists of several dozens of genes and varies in number among species. TAS2Rs respond to various structural toxins, mainly included in herbivorous diet as plant secondary metabolites (PSMs) [61]. Taste receptor gene repertoires are strongly linked to the host diet. For example, *TAS1R1* of the bamboo-eating giant and red pandas and *TAS1R2* of the carnivorous felids such as cats were pseudogenized possibly because they did not require detection of umami (amino acids) or sweet (sugar) compounds in their specialized diet [62,63]. Larger numbers of *TAS2R*s in the genome were observed in species which depend on plant matter rather than animal matter possibly because an herbivorous diet includes a high amount of toxic secondary metabolites, which should be detected as bitterness [64,65]. Koala taste might also show specific adaptation to its obligate diet of toxic eucalyptus leaves, which includes toxic PSMs.

The umami and sweet taste receptor genes (*TAS1R1*, *-2* and *-3*) were identified in the whole genome assemblies of 4 marsupial species (koala *Phascolarctos cinereus*, this study; tammar wallaby *Macropus eugenii*, GCA_000004035.1; Tasmanian devil *Sarcophilus harrisii*, GCA_000189315.1; gray short-tailed opossum *Monodelphis domestica*, GCA_000002295.1) using the BLAST-based method [66]. Mapping of Illumina HiSeq XTen reads of koala “Bilbo” against the koala PacBio assembly (derived from the same individual “Bilbo”) was confirmed to validate the presence of PacBio sequencing and assembly errors, visualized by Integrative Genomics Viewer (IGV) v2.3.97 [67]. Trace BLAST was performed to find Sanger sequencing chromatograms corresponding to *TAS1R*s of non-koala marsupial whole genome sequences. As a result, all exons except some truncations (missing in the data) were detected in the 4 marsupials (Fig. 2, Supplementary Fig. 7 and Supplementary Table 15). Frameshifts before the premature stop codon were detected in koala *TAS1R1* and *TAS1R2* and tammar wallaby *TAS1R1* and *TAS1R3*, but these were caused by assembly error (Supplementary Fig. 7). Unlike in the case of giant and red pandas, the obligately herbivorous diet of koala and tammar wallaby has not led to pseudogenization of *TAS1R*s.

Identification and annotation of the bitter taste receptor genes (*TAS2R*s) were performed in these marsupial assemblies using the BLAST-based method and all intact Euarchontoglires *TAS2R*s as BLAST queries [65,66]. The assembly of a monotreme species, platypus (*Ornithorhynchus anatinus*, GCF_000002275.2) was also analyzed as an outgroup of the marsupials. To determine orthologous and paralogous relationships among all the annotated *TAS2R*s, gene trees were constructed with *TAS2R* representatives of orthologous gene groups (OGGs) of Euarchontoglires [65]. Multiple alignments of nucleotide or amino acid sequences were constructed using E-INS-i in MAFFT v6.857b [42]. The neighbor-joining trees of *TAS2R*s in nucleotides were inferred with the Jukes-Cantor correction using MEGA v7.0.18 [68]. The best-scoring maximum-likelihood tree of *TAS2R*s in amino acids were inferred using RAxML v8.2.11 under the GTR + Γ model with 1000 bootstrap replicates [69]. Amino acid sequences of pseudogenes were deduced after their disrupted sites were excluded based on multiple alignment of nucleotide sequences of phylogenetically-close intact *TAS2R*s. Gene trees were visualized using MEGA [68]. As a result, the number of intact *TAS2R*s in koala (24) was the largest in the analyzed Australian marsupials (tammar wallaby, 16; Tasmanian devil, 19) (Supplementary Table 15) and positioned in a comparatively large repertoire in mammals [64,65]. Gray short-tailed opossum (27) also has a relatively large number and platypus (6) a smaller number of *TAS2R*s possibly due to their specific evolutionary results. The gene tree of all annotated *TAS2R*s showed 27 OGGs in marsupial *TAS2R*s (*TAS2R1*, *-2*, *-4*, *-38*, *-41*, *-60*, *-701-705*, *-710-714* and *-720-730*), where gene symbols of *TAS2R700*s were newly numbered in this study, indicating that the last common ancestor of marsupials had at least 27 *TAS2R*s (Fig. 2; Supplementary Table 15). In comparison with eutherian *TAS2R*s, 3 marsupial-specific expansion events prior to the establishment of marsupial OGGs occurred (the marsupial clusters I, II and III). Marsupial cluster I was orthologous to eutherian *TAS2R3*, whereas marsupial clusters II and III had no clear eutherian orthologs and are positioned in a supercluster including various eutherian *TAS2R*s in addition to *TAS2R704* and *TAS2R725*. Although orthologous and paralogous relationships between koala and tammar wallaby were not necessarily clear due to lack of reliable bootstrap support, massive gene duplication (producing more than 2 koala duplicates) was observed in *TAS2R41*, *-705* (the ortholog of eutherian *TAS2R16*/*TAS2R62*), *-710* (one of the orthologs of eutherian *TAS2R3*) and *-720* groups, contributing to the comparatively large repertoire of koala *TAS2R*s (Fig. 2b-e).

TAS2Rs have a major role of rejection of plant diet including toxic PSMs such as terpenes, phenolics and glycosides, broadly distributed in eucalyptus leaves [70,71]. Cell-based assays reported that human TAS2R16 and mouse TAS2R41 (Tas2r126 in mouse nomenclature) were activated by β-D-glucopyranosides, which characterizes structure of cyanogenic glycosides [61,72]. Although TAS2R orthologs among species do not necessarily have similar agonist profiles [72], the presence of marsupial orthologs of eutherian genes such as *TAS2R16* and *TAS2R41*, and their massive gene expansion might enable koala to reject a high amount of toxic PSMs of eucalyptus diet.

3.8 Genes involved in Development and Reproduction

Orthologues of the vast majority of mouse early developmental genes involved in determining the three germ layers of the early post-conception development were identified in the koala

genome (as in the tammar genome), including those encoding key transcription factor genes such as *POU5F1*, *SOX2*, *NANOG*, *CDX2*, *EOMES*, *GATA4*, *GATA6* and *Brachyury* (*T*) [23]. Genes encoding components of key signalling pathways in early development are largely conserved in mammals, but the platypus and marsupials including the koala additionally have *POU2,* an ancient vertebrate paralogue of *POU5F1* that is not found in eutherian genomes [73]. It will be interesting to determine whether this gene is expressed in a broad range of adult tissues as well as in in pluripotent tissues. Germ cells also express pluripotent genes, and orthologues of genes critical for the specification and development of eutherian germ cells including *DDX4*, *BMP4*, *PRDM1* and *PRDM14* were identified. Interestingly, *STELLA*, a gene and protein critical for germ cell imprinting (located in eutherian genomes between *NANOG* and *GDF3*) does not seem to be present in marsupial genomes [23], including that of the koala.

Ovarian and testicular-specific genes important for gonadal differentiation are conserved in the koala, as in the tammar. *DHH* is a mammal-specific gonadal development gene, and it and the key downstream sexual differentiation genes *SOX9, ATRX, FOXL2, WNT4,* and *RSPO1* are all present in the koala genome. The developing koala fetus depends primarily on the yolk sac placenta for nutrition, but in addition there is a brief period of contact of the embryonic allantois to the uterine endometrium. *IGF2*, *INS* and *PEG10* are imprinted in the tammar placenta and play key roles in early development [74,75]. These genes are present in the koala genome, but their methylation status has not yet been investigated.

Marsupials differ from eutherian mammals primarily in their unique mode of reproduction, in which young are born in an altricial state and then nourished during a period of extended lactation. The koala is a particularly interesting marsupial from a reproductive point of view because it is an induced ovulator [76]. The genome contains a range of genes controlling ovulation on the female side, including *LHB*, *FSHB*, *ERR1* and *ERR2*, and genes encoding prostaglandin synthesis enzymes including *PTGS1*, *PTGS2*, and *PTGES*. We also searched for genes involved in the induction of ovulation from the male side. Ovulation-inducing factor is a component of seminal plasma in other induced ovulators (e.g. llamas, camels), that triggers ovulation in mated females. This factor was subsequently shown to be nerve growth factor (NGF), produced in the seminal vesicles and prostate [77] and present in the seminal plasma of a wide range of species, including spontaneous ovulators [78]. NGF is highly conserved, and when applied cross-species has strong effects on ovarian activity even in species that are not induced ovulators [79]. The *NGF* gene is present in the koala genome, and highly conserved, strongly suggesting a role for prostate-expressed NGF in the induction of ovulation, although functional testing is required to confirm this hypothesis.

