(GIGA)byte

DATA RELEASE

A reference genome for the critically endangered woylie, *Bettongia penicillata ogilbyi*

Emma Peel[1], Luke Silver[1], Parice Brandies[1], Carolyn J. Hogg[1] and Katherine Belov[1,*]

1 School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales, Australia

ABSTRACT

Biodiversity is declining globally, and Australia has one of the worst extinction records for mammals. The development of sequencing technologies means that genomic approaches are now available as important tools for wildlife conservation and management. Despite this, genome sequences are available for only 5% of threatened Australian species. Here we report the first reference genome for the woylie (*Bettongia penicillata ogilbyi*), a critically endangered marsupial from Western Australia, and the first genome within the Potoroidae family. The woylie reference genome was generated using Pacific Biosciences HiFi long-reads, resulting in a 3.39 Gbp assembly with a scaffold N50 of 6.49 Mbp and 86.5% complete mammalian BUSCOs. Assembly of a global transcriptome from pouch skin, tongue, heart and blood RNA-seq reads was used to guide annotation with Fgenesh++, resulting in the annotation of 24,655 genes. The woylie reference genome is a valuable resource for conservation, management and investigations into disease-induced decline of this critically endangered marsupial.

Subjects Genetics and Genomics, Animal Genetics, Genetics

DATA DESCRIPTION

Background and context

Globally, we are experiencing a biodiversity crisis, as more than 1 million species currently face extinction [1]. Australia is one of 17 megadiverse countries [2], and has a high level of endemism, with 87% of mammals, 94% of frogs and 93% of reptiles being endemic to Australia [3]. Despite this, Australia has one of the worst mammal extinction rates in the world, with over 10% of endemic terrestrial mammals driven to extinction within the past 200 years [4]. The International Union for Conservation of Nature (IUCN) currently lists over 1000 Australian animals as critically endangered, endangered or vulnerable as of August 2021 [5], yet genome sequences are only available for 5% of these species.

Reference genomes are a valuable conservation tool that allow researchers and managers to answer vital biological questions and inform management policy [6]. Development of new sequencing technologies and their subsequent decrease in cost allows the genomes of threatened species to be sequenced [7–9]. Individual research groups can now assemble highly contiguous genomes, which can then be used to answer various biological, evolutionary and conservation questions [10–12].

The woylie, or brush-tailed bettong (*Bettongia penicillata ogilbyi*, NCBI:txid881300), is a small marsupial of the Potoroidae family (Figure 1) [13]. Marsupials are one of three

Submitted: 15 September 2021
Accepted: 08 December 2021
Published: 10 December 2021

* Corresponding author. E-mail: kathy.belov@sydney.edu.au

Published by GigaScience Press.

Preprint submitted at https://doi.org/10.1101/2021.12.07.471656

Gigabyte, 2021, 1–15

Figure 1. Woylie *Bettongia penicillata ogilbyi* from Western Australia. Photo credit: Sabrina Trocini.

lineages of mammals, the others being eutherians (e.g. humans and mice) and monotremes (e.g. platypus and echidna). Since their divergence from eutherian mammals around 156 million years ago, marsupials have evolved into over 300 species, most of which are endemic to Australia [14]. Marsupials differ to other mammals in several ways; the most prominent is their pouch. After a short gestation of only 20 days, woylies give birth to altricial young, which develop within the pouch for 100 days [13]. The complex milk profile of the mother provides nutrient and immune support throughout pouch life [15]. Similar to wallabies and kangaroos, woylies undergo embryonic diapause, whereby embryonic development is suspended by the suckling pouch young, and resumes when the young exits the pouch [13]. Woylies are ecosystem engineers, with individuals displacing on average 4.8 tonnes of soil per year [16]. This bioturbation is essential for ecosystem health and function, as it activates and disperses the mycorrhizal fungi comprising a large portion of their diet [17, 18], alters soil nutrient composition and water penetration [19], and aids seed dispersal [20].