Unlike most eutherian mammals, male marsupials lack seminal vesicles and ampullary glands, and so seminal fluid is produced by the prostate and bulbourethral (Cowpers) glands [80]. Male koala ejaculates contain a fraction with coagulatory activity that may have copulatory plug function [81-85]. However, the genes encoding major copulatory plug proteins in some eutherian ejaculates are absent from the koala genome. Prostate *TGM4*, which is a pseudogene in gray short-tailed opossum [86], was not identified in the genome; nor were *SEMG1* and *SEMG2*, possibly because koalas lack seminal vesicles, where these proteins are normally produced [reviewed in [87]]. Instead, seminal fluid coagulation may be mediated in part by the production of polyamines in the prostate, which are implicated in semen coagulation in eutherians [88]. The

genes encoding the production of prostate polyamines spermine and spermidine (*ODC1*, *SAT1*, *SAT2*, *SMOX*, *SRM*, *SMS*) are present in the koala genome. Seminal fluid coagulation and copulatory plug formation is associated with the intensity of sperm competition concomitant with polyandry [89]. Female koalas may mate multiply but this is not well understood [90], and semen coagulation in the koala may either serve as a mechanism evolved in response to sperm competition, and/or to prevent sperm leakage from the female reproductive tract, which may be important in this arboreal species [83].

3.9 Genes involved in Lactation

The changing composition of marsupial milk during the growth of altricial young has been of great interest as changes in bioactive profiles stimulate development and protect young from infection [91]. Key milk proteins involved in nutrition and transport such as caseins and lipocalins such as beta-lactoglobulin, alpha-lactalbumin and lactotransferrin that are found in all mammals, were identified in the koala. As for other marsupials, no duplications are seen within the three koala casein families (alpha, kappa and beta); this contrasts with duplications in monotremes and eutherian mammals.

Several milk proteins are specific to marsupials. Trichosurin (TRSN) is a protein of unknown function which is expressed across lactation in possums and wallabies. Late Lactation Proteins (LLPs), thought to be important in nutrition, are highly expressed in late lactation across studied marsupials. The genomic context of these proteins in marsupials could not be determined in the previous fragmented marsupial genome assemblies. In the koala genome we show four *LLP* genes tightly linked to both Trichosurin and Beta-lactoglobulin (Supplementary Fig. 8). These four *LLP* genes may allow marsupials to fine-tune milk protein composition across the stages of lactation to meet the changing needs of their young. More broadly this region contains 6 genes closely related to Trichosurin, the milk expressed Lipocalin-2, and 11 additional lipocalin genes. It appears that this group of lipocalin genes, particularly within the LLP and Trichosurin-like groups, has expanded within the marsupial lineage by gene duplication, producing 1 to 4 LLP proteins. *LLP* genes have a complex evolutionary history with duplications and deletions, and a phylogenetic tree of the sequences (Supplementary Fig. 8) indicates that some LLP genes are conserved among Australian marsupials, while lineage specific duplications occur in several species including koala. Although the functions of these proteins have not been clearly established, these duplications may allow marsupials to fine tune milk protein composition across the stages of lactation to meet the changing needs of offspring.

Several marsupial-specific milk proteins of unknown function, including Very Early Lactation Protein (VELP) and Marsupial Milk 1 (MM1), are highly expressed in marsupial milk. We used their genomic position in the koala genome, to speculate on their probable function. Koala VELP shows homology to a eutherian antimicrobial protein known as Glycam1 or PP3 [92]. In the koala genome VELP is located in a region of shared synteny with Glycam1 (Supplementary Fig. 8) and MM1 is located close to VELP (Supplementary Fig. 8). This region in eutherians contains an array of short glycoproteins including lacritin, dermicidin, Glycam1 and Mucin-like 1, which have antimicrobial properties and are found in secretions such as milk, tears and sweat. We propose that MM1 may be related to this group of genes, and has an antimicrobial role in

marsupial milk. Three short novel genes were also seen in this region, including one that showed homology to lacritin and dermicidin. Although these proteins are highly divergent, their location and similar length and structure indicate that this group of genes may also be related to the eutherian antimicrobial peptides.

3.10 Characterization of koala immune genes

Twenty-three MHC class I and 23 class II genes and pseudogenes were identified in the PacBio assembly of koala genome. The core MHC region on scaffold #255 contains 138 genes (Supplementary Figure 9). It includes 8 class I and 11 class II loci (including 4 class I pseudogenes and 1 class II pseudogene) (Supplementary Table 12). The MHC class II beta subfamilies *DBB*, *DCB*, *DMB*, and class II alpha subfamilies *DBA, DAA, DMA* and *DCA* are all located within this core region. Twelve MHC II *DAB* subfamily genes and 15 MHC I genes are located outside this core region (located on 10 and 9 scaffolds respectively). While most of these genes are putative pseudogenes, 2 *DAB* and 7 MHC I genes appear functional with in-frame coding sequences. Among the 11 putatively functional MHC I genes, 3 are predicted to be classical while the other 8 showed characteristics of non-classical genes [93]. Without mapping it is not possible to determine whether the koala MHC I loci that are not on scaffold #255 are co-located on the same chromosome as the MHC proper, as in the gray short-tailed opossum [94], or are located on multiple other autosomes, as in the tammar wallaby [95].

The koala is the first Australian marsupial for which the TCR loci have been fully assembled and annotated (Supplementary Figure 9). The koala TCR system contains the conventional α, β, δ, and γ chains, and an additional isotype class known as the TCR μ chain (encoded by *TRM*), which is unique to marsupials and monotremes and missing in eutherian mammals. The koala TRA/D, TRB, TRG, and TRM loci are each located on a single scaffold (scaffold 1, 71, 153, and 59, respectively) in the long-read assembly. The TRA/D locus contains 94 putatively functional gene segments, including 52 V_α, 2 D_δ, 3 J_δ, 1 C_δ, 35 J_α, and 1 C_α. The TRB locus contains 33 V_β segments, and three sets of D_β, J_β, and C_β segments arranged in tandem cassettes, each comprising 1 D_β, 2 to 4 J_β, and 1 C_β. The TRG locus consists of 4 V_γ, 4 J_γ, and 1 C_γ. Four sets of TRM gene segment clusters, each consisting of 1 V, 1 or 2 D, 1 J, 1 Vj (joined VJ segment specific to TCR μ chain), and 1 C, are found in the koala; two sets are probably functional because all segments contain an intact open reading frame.

Koala IG loci comprise 36 IGHV, four IGHC (one each of M, G, E, and A), 78 IGKV, three IGKJ, one IGKC, 11 IGLV, four IGLJ and four IGLC gene segments. Compared to other mammalian IG systems, heavy chain isotype D (IgD) appears to be missing in the koala. This is consistent with previous findings in the gray short-tailed opossum, in which the IGHD gene has been lost, possibly due to the insertion of several endogenous retroviruses and long interspersed elements [96].

Additional immune genes known to play a role in resistance and susceptibility to chlamydia infection in other species were also characterized including CD4, CD8-a, CD8-b, IFN-gamma, TNF-a, IL-1B, IL-2, IL-4, IL-6, IL-10, IL-12A, IL-17A, IL-17B and IL-23A (Supplementary Table 18).

3.11 Immune gene diversity and responses to chlamydia vaccine

We resequenced eight koalas from south east Queensland on an Illumina NextSeq500 using 150 bp paired-end reads and obtaining approximately 10-fold coverage (genbank project accession number (PRJEB25952). Reads were aligned using Burrows-Wheeler aligner v 0.7.15 [97] to the phaCin_unsw_v4.1 genome as the reference and SNPs were called using SAMtools v 1.3 [98] with minimum base and mapping quality of at least 20. Candidate gene sequences (± 10 kb flanking sequence) were annotated in Geneious [99] to identify SNPs with a minimum variant frequency of 0.2 (to increase the likelihood that any SNPs detected were representative of true diversity and not sequencing errors) (Supplementary Table 19). Diversity at the target immune genes was characterized following Morris, et al. [100] (Supplementary Table 20). Diversity at the characterized MHCII genes was also evaluated due to MHCII variants previously been associated with chlamydial disease in the koala [101] (Supplementary Tables 19-20).

MHCII genes had on average 350 total SNPs and the remaining genes had on average 74 total SNPs, with ~ 25% of these being within the gene (90% intronic, 10% exonic) and ~75% being within the 10kb flanking regions (Supplementary Table 20). Number of haplotypes ranged from 1-9 (out of 9 koalas), with MHC genes having the largest numbers of haplotypes (Supplementary Table 20). Average haplotype diversity was 0.5, average number of nucleotide differences was 3.8 and average nucleotide diversity was 0.005 across the immune genes (Supplementary Table 20). Tajima's D and Z tests of selection were not significant for any of the genes (Supplementary Table 20).

A chlamydia vaccine trial [102], with approval from the Queensland University of Technology Animal Ethics Committee (#07600000559) was conducted at Lone Pine whereby five or the resequenced koalas were exposed to a vaccine (MOMP antigen plus ISC adjuvant) and their response to the vaccination was monitored. Antibody responses were quantified by ELISA and lymphocyte proliferation responses to chlamydial antigen were also assessed. The responses of the five koalas were then nominated as either case animals (strong response) or control animals (weak response). Basic case/control association tests were performed on the antibody and T-cell response data in PLINK to identify any SNPs that may be associated with differential vaccine responses. Three SNPs were found to significantly differ between the 1 case (weak) and 3 control (strong) antibody vaccine responses including: a SNP downstream of the MHCII DMA gene**,** a SNP downstream of the MHCII DMB gene and an intronic CD8-a SNP (Supplementary Table 21). No SNPs were found to be statistically significant at the 0.05 alpha level for the T-cell vaccine responses.

3.12 RNASeq analysis of koala conjunctival tissue samples

Conjunctival tissue samples were collected from 26 koalas euthanized due to injury or disease by veterinarians at Australia Zoo Wildlife Hospital, Currumbin Wildlife Hospital and Moggill Koala Hospital. The collection protocol was approved by the University of the Sunshine Coast Animal Ethics Committee (AN/S/15/36). Health assessments of the eye were performed by an experienced veterinarian and classified as either 'healthy' (n=13) or 'diseased' (n=13) based on evidence of gross pathology consistent with ocular chlamydiosis [103]. Conjunctival tissue samples

from each animal were placed directly in RNALater (Qiagen, Germany) buffer overnight at 4°C prior to storing at -80°C for later use. RNA was extracted using an RNeasy Mini Kit (Qiagen, Germany), according to the manufacturer's instructions, with an on- column DNase treatment to eliminate contaminating DNA from the sample. The concentration and quality of the isolated RNA was determined using a NanoDrop ND-1000 160 Spectrophotometer and Agilent BioAnalyser (Agilent, USA). Library construction and sequencing were performed by The Ramaciotti Centre (UNSW, Kensington, NSW) with TruSeq stranded mRNA chemistry on a NextSeq500 (Illumina, USA). Reads were mapped to the phCin_unsw_v4.1 assembly using the default parameters of STAR [104] and counts summed over features using featureCounts [105] . Differentially expressed genes were called using DESeq2 [106] as implemented in the SARTools package [107]. Reads have been deposited in the SRA under the accession BioProject PRJEB19389.

4 Koala Retrovirus (KoRV)

In the long-read assembly, KoRV insertions were divided into two groups. The first comprises 58 KoRV-only insertions, all of which belong to subgroup KoRV-A (the most prevalent KoRV subtype). Most (49/58) of these are near full-length provirus, although solo LTRs and forms with deletions are also present. The second group (15 loci) comprises a recombinant [72], formed between KoRV and an ancient retroelement present in the koala genome. KoRV terminal fragments (including both intact LTRs) remain, but most of the central, protein-coding region of KoRV has been replaced [73]. This recombinant form is very likely endogenous, and has been observed in museum koala samples dating to 19th century, suggesting it is not recent in origin [73].

Three recombinant KoRVs (recKoRVs) were mapped to hybridization capture enriched Illumina data generated from historical koalas reported in [108]. Supplementary Table 24 illustrates the number of reads mapping for each recombinant to each of the historical koalas. The first column indicates in which koala genome data set the breakpoints were identified, although the two breakpoints identified in Birkie were originally identified in Pci-SN265. Koala samples that were positive for the identified breakpoints are indicated with an X mark. All koalas in columns subsequent to Pci-SN265 are historical koala samples. The 4 recombination events represent three recombinant KoRVs. The two breakpoints found in Birkie and Pci-SN265 represent a single contiguous double recombinant KoRV.

5 Koala population genomics

5.1 Historical population size

Demographic history was inferred from the diploid sequence of each of the three koalas, using a pairwise sequential Markovian coalescent (PSMC) method [109]. We conducted a range of preliminary analyses and found that PSMC plots were not sensitive to the values chosen for the maximum number of iterations (*N*), the number of free atomic time intervals (*p*), the maximum

time to the most recent common ancestor (*t*), and the initial value of rho (*r*). Based on these investigations, our final PSMC analyses of the three genome sequences used values of *N*=25, *t*=5, *r*=1, and *p*=4+25*2+4+6. The number of atomic time intervals is similar to that recommended for analyses of modern human genomes [109], which are similar in size to the koala genomes. We determined the variance in estimates of N_e using 100 bootstrap replicates. Replicate analyses in which we varied the values of *p*, *t*, and *r* produced PSMC plots that were broadly similar to those using our chosen 'optimal' settings (Supplementary Fig. 10).

The plots of demographic history were scaled using a generation length of 7 years, corresponding to the midpoint of the range of 6 to 8 years estimated for the koala [110] and the midpoint of the estimates of the human mutation rate (1.45×10^{-8} mutations per site per generation; summarized by [111]) and mouse mutation rate (5.4×10^{-9} mutations per site per generation [112]) was applied in the absence of a mutation rate estimate for koala (Supplementary Fig. 10). The koala mutation rate is likely to be closer to that of humans, based on greater similarity in genome size, life history, and effective population size, relative to mouse [111].

5.2 Contemporary population analysis

Fifty-six koalas (49 wild, 7 captive) were sampled throughout the distribution using a hierarchical approach to allow examination of genetic relationships at a range of scales, from familial to range-wide. All individuals were sequenced using a target capture approach described in [113], with a kit targeting 2,167 marsupial exon sequences. Illumina sequence reads were quality-filtered and trimmed (see [113] for details) and mapped to the koala genome (Bowtie2, v2.2.4 [11]). A panel of 4257 SNP sites was identified (using GATK version 3.3-0-g37228af [114]) that showed expected levels of relatedness and differentiation among the sampled individuals. A set of SNP sites was identified (using GATK version 3.3-0-g37228af [114]) and showed expected levels of relatedness among individual samples. A panel of 1,200 SNPs (obtained by mapping to targets, filtering, and selecting one SNP per target) showed fine-scale regional differentiation in wild individuals consistent with evolutionary history and recent population management (Fig. 3b). Inbreeding estimates for each region were calculated using COANCESTRY [115] and linear modelling was conducted in R [116] to determine whether inbreeding levels differed significantly across regions.

5.3 Genome-wide heterozygosity

The eight resequenced koala genomes from south east Queensland were examined using PLINK v 1.07 [117] to calculate heterozygosity in each individual using sites with a minimum depth of four reads.