Historically, woylies inhabited much of central and southern Australia; however, populations have contracted to 1% of their former range owing to habitat loss and fragmentation [4, 21]. Between 1999 and 2006, woylie populations significantly declined by more than 90% [22], resulting in their current IUCN listing as critically endangered [5]. The decline is thought to be caused by a combination of predation by introduced feral cats (*Felis catus*) and foxes (*Vulpes vulpes*), and an unknown disease linked to parasites such as trypanosomes [23–25]. Currently, there are two remaining indigenous populations in the Upper Warren and Dryandra regions of Western Australia [22]. In response to the decline, several translocations and reintroductions have been conducted within Western Australia (WA), South Australia (SA) and New South Wales (NSW), guided by genetic assessments of

diversity and population structure using microsatellite [26, 27] and genomic-based methods [28].

In this study, we present the first *de novo* reference genome assembly of the woylie, as well as four tissue transcriptomes. This is the first genome sequenced within the Potoroidae family, which will be a valuable tool for investigating disease-induced decline, basic biology, and to aid conservation.

METHODS

Sample collection and sequencing

Spleen, heart, kidney and tongue were opportunistically sampled from a single wild female woylie (woy01), as well as pouch skin from a second wild female woylie (woy02), both of which died by vehicle strike at Manjimup, Western Australia (WA) in 2018. In addition, 500 µL of peripheral blood was collected into RNAprotect Animal Blood tubes (Qiagen) from a third wild male woylie (woy03) from Balban, WA, in 2018 during routine trapping and health examinations. All samples were collected under the Western Australian Government Department of Biodiversity, Conservation and Attractions animal ethics 2018-22F and scientific licence number NSW DPIE SL101204.

High-molecular-weight (HMW) DNA was extracted from woy01 kidney using the Nanobind Tissue Big DNA kit (Circulomics), and quality assessed using the NanoDrop 6000 with an A260/280 of 1.91 and A260/230 of 2.37. HMW DNA was submitted to the Australian Genome Research Facility (Brisbane) for Pacific Biosciences (PacBio) HiFi sequencing. Briefly, the DNA was sheared using the Megaruptor2 kit to generate 15 to 20-Kbp (kilobase pair) fragments. The BluePippin SMRTbell Library Kit was then used to select DNA fragments longer than 15 Kbp, which were used as input to the SMRTbell Express Template Prep Kit 2.0. The resulting PacBio HiFi SMRTbell libraries were sequenced across two single-molecule real-time (SMRT) cells on the PacBio Sequel II. This resulted in 37 Gbp (gigabase pairs) of raw data.

For 10X Chromium linked-read sequencing, HMW DNA was extracted from 25 mg of woy01 spleen using the MagAttract HMW DNA kit (Qiagen), and quality was assessed using the NanoDrop 6000 with an A260/280 and A260/230 of 1.8–2.3. HMW DNA was submitted to the Ramaciotti Centre for Genomics (UNSW) for 10X Chromium genomics library preparation, and 150-bp (base pair) paired-end (PE) reads were sequenced on an Illumina NovaSeq 6000 S1 flowcell. This generated 137 Gbp of raw data.

Total RNA was extracted from 25 mg of woy01 tongue and heart, woy02 pouch skin, using the RNeasy Plus Mini Kit (Qiagen). In addition, total RNA was extracted from 500 µL of woy03 peripheral blood using the RNAprotect Animal Blood Kit (Qiagen). In all extractions, contaminating DNA was removed through on-column digestion using the RNase-free DNase I set (Qiagen). RNA purity was assessed using the NanoDrop 6000, with all samples displaying an A260/280 and A260/230 of 1.9–2.2. RNA concentration and integrity were measured using an RNA Nano 6000 chip (Agilent Technologies), with all samples displaying an RNA integrity number (RIN) greater than 7. Total RNA was submitted to the Ramaciotti Centre for Genomics (University of New South Wales) for TruSeq mRNA library preparation. All tissue libraries were sequenced as 150-bp PE reads across one lane of an S1 flowcell on the NovaSeq 6000, while the blood library was sequenced as 150-bp PE reads across an SP flowcell on the NovaSeq 6000. This resulted in 23–27 GB (gigabytes) raw data per sample. All genomic and transcriptomic data generated in this study are summarised in Table 1.

Table 1. Summary of sequencing data generated in this study.

Sequencing platform	Data type	Individual/tissue	Raw data (GB)
PacBio Sequel II	HiFi reads	woy01 kidney	37
Illumina NovaSeq6000	10x linked-reads	woy01 spleen	137
Illumina NovaSeq6000	RNA-seq reads	woy01 tongue and heart woy02 pouch skin woy03 blood	23–27 per sample

Genome assembly and annotation

Raw sequencing data was quality checked using SMRT Link v9.0.0.92188 [29] and fastQC v0.11.8 (RRID:SCR_014583) [30]. Sequences were assembled d*e novo* using Improved Phased Assembler (IPA) v1.1.2 [31] with default parameters on the Nimbus cloud service provided by the Pawsey Supercomputing Centre (virtual machine – 64 vCPUs; 256 GB RAM; 3 TB Storage). Assembly statistics were obtained using BBmap v37.98 (RRID:SCR_016965) [32] and assembly completeness assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) v4.0.6 and v3.1.0 (RRID:SCR_015008) [33]. The assembly was then filtered to remove duplicate haplotigs using purgedups v1.0.1 [34, 35]. The *de novo* assembly was then scaffolded with 10x linked reads using 10x Genomics Long Ranger v2.2.2 (RRID:SCR_018925) and arcs v1.1.1 [36]. Additional gap filling was performed with the HiFi data using PBJelly v14.1 (RRID:SCR_012091) [9]. The genome was polished with Pilon v1.20 (RRID:SCR_014731) [37] by converting the 10x linked reads to standard Illumina reads by removing 10x adapter sequences. For annotation, a custom repeat database was generated using RepeatModeler v2.0.1 (RRID:SCR_015027) [38], then RepeatMasker v4.0.6 (RRID:SCR_012954) [39] was used to mask repeats, excluding low complexity regions and simple repeats. Functional completeness was assessed using BUSCO v4.0.6 and v3.1.0 (RRID:SCR_015008) against the mammalian database [40]. Genome annotation was then performed using Fgenesh++ v7.2.2 (RRID:SCR_018928) [41] with general mammalian pipeline parameters and an optimised gene-finding matrix from Tasmanian devils (*Sarcophilus harrisii*). Transcripts with the longest open reading frame for each predicted gene were extracted from the global transcriptome for mRNA-based gene predictions. The National Center for Biotechnology Information (NCBI) non-redundant protein database was used for protein-based gene predictions [42]. Statistics for protein-coding genes were calculated using genestats [43].

Transcriptome assembly and annotation

Raw RNA-seq data was quality checked using fastQC v0.11.8 (RRID:SCR_014583) [30], then and length and quality trimmed using Trimmomatic v0.38 (RRID:SCR_011848) [44] with the following flags: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25. Illumina TruSeq sequencing adapters were removed from the dataset (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10), as well as reads shorter than 25 bp (MINLEN:25). Reads were quality trimmed and removed where the average quality score fell below 5 within a 4-bp sliding window (SLIDINGWINDOW:4:5), as well as at the 5′ (LEADING:5) and 3′ (TRAILING:5) end of the read. Over 99.7% of reads were retained for all datasets post-trimming.