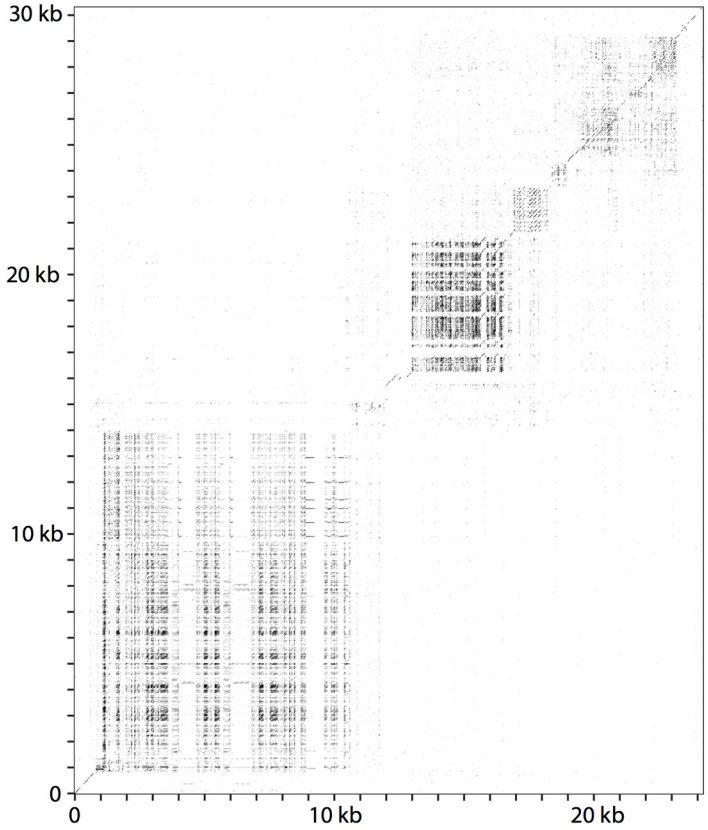
Of 5,077,810 variable sites within eight resequenced genomes, 66% were heterozygous on average across all eight samples. Considering a genome size of ~3.5 Gb, this equates to 1 SNP every 689 bases, which is a higher rate of variable sites than found in species with known low genetic diversity or suffering from the effects of population bottlenecks [118,119]. Note that samples were from south-east Queensland and genome-wide heterozygosity may be expected to be lower if samples from the southern part of the distribution were to be included (Fig. 3c).

Supplementary Note Figures

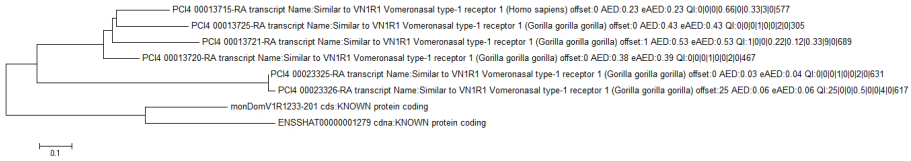
Supplementary Note Figure 1: Sequence similarity plot of koala *Rsx* to gray short-tailed opossum *Rsx*

Supplementary Note Figure 2: Neighbor-joining tree of 6 koala vomeronasal 1 receptor type 1 genes, plus 1 each from Tasmanian devil and gray short-tailed opossum

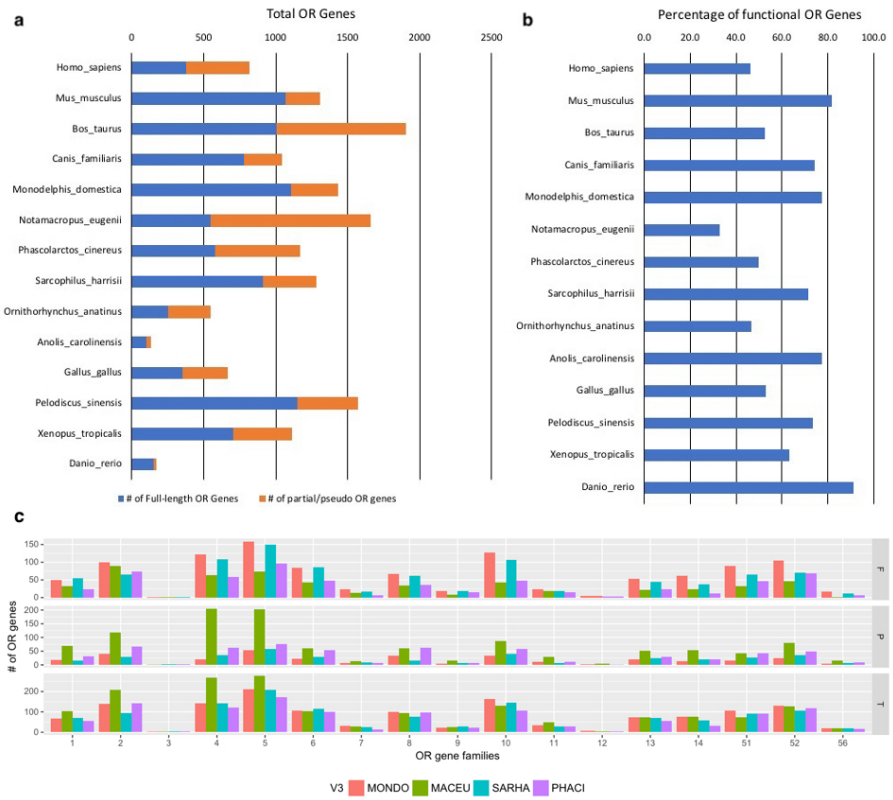
Supplementary Note Figure 3: Comparison of *OR* genes across marsupial species



Supplementary Note Figure 1: Sequence similarity plot of koala *Rsx* to gray short-tailed opossum *Rsx*. The three repeat arrays detected in koala are all present in gray short-tailed opossum.



Supplementary Note Figure 2. Neighbor-joining tree of 6 koala vomeronasal 1 receptor type 1 genes, plus 1 each from Tasmanian devil and gray short-tailed opossum. Koala transcripts have a prefix “PCI4”, opossum sequence is indicated by “monDom” and Tasmanian devil by “ENSSHAT”.



Supplementary Note Figure 3. Comparison of *OR* genes across marsupial species. a Number of full-length and partial/pseudo OR genes in the koala and other vertebrates. **b** Percentage of total OR genes that are full-length/functional in the koala and other vertebrates. **c** Number of full-length (F), partial (P) and total (T) class I (families 51-56) and class II (families 1-13) OR genes in the koala (PHACI) and other marsupials, including the gray short-tailed opossum (MONDO), the Tammar wallaby (MACEU) and the Tasmanian Devil (SARHA).

Comment [PB1]: Will make a high quality version of this figure once we hear back from Hardip.

Supplementary Tables

Supplementary Table 1: Genome Continuity of the DISCOVAR only Assembly.

Supplementary Table 2: Comparison of koala genome assembly statistics phaCin_unsw_v4.1 (long read PacBio only), phaCin_tgac_v2.0 (short-read and BioNano) and Koala v1.1 generated in this study.

Supplementary Table 3: Comparison of koala genome assembly statistics phaCin_unsw_v4.1 (long read PacBio only) and phaCin_tgac_v2.0 (short-read and BioNano) showing the greater contiguity of the PacBio assembly.

Supplementary Table 4: Assembly Completeness in BUSCO.

Supplementary Table 5: Sequence (Mb) virtually assigned to koala chromosomes compared to Tasmanian devil (ancestral karyotype used as a point of reference) and gray short-tailed opossum.

Supplementary Table 6: Mapping statistics for CENP-A, CREST and Input ChIP-seqdatasets.

Supplementary Tables 7-10: De novo and Mars summary statistics for ChIP-Seq for single koala sample using two IPs.

Supplementary Table 11: Comparison of koala repeat regions with other marsupial, eutherian and monotreme genomes.

Supplementary Table 12: SLAC and MEME codon-based output for single koala sample.

Supplementary Table 13: MEME output by-codon and by-branch for single koala sample.

Supplementary Table 14: The aquaporins of koala showing scaffold position, reading strand and the number of exons.

Supplementary Table 15: List of annotated taste receptor genes in koala and other marsupial assemblies.

Supplementary Table 16: List of annotated genes involved in koala lactation.

Supplementary Table 17: MHC, TCR, and IG genes annotated using BLAST and HMMER.

Supplementary Table 18: List of annotated koala immune genes.

Supplementary Table 19: Summary of SNPs identified in candidate immune genes.

Supplementary Table 20: Diversity statistics for the coding regions of the candidate immune genes.