A global transcriptome for the woylie was produced by aligning trimmed reads from four tissues (heart and tongue from woy01, pouch skin from woy02 and whole blood from woy03) to the final genome assembly using hisat2 v2.1.0 (RRID:SCR_015530) [45] using

default parameters. This produced sam files, which were then sorted using SAMTOOLS v1.6 (RRID:SCR_002105) [46] to produce bam files for each tissue. StringTie v2.1.4 (RRID:SCR_002105) [47] was used to produce gtf files for each tissue. Tama merge [48, 49] was used to merge aligned reads from each tissue to a global transcriptome with a 5′ threshold of 3 and 3′ threshold of 500. Transcripts were then removed if they were only present in one tissue sample or had a fragments per kilobase of transcript per million (FPKM) of less than 0.1. CPC2 [50, 51] was used to predict whether a transcript was a coding gene, and to filter out transcripts with low expression. TransDecoder v2.0.1 (RRID:SCR_017647) [52] was then used to determine coding regions and open reading frames within the transcripts. The number of full-length protein coding genes was determined by using BLAST (Basic Local Alignment Search Tool) [53] to identify top hits of the full-length TransDecoder-predicted proteins against the Swiss-Prot non-redundant database, available at UniProt [54]. Functional completeness was assessed using BUSCO v4.0.6 and v3.1.0 (RRID:SCR_015008) against the mammalian database [40]. To determine read representation and generate transcript counts, trimmed reads were mapped back to the global transcriptome assembly using bowtie2 v2.4.4 (RRID:SCR_005476) [55] with default parameters, except a maximum of 20 distinct valid alignments for each read. These alignments were used as input to Salmon v1.4.0 [56] to generate transcript per million (TPM) counts for each tissue. TransDecoder-predicted proteins expressed in the pouch skin with hits to Swiss-Prot (e-value threshold of e^{-5}) were used as input to Panther (RRID:SCR_004869) [57], where they were assigned Gene Ontology (GO) slim terms under the Biological Process category.

RESULTS AND DISCUSSION

Genome

The *de novo* woylie genome assembly was 3.39 Gbp in size, like other marsupial genomes (Table 2). The genome was assembled into just over 1000 scaffolds, with a scaffold N50 of 6.94 Mbp, and is more contiguous than the tammar wallaby genome, the closest relative with an available genome [58]. Gaps made up 0.40% of the genome; fewer than antechinus (*Antechinus stuartii*) (2.75%) [59] but higher than koala (*Phascolarctos cinereus*) (0.1%), which is not surprising given the numerous sequencing technologies used to generate the high-quality koala genome assembly [60]. The high scaffold N50 for the woylie genome relative to other assembly statistics is likely to be associated with the presence of long contigs derived from the long HiFi reads. The longest contig in the assembly was 12 Mbp, with eight contigs longer than 5 Mbp. Following scaffolding with 10x linked-reads, three scaffolds in the assembly were longer than 25 Mbp (longest 35.66 Mbp) and 72 were longer than 10 Mbp. Despite this, the high scaffold N50 may be attributed to scaffolding error. The genome presented here is a high-quality draft assembly and provides a basis for future improvement.

Repeat elements comprised 53.05% of the woylie genome, similar to the tammar wallaby (*Notamacropus eugenii*) (52.8%) [58], but higher than antechinus (44.82%) [59] and koala (47.5%) [60]. Repeat families numbering 1184 were identified in the woylie genome, with long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) being the most numerous (Table 3), as in other marsupial genomes [59, 60, 64]. Retrotransposon-like elements (RTE) were also identified in the woylie genome, as observed in other marsupial genomes and some mammals such as ruminants [59, 65, 66].

Table 2. Assembly metrics for the woylie genome, compared with other published marsupial genomes.

Metric	Woylie (present study)	Koala [60–62]	Antechinus [59]	Tasmanian devil [63]	Tammar wallaby [58]
Year	2021	2018	2020	2012	2011
Genome size (Gbp)	3.39	3.19	3.31	3.17	2.7
No. scaffolds	1116	1318	30,876	237,291	277,711
No. contigs	3016	1935	106,199	35,974	1,174,382
Scaffold N50 (Mbp)	6.94	480.11	72.7	1.8	0.036
Contig N50 (Mbp)	1.99	11.4	0.08	0.02	0.002
GC (%)	38.64	39.05	36.20	36.04	38.8

Table 3. Repeats identified in the woylie genome.

Repeat type	Number	Length (bp)	Sequence masked
			Percentage (%)
SINE			
ALUs	16,376	3,542,886	0.10
MIRs	2,012,972	349,844,958	10.31
LINE			
LINE1	1,054,578	588,827,450	17.35
LINE2	1,261,114	282,854,529	8.34
CR1	553,626	102,521,146	3.02
LTR			
ERVL	16,900	7,241,092	0.21
ERV1	22,176	7,069,536	0.21
ERV2	22,364	11,265,604	0.33
DNA elements			
hAT-Charlie	141,744	23,630,550	0.70
TcMar-Tigger	36,287	8,389,535	0.25
Other			
Unclassified	1,017,337	231,493,940	6.82
Satellite DNA	29,228	4,458,406	0.13
Small RNA	615	53,205	0.00

Interestingly, primate-specific ALU repeats were identified in the woylie genome. ALU repeats are a type of SINE that comprise over 10% of the human genome and are involved in genome evolution and disease [67]. These repeat elements contributed only 0.10% to the woylie genome, and were also identified in the antechinus genome (0.04%) [59]. As this may represent an inaccurate repeat annotation, further work is required to confirm the presence of ALU repeats in marsupials.

Fgenesh++ predicted 41,868 genes in the woylie genome, of which 24,655 had BLAST hits to eukaryote genes in the NCBI non-redundant database. Of these 24,655 genes, 15,904 were supported by mRNA evidence, and 1309 by protein evidence. This is higher than the number of protein-coding genes annotated by NCBI in the koala genome (20,103) [60] and Tasmanian devil (20,053) [63] (Table 4). The higher number of genes annotated in the woylie genome is likely due to incomplete RNA-seq evidence, and hence gene models, used for gene prediction by Fgnesh++. In addition, fragmentation of the genome causes fragmentation of gene sequences, which can result in an overinflated gene count [68, 69]. Statistics for protein-coding genes annotated within the woylie genome also reflected deficiencies in gene models, as mean gene and exon length, and mean exon number per gene differed to the NCBI annotation of the koala and devil genome (Table 4). The NCBI annotation pipeline uses mRNA and protein evidence from multiple public scientific

Table 4. Statistics for protein-coding genes annotated in the woylie genome compared to NCBI annotations of the koala and Tasmanian devil genomes. Accession numbers provided.

Parameter	Woylie (this study)	Koala (GCF_002099425.1)	Tasmanian devil (GCF_000189315.1)
No. protein-coding genes	24,655	20,103	20,053
Mean gene length (bp)	24,136.6	55,640	45,875
Mean exon length (bp)	550	291	269
Mean intron length (bp)	8559	7261	5936
Mean exon number per gene	3.58	11.11	10.36

databases for gene prediction, resulting in an annotated gene number that more closely resembles humans (~20,000) [70].

Transcriptome

The woylie global transcriptome assembly of four tissues (blood, heart, pouch skin and tongue) contained 145,939 transcripts, with an average transcript length of 7739 bp and transcript N50 of 15,469 bp. TransDecoder predicted 151,147 coding regions within the global transcriptome, of which 74% were complete (contained a start and stop codon) and 75.4% had BLAST hits to the Swiss-Prot non-redundant database. The number of TransDecoder proteins expressed in the pouch skin with BLASTp hits to Swiss-Prot was 89,707, of which 21,856 represented unique Swiss-Prot entries. Of these, 17,588 were assigned GO-slim terms under the biological process (BP) category, with cellular processes (29.8%), metabolic processes (18.9%) and biological regulation (16.7%) being the most common GO terms (Figure 2). Of the GO-slim terms under the BP category, 1.4% were involved in immune system processes, such as immune cell development, activation, and antigen processing (Figure 2).