Supplementary Table 21: Chlamydia vaccine immune response case/control association test results

Supplementary Table 22: Locations of KoRV loci in phaCin_unsw_v4.1 assembly. The data are in BED format (with 0-based numbering of the start field).

Supplementary Table 23: phaCin_unsw_4.1 KoRV insertions with breakpoint sequences which are also present in the Birkie genome sequence library.

Supplementary Table 24: Recombinant KoRV (recKoRV) breakpoints found in historical koala genome sequences.

Supplementary Table 25: Genes with KoRV insertions.

Supplementary Table 1 | Genome Continuity of the DISCOVAR only Assembly. From the assembly calculations, 64x coverage assembly provided the best overall contiguity (and matched Discovar's recommended 60x coverage) and hence this assembly was used for all subsequent analysis. The assembly had an estimated genome size of 3.165 Gb (3.104 Gb using 10Kb+ contigs) and an N50 of 308 Kb. Scaffold length (in kilobases) for each 'N-size' in multiples of ten between N10 and N100 (longest scaffold).

N-size	N90	N80	N70	N60	N50	N40	N30	N20	N10	Longest scaffold
Scaffold length (Kb)	71	131	186	245	308	385	484	613	835	2343

Supplementary Table 2 | Comparison of koala genome assembly statistics phaCin_unsw_v4.1 (long read PacBio only), phaCin_tgac_v2.0 (short-read and BioNano) and Koala v1.1 generated in this study

Species	Genome size (Gb)	G+C content (%)	No. scaffolds	Scaffold N50	Data type	Coverage
Koala phaCin_unsw_v4.1 (Female – Bilbo) *homozygous/ **heterozygous	3.19	39.0	1906* 5525** (contigs)	11,587,828 (contig)	Pacific Biosciences RS II platform Illumina 150bpPE HiSeq X Ten	57.3x 34x
Koala phaCin_tgac_v2.0 (Female – Pacific Chocolate)	3.60	36.8	796,464	798,273	250bp paired-end Illumina HiSeq2500 & BioNano	60x
Koala v1.1 *(Phascolarctos cinereus)* (Male – Birke)	3.09	38.9	599,721	11,863	100 bp paired-end Illumina HiSeq2000	100x

Supplementary Table 3 | Comparison of koala genome assembly statistics phaCin_unsw_v4.1 (long read PacBio only) and phaCin_tgac_v2.0 (short-read and BioNano) showing the greater contiguity of the PacBio assembly.

	phaCin_unsw_v4.1 (primary contigs only)		phaCin_tgac_v2.0 (scaffolds)	
Number of contigs	1906		796464	
Number of contigs in scaffolds	0		28516	
Number of contigs not in scaffolds	1906		791109	
Total size of contigs / scaffolds	3192565135		3603908465	
Longest contig / scaffold	40558015		5231295	
Shortest contig / scaffold	353		200	
Number of contigs > 1K nt	1901	100.0%	64589	7.9%
Number of contigs > 10K nt	1677	93.0%	35271	4.3%
Number of contigs > 100K nt	522	29.0%	10649	1.3%
Number of contigs > 1M nt	383	21.2%	0	0%
Number of contigs > 10M nt	105	5.8%	0	0%
Mean contig size / scaffold size	1675008		4525	
Median contig size	31699		236	
N50 contig length	11587828		798273 (scaffold)	
L50 contig count	85		1246 (scaffold)	
contig %A	30.47		28.68 (scaffold)	
contig %C	19.52		18.41 (scaffold)	
contig %G	19.53		18.41 (scaffold)	
contig %T	30.47		28.69 (scaffold)	
contig %N	0.00		5.82 (scaffold)	
contig %non-ACGTN	0.00		0.00	
Number of contig non-ACGTN nt	0		0	

Supplementary Table 4 | Assembly Completeness in BUSCO.

	phaCin_unsw_v4.1	koala_tgac_v2.0
Complete BUSCOs	2144	1858
Complete and single-copy BUSCOs	2094	1827
Complete and duplicated BUSCOs	50	31
Fragmented BUSCOs	587	589
Missing BUSCOs	292	576
Total BUSCO groups searched	3023	3023

Supplementary Table 5 | Sequence (Mb) virtually assigned to koala chromosomes compared to Tasmanian devil (ancestral karyotype used as a point of reference) and gray short-tailed opossum.

Tasmanian devil Ancestral (2n=14)		Koala Derived (2n=16)		Gray short-tailed opossum Derived (2n=18)	
Chromosome	**Mb**	**Chromosome**	**Mb**	**Chromosome**	**Mb**
1	740	1	733	3 + 6	819
2	684	4 + 7	663	1	748
3	641	2	601	4 + 7	695
4	487	3	480	2	541
5	300	5	225	8	312
6	263	6	207	5	304
X	86	X	68	X	79
Total	3201	Total	2978	Total	3498

Supplementary Table 6 | Mapping statistics for CENP-A, CREST and Input ChIP-seqdatasets.

	Alignment statistics to phaCin_unsw_v4.1					
	CENP-A	**%**	**CREST**	**%**	**INPUT**	**%**
paired reads	21957887	100.00	33421909	100.00	39171541	100.00
aligned concordantly 0 times	913758	4.16	982932	2.94	932658	2.38
aligned concordantly 1 times	18496083	84.23	28706056	85.89	34282257	87.52
aligned concordantly >1 times	2548046	11.60	3732921	11.17	3956626	10.10
pairs aligned concordantly 0 times	913758		982932		932658	
aligned discordantly 1 time	360763	39.48	281291	28.62	298022	31.95
pairs aligned 0 times concordantly or discordantly	552995		701641		634636	
mates make up the pairs	1105990		1403282		1269272	
aligned 0 times	656148	59.33	973476	69.37	879039	69.26
aligned exactly 1 time	237228	21.45	230151	16.40	205418	16.18
aligned >1 times	212614	19.22	199655	14.23	184815	14.56
overall alignment rate		98.51		98.54		98.88

Supplementary Tables 7-10 | *De novo* and Mars summary statistics for ChIP-Seq for single koala sample for two IPs. Separate excel file.

Supplementary Table 11 | Comparison of koala repeat regions with other marsupial, eutherian and monotreme genomes.

	Koala	Gray short-tailed opossum	Tammar wallaby	Platypus	Human	Mouse	Gorilla
SINE	8.9	11.9	11.8	21.0	13.4	7.8	11.8
LINE	32.1	30.1	29.6	20.8	21.2	20.3	21.0
LTR	0.2	0.2	0.3				
ERV	1.0	9.8	1.7	0.2	9.0	12.0	8.8
Gypsy	0.2				0.2		0.2
DNA	1.6	2.4	2.2	0.8	3.7	1.1	2.7
Satellites	0.1			0.9	2.8	0.2	1.3
Simple Repeats	1.8	2.0	2.3	1.1	1.5	3.1	2.0

[Large data file (>80 Mb) of repeat calls is available. File name: Repeat_calls.gff upon request]

Supplementary Table 12 | SLAC and MEME codon-based output for single koala sample.
Separate excel file.

Supplementary Table 13 | MEME output by-codon and by-branch for single koala sample.
Separate excel file.

Supplementary Table 14 | The aquaporins of koala showing scaffold position, reading strand and the number of exons.