As this study presents the first marsupial pouch skin transcriptome, and given the important protective role of the pouch, we investigated the top transcripts expressed in this tissue. Of the top 10 transcripts expressed in the pouch skin with hits to Swiss-Prot, four were involved in innate immune defence (Figure 3). The most highly abundant pouch skin transcript was lysozyme C. Lysozyme is an ancient antimicrobial enzyme conserved throughout evolution [71], which degrades the peptidoglycan layer of bacterial cell membranes. Calcium binding proteins of the S100 family, such as S100-A9 and S100A15A, and surfactant-associated protein D (SP-D), were also highly expressed in the pouch skin. These proteins are involved in innate immunity, and are chemotactic [72], antimicrobial [73–75] and modulate inflammation [76]. Marsupial young, including the woylie, are born immunologically naïve without mature immune tissues or cells [77]. The abundance of innate immune proteins in the pouch skin transcriptome highlights the importance of the pouch in protecting naïve young during development. As adaptive immunity does not completely mature until 100 days after birth in some species, the young rely on passive immunity from the milk, rapid development of the innate immune system, and antimicrobial compounds from the pouch for protection against pathogens [78–82]. Antimicrobial compounds expressed in the pouch skin likely contribute to changes in the pouch microbiome throughout lactation in marsupials [83, 84], and may selectively eliminate pathogens via direct antibacterial activity [79, 85, 86].

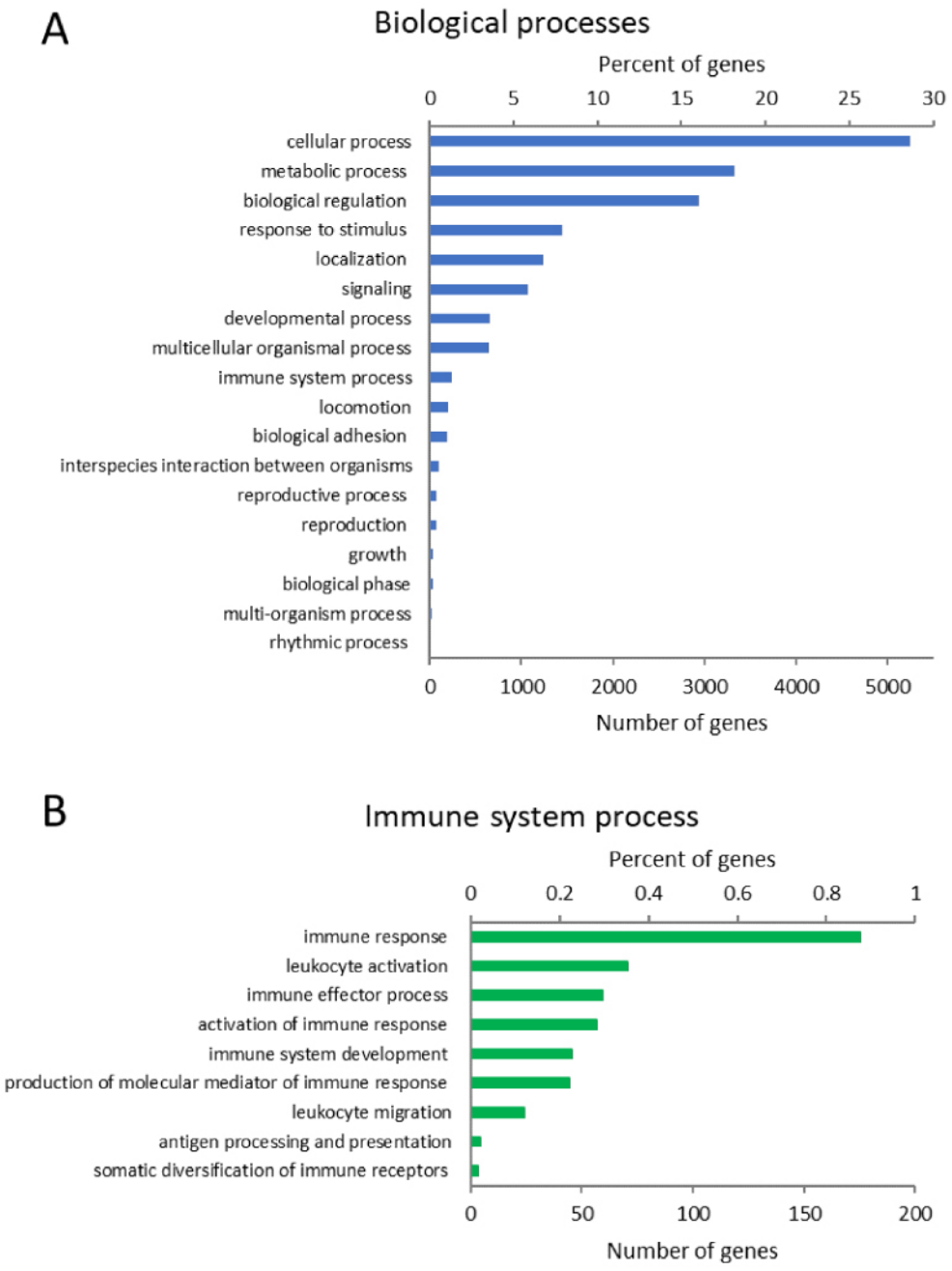

Figure 2. Top GO terms assigned to proteins expressed in the pouch skin with Swiss-Prot hits. (A) Proteins identified under the biological process category and (B) the immune system process term. Both the number of genes and percentage of total genes assigned to each category are provided.

DATA VALIDATION AND QUALITY CONTROL

BUSCO was used to assess functional completeness by searching for complete single-copy gene orthologs within the genome assembly, Fgenesh++ predicted proteins and the global transcriptome assembly. The genome contained 86.5% of complete mammalian BUSCOv4

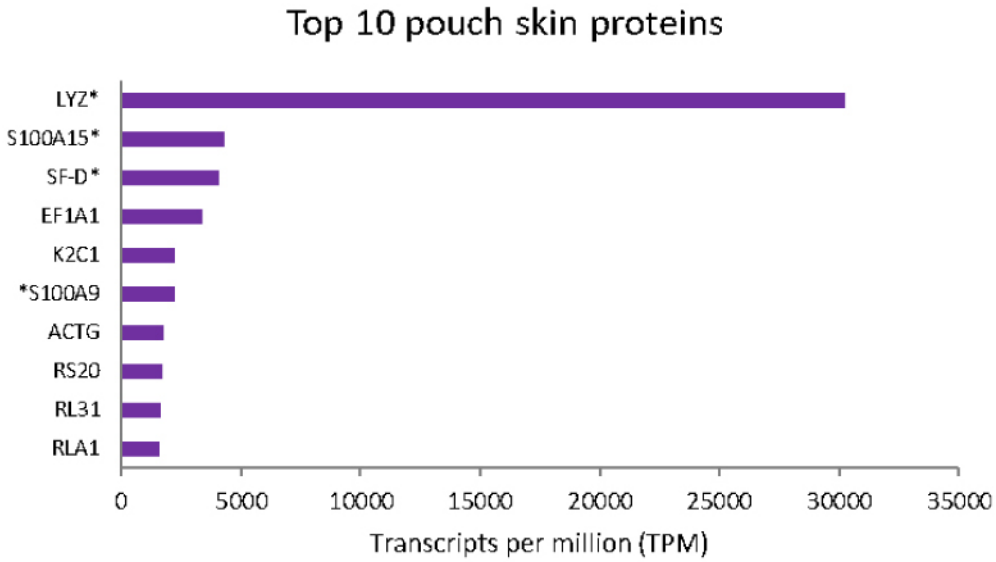

Figure 3. Transcript per million (TPM) counts of the top 10 proteins expressed in the pouch skin with hits to Swiss-Prot. Proteins involved in innate immunity are indicated by *.

genes, comparable to other marsupial genomes [58–60, 63, 87]. Fgenesh++-predicted proteins and the global transcriptome also displayed a high level of completeness, with 80.4% and 80.8% of complete mammalian BUSCOv4 identified, respectively.