Protein	Identification	Scaffold	Start	Finish	Strand	Exons
Classical						
AQP0	PCI4_00006840-RA	phaCin_unsw_v4.1.fa.scaf00038	1256395	1262891	-	7
AQP1	PCI4_00017566-RA	phaCin_unsw_v4.1.fa.scaf00153	6910847	6928809	+	4
AQP2	PCI4_00018407-RA	phaCin_unsw_v4.1.fa.scaf00166	1222496	1291207	+	5
AQP4	PCI4_00019756-RA	phaCin_unsw_v4.1.fa.scaf00185	430703	443685	+	5
AQP5	PCI4_00018408-RA	phaCin_unsw_v4.1.fa.scaf00166	1299065	1303890	+	5
AQP5	PCI4_00002921-RA	phaCin_unsw_v4.1.fa.scaf00013	9042752	9043549	-	2
AQP6	PCI4_00018409-RA	phaCin_unsw_v4.1.fa.scaf00166	1318199	1322680	+	4
Aquaglyceroporins						
AQP3	PCI4_00024970-RA	phaCin_unsw_v4.1.fa.scaf00358	117408	125635	-	6
AQP7	PCI4_00004566-RA	phaCin_unsw_v4.1.fa.scaf00022	17589316	17636945	-	2
AQP7	PCI4_00026460-RA	phaCin_unsw_v4.1.fa.scaf01349	4604	17705	-	6
AQP9	PCI4_00000317-RA	phaCin_unsw_v4.1.fa.scaf00001	27914219	27974886	+	7
AQP10	PCI4_00007559-RA	phaCin_unsw_v4.1.fa.scaf00043	4098538	4105632	-	5
Aquaammoniaporins						
AQP8	PCI4_00009519-RA	phaCin_unsw_v4.1.fa.scaf00060	9801316	9810135	+	4
AQP11	PCI4_00006543-RA	phaCin_unsw_v4.1.fa.scaf00036	612875	629390	-	3
Unorthodox Aquaporins						
AQP12	PCI4_00004811-RA	phaCin_unsw_v4.1.fa.scaf00024	11814861	11820666	+	3

Supplementary Table 15 | List of annotated taste receptor genes in koala and other marsupial assemblies. Separate excel file.

Supplementary Table 16 | List of annotated genes involved in koala lactation. Separate excel file.

Supplementary Table 17 | MHC, TCR, and IG genes annotated using BLAST and HMMER.

Gene family	Number of genes/pseudogenes	Number of scaffolds
MHC I	23	9
MHC II	23	10
TRA/D	94	1
TRB	48	1
TRG	9	1
TRM	20	1
IGH	40	15
IGK	82	12
IGL	19	1

Supplementary Table 18 | List of annotated koala immune genes. Separate excel file.

Supplementary Table 19 | Summary of SNPs identified in candidate immune genes.

Gene Family	Gene Name	Length Examined (kb)*	CDS Length (bp)	Total No. of SNPs†			
				Within 10 kb	Intron	Exon (s)	Exon (ns)
MHC Class II	PhciDMA	22	795	78	6	1	1
	PhciDMB	24	807	85	9	1	0
	PhciDBA	23	771	479	31	3	17
	PhciDBA	23	771	300	28	2	15
	PhciDAA	25	771	265	47	0	6
	PhciDCB	29	849	87	45	0	3
	MHCIIDAB	29	801	234	237	6	23
	MHCIIDAB	29	801	522	237	9	27
Glycoprotein	CD4	45	1395	36	19	0	3
Cytokine	CD8-a	38	969	55	53	0	2
	CD8-b	40	636	36	37	0	1
	IFN-gamma	24	495	99	2	0	0
	TNF-a	22	699	15	0	0	0
	IL1B	25	813	50	11	1	0
	IL2	25	315	69	12	1	0
	IL4	27	471	51	11	0	0
	IL6	25	648	93	14	1	0
	IL10	25	621	33	6	1	0
	IL12A	26	636	91	34	2	0
	IL17A	23	462	68	8	0	0
	IL17B	34	597	46	37	0	0
	IL23A	22	558	32	4	0	0

*Includes gene + 10 kb upstream and downstream

†Minimum variant frequency = 0.2

Supplementary Table 20 | Diversity statistics for the coding regions of the candidate immune genes.

Gene Family	Gene Name	SNPs	*h*	*hd*	*k*	π	Tajima's D*	Z Statistic	Prob
MHC Class II	PhciDMA	2	4	0.69	0.89	1.12E-03	0.72	1.46	0.15
	PhciDMB	1	2	0.50	0.50	6.20E-04	0.99	1.00	0.32
	PhciDBA	20	8	0.97	9.22	1.20E-02	1.25	-0.26	0.80
	PhciDBA	17	7	0.92	8.39	1.09E-02	1.66	0.72	0.48
	PhciDAA	6	7	0.94	2.67	3.46E-03	0.91	-2.64	0.01
	PhciDCB	3	4	0.78	1.28	1.51E-03	0.60	-0.29	0.77
	MHCIIDAB	29	9	1.00	13.61	1.71E-02	0.97	-1.08	0.28
	MHCIIDAB	36	9	1.00	18.69	2.34E-02	1.71	0.42	0.68
Glycoprotein	CD4	3	4	0.58	1.17	8.40E-04	0.22	-1.19	0.24
	CD8-a	2	4	0.78	1.06	1.09E-03	1.49	-0.68	0.50
	CD8-b	1	2	0.56	0.56	8.70E-04	1.40	-1.08	0.28
Cytokine	IFN-gamma	0	1	0.00	-	-	-	0.00	1.00
	TNF-a	0	1	0.00	-	-	-	0.00	1.00
	IL1B	1	2	0.39	0.39	4.80E-04	0.16	0.99	0.32
	IL2	1	2	0.39	0.39	1.23E-03	0.16	-1.02	0.31
	IL4	0	1	0.00	-	-	-	0.00	1.00
	IL6	1	2	0.50	0.50	7.70E-04	0.99	-1.13	0.26
	IL10	1	2	0.39	0.39	6.30E-04	0.16	-0.97	0.33
	IL12A	2	3	0.72	0.89	1.40E-03	0.72	-1.48	0.14
	IL17A	0	1	0.00	-	-	-	0.00	1.00
	IL17B	0	1	0.00	-	-	-	0.00	1.00
	IL23A	0	1	0.00	-	-	-	0.00	1.00

h = number of inferred haplotypes, *hd* = haplotype diversity, *k* = average number of nucleotide differences, π = nucleotide diversity.

*All p > 0.10

Supplementary Table 21 | Chlamydia vaccine immune response case/control association test results

Vaccine Response	SNP	CHISQ	P*	Gene	Location	A1	A2	Type
Antibody (1 cases, 4 controls and 3 missing)	Scaf00255_319720	4.444	0.035	MHCII DMA	Downstream	A	G	Transition
	Scaf00255_345446	4.444	0.035	MHCII DMB	Downstream	C	T	Transition
	Scaf00192_5793477	4.444	0.035	CD8-a	Intronic	T	C	Transition

*Only SNPs significant at the 0.05 alpha level are shown

Supplementary Table 22 | Locations of KoRV loci in phaCin_unsw_v4.1 assembly. The data is in BED format (with 0-based numbering of the start field). Separate excel file.

Supplementary Table 23 | phaCin_unsw_4.1 KoRV insertions with breakpoint sequences which are also present in the Birkie genome sequence library.

KoRV insertion id	Example Birke library read id	
	Covering 5' breakpoint	Covering 3' breakpoint
KoRV:23	D3VG1JS1:170:D24TWACXX:6:2316:14288:91457/1	D3VG1JS1:170:D24TWACXX:8:2204:11905:45020/1
KoRV:50	D3VG1JS1:170:D24TWACXX:8:1211:14406:64319/2	D3VG1JS1:170:D24TWACXX:6:2107:18748:31401/2
KoRV:54	D3VG1JS1:170:D24TWACXX:5:1306:5165:84750/2	D3VG1JS1:170:D24TWACXX:3:1314:6983:48313/1
KoRV:5	D3VG1JS1:170:D24TWACXX:6:2212:4306:37443/1	D3VG1JS1:170:D24TWACXX:8:1302:12843:74096/2
recKoRV3:1	not found	D3VG1JS1:170:D24TWACXX:3:2210:20277:46676/1
recKoRV3:2	D3VG1JS1:170:D24TWACXX:8:1308:17696:38038/1	not found

Supplementary Table 24 | Recombinant KoRV (recKoRV) breakpoints found in historical koala genome sequences.

Breakpoint identified in koala	bp Sequence	# Reads Mapped	Pci-SN265	Pci-QM-J6480	Pci-um3435	Pci-MCZ12454	Pci-MCZ8574	Pci-maex1738	Pci-582119
PC	GGTAGTAGAACCCTTTAGGCCATGG AAAACTCCAGATTCCTTAGcaagactcca taagtaaactagaggattccttaacctccctgtctgaa gtagtgctc	19	X					X	
PC	ggagacggacagcctgttcctcctcgcagaaccccc gccctatccaacatcccctccCTGTCCTGTAT ACTCCCTATATAAACCCTAGCTCAA	22	X			X	X		
Bi	ttaatccttctattatttctgcagtcaaaaggattgtcttcc aggagactgTTGGACCTGGTGAAAGGG TCTAGAAAGAAGAAAGATTTTTTTATT GGGTC	>808	X	X	X	X	X	X	X
Bi	TTTAAAATTAGCTTGGGTCGTCAACT CGCCTATAGGGATGAGAGTGGCCCC aggagactcaaggaaaggttagataagaggcagtt agagcaccaaaagaa	>606	X	X	X	X	X	X	X

Supplementary Table 25 | Genes with KoRV insertions. Separate excel file.

References

1. Weisenfeld, N.I. *et al.* Comprehensive variation discovery in single human genomes. *Nature Genetics* **46**, 1350-1355 (2014).
2. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963 (2014).
3. Mapleson, D., Accinelli, G.G., Kettleborough, G., Wright, J. & Clavijo, B.J. KAT: A K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, btw663 (2016).
4. Deakin, J.E. *et al.* Anchoring genome sequence to chromosomes of the central bearded dragon (*Pogona vitticeps*) enables reconstruction of ancestral squamate macrochromosomes and identifies sequence content of the Z chromosome. *BMC Genomics* **17**(2016).
5. De Leo, A. *et al.* Comparative chromosome painting between marsupial orders: relationships with a 2n= 14 ancestral marsupial karyotype. *Chromosome Research* **7**, 509-517 (1999).
6. Rens, W. *et al.* Reversal and convergence in marsupial chromosome evolution. *Cytogenetic and Genome Research* **102**, 282-290 (2004).
7. Murchison, E.P. *et al.* Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* **148**, 780-791 (2012).
8. Mikkelsen, T.S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167-177 (2007).
9. Carone, D.M. *et al.* A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. *Chromosoma* **118**, 113-125 (2009).
10. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170 (2014).
11. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).
12. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015. (2015).
13. R Core Development Team. R: A language and environment for statistical computing. . (R Foundation for Statistical Computing, Vienna, Austria. , 2008).
14. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
15. Brown, J.D. & O'Neill, R.J. The Evolution of Centromeric DNA Sequences. *eLS* (2014).
16. Earnshaw, W.C. & Rothfield, N. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma* **91**, 313-321 (1985).
17. Nagaki, K. *et al.* Sequencing of a rice centromere uncovers active genes. *Nature Genetics* **36**, 138-145 (2004).
18. Zhang, Y. *et al.* Structural features of the rice chromosome 4 centromere. *Nucleic Acids Research* **32**, 2023-2030 (2004).

19. Carbone, L. *et al.* Centromere remodeling in *Hoolock leuconedys* (Hylobatidae) by a new transposable element unique to the gibbons. *Genome Biology and Evolution* **4**, 760-770 (2012).

20. Sperber, G.O., Airola, T., Jern, P. & Blomberg, J. Automated recognition of retroviral sequences in genomic data—RetroTector©. *Nucleic Acids Research* **35**, 4964-4976 (2007).

21. Brown, C.J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527-542 (1992).

22. Grant, J. *et al.* Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* **487**, 254-258 (2012).

23. Renfree, M.B. *et al.* Genome sequence of an Australian kangaroo, *Macropus eugenii,* provides insight into the evolution of mammalian reproduction and development. *Genome Biology* **12**, 1 (2011).

24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580 (1999).

25. Bembom, O. seqLogo: Sequence logos for DNA sequence alignments. . (R package version 1.40.0. , 2016).

26. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R. & Hofacker, I.L. The vienna RNA websuite. *Nucleic Acids Research* **36**, W70-W74 (2008).

27. Deakin, J., Graves, J. & Rens, W. The evolution of marsupial and monotreme chromosomes. *Cytogenetic and Genome Research* **137**, 113-129 (2012).

28. Hobbs, M. *et al.* A transcriptome resource for the koala (*Phascolarctos cinereus*): insights into koala retrovirus transcription and sequence diversity. *BMC Genomics* **15**, 1 (2014).

29. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).

30. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* **13**, 329-342 (2012).

31. Smit, A., Hubley, R. & Green, P. RepeatModeler Open-1.0. 2008-2015. (2014).

32. Boutet, E. *et al.* UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Plant Bioinformatics: Methods and Protocols*, 23-54 (2016).

33. O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 733-745 (2015).

34. Wong, E.S., Papenfuss, A.T. & Belov, K. Immunome database for marsupials and monotremes. *BMC Immunology* **12**, 48 (2011).

35. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652 (2011).

36. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

37. Borodovsky, M. & Lomsadze, A. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Current Protocols in Bioinformatics*, 4.5. 1-4.5. 17 (2011).

38. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435-W439 (2006).

39. MacManes, M.D. Establishing evidenced-based best practice for the de novo assembly and evaluation of transcriptomes from non-model organisms. *bioRxiv*, 035642 (2016).

40. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research* **26**, 1134-1144 (2016).

41. Li, L., Stoeckert, C.J. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178-2189 (2003).

42. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).

43. Vilella, A.J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research* **19**, 327-335 (2009).

44. Pond, S.L.K. & Muse, S.V. HyPhy: hypothesis testing using phylogenies. in *Statistical Methods in Molecular Evolution* 125-181 (Springer, 2005).

45. Delport, W., Poon, A.F., Frost, S.D. & Pond, S.L.K. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455-2457 (2010).

46. Delport, W. *et al.* CodonTest: modeling amino acid substitution preferences in coding sequences. *PLoS Computational Biology* **6**, e1000885 (2010).

47. Pond, S.L.K., Posada, D., Gravenor, M.B., Woelk, C.H. & Frost, S.D. GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096-3098 (2006).

48. Pond, S.L.K. & Frost, S.D. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution* **22**, 1208-1222 (2005).

49. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics* **8**, e1002764 (2012).

50. Glusman, G. *et al.* The olfactory receptor gene superfamily: data mining, classification, and nomenclature. *Mammalian genome* **11**, 1016-1023 (2000).

51. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

52. Van Dongen, S. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications* **30**, 121-141 (2008).

53. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).

54. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).

55. Wheeler, T.J. & Eddy, S.R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487-2489 (2013).

56. Lipman, D.J. & Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435-1441 (1985).

57. Finn, R.N., Chauvigné, F., Hlidberg, J.B., Cutler, C.P. & Cerdà, J. The lineage-specific evolution of aquaporin gene clusters facilitated tetrapod terrestrial adaptation. *PLoS One* **9**, e113686 (2014).

58. Gilbertson, T., Kim, I., Siears, N., Zhang, H. & Liu, L. The water response in taste cells: expression of aquaporin-1,-2 and-5 and the characterization of hypoosmic-induced currents in mammalian taste cells. *Chemical Senses* **24**, 596 (1999).

59. Gilbertson, T.A., Baquero, A.F. & Spray-Watson, K.J. Water taste: the importance of osmotic sensing in the oral cavity. *Journal of Water and Health* **4**, 35-40 (2006).

60. Watson, K.J. *et al.* Expression of aquaporin water channels in rat taste buds. *Chemical Senses* **32**, 411-421 (2007).

61. Meyerhof, W. *et al.* The molecular receptive ranges of human TAS2R bitter taste receptors. *Chemical Senses* **35**, 157-170 (2010).

62. Hu, Y. *et al.* Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proceedings of the National Academy of Sciences* **114**, 1081-1086 (2017).

63. Li, X. *et al.* Pseudogenization of a sweet-receptor gene accounts for cats' indifference toward sugar. *PLoS Genetics* **1**, e3 (2005).

64. Li, D. & Zhang, J. Diet shapes the evolution of the vertebrate bitter taste receptor gene repertoire. *Molecular Biology and Evolution* **31**, 303-309 (2014).

65. Hayakawa, T., Suzuki-Hashido, N., Matsui, A. & Go, Y. Frequent expansions of the bitter taste receptor gene repertoire during evolution of mammals in the Euarchontoglires clade. *Molecular Biology and Evolution* **31**, 2018-2031 (2014).

66. Kishida, T., Thewissen, J., Hayakawa, T., Imai, H. & Agata, K. Aquatic adaptation and the evolution of smell and taste in whales. *Zoological letters* **1**, 9 (2015).

67. Robinson, J.T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24-26 (2011).

68. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33**, 1870-1874 (2016).

69. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* **57**, 758-771 (2008).

70. Eschler, B., Pass, D., Willis, R. & Foley, W. Distribution of foliar formylated phloroglucinol derivatives amongst Eucalyptus species. *Biochemical Systematics and Ecology* **28**, 813-824 (2000).

71. Gleadow, R.M., Haburjak, J., Dunn, J., Conn, M. & Conn, E.E. Frequency and distribution of cyanogenic glycosides in Eucalyptus L'Hérit. *Phytochemistry* **69**, 1870-1874 (2008).

72. Lossow, K. *et al.* Comprehensive analysis of mouse bitter taste receptors reveals different molecular receptive ranges for orthologous receptors in mice and humans. *Journal of Biological Chemistry* **291**, 15358-15377 (2016).

73. Frankenberg, S., Pask, A. & Renfree, M.B. The evolution of class V POU domain transcription factors in vertebrates and their characterisation in a marsupial. *Developmental Biology* **337**, 162-170 (2010).

74. Renfree, M.B., Ager, E.I., Shaw, G. & Pask, A.J. Genomic imprinting in marsupial placentation. *Reproduction* **136**, 523-531 (2008).

75. Ager, E. *et al.* Insulin is imprinted in the placenta of the marsupial, *Macropus eugenii*. *Developmental Biology* **309**, 317-328 (2007).

76. Johnston, S., McGowan, M., O'Callaghan, P., Cox, R. & Nicolson, V. Studies of the oestrous cycle, oestrus and pregnancy in the koala (*Phascolarctos cinereus*). *Journal of Reproduction and Fertility* **120**, 49-57 (2000).

77. Harper, G.P., Glanville, R. & Thoenen, H. The purification of nerve growth factor from bovine seminal plasma. Biochemical characterization and partial amino acid sequence. *Journal of Biological Chemistry* **257**, 8541-8548 (1982).

78. Ratto, M.H. *et al.* The nerve of ovulation-inducing factor in semen. *Proceedings of the National Academy of Sciences* **109**, 15042-15047 (2012).

79. Bogle, O.A., Ratto, M.H. & Adams, G.P. Evidence for the conservation of biological activity of ovulation-inducing factor in seminal plasma. *Reproduction* **142**, 277-283 (2011).

80. Rodger, J.C. & Hughes, R. Studies of the accessory glands of male marsupials. *Australian Journal of Zoology* **21**, 303-320 (1973).

81. Wildt, D.E. *et al.* Semen characteristics in free-living koalas (*Phascolarctos cinereus*). *Journal of Reproduction and Fertility* **92**, 99-107 (1991).

82. Johnston, S., McGowan, M., Carrick, F., Cameron, R. & Tribe, A. Seminal characteristics and spermatozoal morphology of captive Queensland koalas (*Phascolarctos cinereus adustus*). *Theriogenology* **42**, 501-511 (1994).

83. Johnston, S., O'Callaghan, P., McGowan, M. & Phillips, N. Characteristics of koala (*Phascolarctos cinereus adustus*) semen collected by artificial vagina. *Journal of Reproduction and Fertility* **109**, 319-323 (1997).

84. Johnston, S., McGowan, M., O'Callaghan, P., Cox, R. & Nicolson, V. Natural and artificial methods for inducing the luteal phase in the koala (*Phascolarctos cinereus*). *Journal of Reproduction and Fertility* **120**, 59-64 (2000).

85. Johnston, S., McGowan, M., Phillips, N. & O'Callaghan, P. Optimal physicochemical conditions for the manipulation and short-term preservation of koala (*Phascolarctos cinereus*) spermatozoa. *Journal of Reproduction and Fertility* **118**, 273-281 (2000).

86. Tian, X., Pascal, G., Fouchécourt, S., Pontarotti, P. & Monget, P. Gene birth, death, and divergence: the different scenarios of reproduction-related gene evolution. *Biology of reproduction* **80**, 616-621 (2009).

87. Robert, M. & Gagnon, C. Semenogelin I: a coagulum forming, multifunctional seminal vesicle protein. *Cellular and Molecular Life Sciences CMLS* **55**, 944-960 (1999).

88. Romijn, J. Polyamines and transglutaminase actions: Polyamine und Transglutaminase‐Wirkungen. *Andrologia* **22**, 83-91 (1990).

89. Ramm, S.A., Parker, G.A. & Stockley, P. Sperm competition and the evolution of male reproductive anatomy in rodents. *Proceedings of the Royal Society of London B: Biological Sciences* **272**, 949-955 (2005).

90. Bercovitch, F., Tobey, J., Andrus, C. & Doyle, L. Mating patterns and reproductive success in captive koalas (*Phascolarctos cinereus*). *Journal of Zoology* **270**, 512-516 (2006).

91. Sharp, J.A. *et al.* The tammar wallaby: a marsupial model to examine the timed delivery and role of bioactives in milk. *General and Comparative Endocrinology* **244**, 164-177 (2017).

92. Morris, K.M. *et al.* Characterisation of the immune compounds in koala milk using a combined transcriptomic and proteomic approach. *Scientific Reports* **6: 35011**(2016).

93. Cheng, Y. *et al.* Characterisation of MHC class I genes in the koala. *Immunogenetics*, 1-9 (2017).

94. Belov, K. *et al.* Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. *PLoS Biology* **4**, e46 (2006).

95. Deakin, J., Siddle, H.V., Cross, J., Belov, K. & Graves, J. Class I genes have split from the MHC in the tammar wallaby. *Cytogenetic and Genome Research* **116**, 205-211 (2007).

96. Wang, X., Olp, J.J. & Miller, R.D. On the genomics of immunoglobulins in the gray, short-tailed opossum *Monodelphis domestica*. *Immunogenetics* **61**, 581-596 (2009).

97. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

98. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

99. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).

100. Morris, K.M., Wright, B., Grueber, C.E., Hogg, C. & Belov, K. Lack of genetic diversity across diverse immune genes in an endangered mammal, the Tasmanian devil (Sarcophilus harrisii). *Molecular Ecology* **24**, 3860-3872 (2015).

101. Lau, Q.T., Griffith, J.E. & Higgins, D.P. Identification of MHCII variants associated with chlamydial disease in the koala (Phascolarctos cinereus). *PEERJ* **2**, e443 (2014).

102. Kollipara, A. *et al.* Antigenic specificity of a monovalent versus polyvalent MOMP based Chlamydia pecorum vaccine in koalas (Phascolarctos cinereus). *Vaccine* **31**, 1217-1223 (2013).

103. Polkinghorne, A., Hanger, J. & Timms, P. Recent advances in understanding the biology, epidemiology and control of chlamydial infections in koalas. *Veterinary Microbiology* **165**, 214-223 (2013).

104. Dobin, A. & Gingeras, T.R. Mapping RNA‐seq Reads with STAR. *Current Protocols in Bioinformatics*, 11.14. 1-11.14. 19 (2015).

105. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2013).

106. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

107. Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. & Dillies, M.-A. SARTools: a DESeq2-and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. *PloS One* **11**, e0157022 (2016).

108. Tsangaras, K. *et al.* Hybridization capture reveals evolution and conservation across the entire koala retrovirus genome. *PLoS One* **9**, e95633 (2014).

109. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).

110. Phillips, S.S. Population trends and the koala conservation debate. *Conservation Biology* **14**, 650-659 (2000).

111. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**, 704-714 (2016).

112. Uchimura, A. *et al.* Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Research* **25**, 1125-1134 (2015).

113. Bragg, J.G., Potter, S., Bi, K. & Moritz, C. Exon capture phylogenomics: efficacy across scales of divergence. *Molecular Ecology Resources* (2015).

114. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303 (2010).

115. Wang, J. COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources* **11**, 141-145 (2011).

116. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. URL h ttp. *www. R-project. org* (2016).

117. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575 (2007).

118. Miller, W. *et al.* Genetic diversity and population structure of the endangered marsupial *Sarcophilus harrisii* (Tasmanian devil). *Proceedings of the National Academy of Sciences* **108**, 12348-12353 (2011).

119. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).