High mapping rates were observed for both the genome and global transcriptome assemblies, indicating high sequencing accuracy and low contaminating DNA. 99.8% of HiFi reads and 88.4% of 10x Chromium Illumina reads mapped to the genome assembly. Similarly, 80.25% (blood), 70.96% (pouch skin), 65.30% (tongue) and 60.43% (heart) of RNA-seq reads mapped to the global transcriptome assembly. The lower mapping rate for heart and tongue against the global transcriptome is not unexpected, as reads which map to unannotated transcripts are lost [88]. Alignment of reads from heart and tongue to the genome was higher, with 77.79% and 81.70% of reads mapped, respectively.

REUSE POTENTIAL

Genomes are valuable tools for wildlife conservation and management [6, 89, 90]. In marsupials, Tasmanian devils [63] and koalas [60] are two examples where genomes have been used to investigate genetic diversity, population structure, adaptation and disease [91–93]. The woylie reference genome is the first genome available for the Potoroidae family of marsupials. Not only will this resource facilitate basic biological research of bettongs and potoroos, but also provide a tool for population genomics studies of woylies and other species within the Potoroidae family. The woylie reference genome has already been used alongside reduced representation sequencing data of woylie populations across Australia to investigate population structure and inbreeding [28].

Infectious diseases threaten wildlife globally, with devastating consequences, such as chytridiomycosis in amphibians and devil facial tumour disease in Tasmanian devils [94]. Genetic diversity within immune genes is essential for adaptation to new and emerging diseases [95]. The cause of the rapid decline of woylie populations in the Upper Warren region of WA remains unknown, but an unknown disease has been hypothesised [96].

The woylie reference genome will enable characterisation of immune genes, an essential first step in determining genetic diversity within these genomic regions and detecting pathogen-driven signatures of selection. Our current understanding of woylie immune genes is extremely limited. The long-read sequencing used to generate the woylie reference genome will enable characterisation of complex immune gene families, such as the major histocompatibility complex. This immunogenetic information will be essential for determining the health of existing populations and mitigating potential future disease outbreaks.

DATA AVAILABILITY

The reference genome and global transcriptome assemblies supporting the results of this article are available through the Amazon Web Services Open Datasets Program [97]. The genome assembly and all raw sequencing reads including the PacBio HiFi reads, 10x linked-reads and RNA-seq reads are available through NCBI under the BioProject accession PRJNA763700. Annotations, alignments and other results are available via the *GigaScience* GigaDB repository [98].

DECLARATIONS

LIST OF ABBREVIATIONS

BLAST: Basic Local Alignment Search Tool; bp: base pair; Benchmarking Universal Single-Copy Orthologs (BUSCO); BP: biological process (BP), GO: Gene Ontology; Gbp: gigabase pair; HMW: high molecular weight; IUCN: International Union for Conservation of Nature; LINE: long interspersed nuclear element; Mbp: megabase pair; NCBI: National Center for Biotechnology Information; NSW: New South Wales; PE: paired end; SINE: short interspersed nuclear element; SA: South Australia; WA: Western Australia.

ETHICAL APPROVAL

All samples were collected under the Western Australian Government Department of Biodiversity, Conservation and Attractions animal ethics 2018-22F and scientific licence number NSW DPIE SL101204.

CONSENT FOR PUBLICATION

Not applicable.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This work has been funded by the Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science (CE200100012) and Discovery Project (DP180102465), and the Presbyterian Ladies' College Sydney. PS and LS are supported by an Australian postgraduate award.

AUTHORS' CONTRIBUTIONS

EP conducted the DNA and RNA extractions, the gene ontology analysis and drafted the manuscript. LS, PB and EP assembled and annotated the genome and transcriptomes. KB and CJH designed the study. All authors viewed, commented on and agreed to publication of the manuscript.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Adrian Wayne (WA DBCA) and Anke Seidlitz (WA DBCA) for providing samples.

REFERENCES

1 **Diaz S**, **Settle J**, **Brondizio ES et al. IPBES**. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. 2019; Bonn, Germany.

2 **Mittermeier RA**, **Mittermeier CG**, Megadiversity: Earth's Biologically Wealthiest Nations. Mexico City, Mexico: Cemex, 2005.

3 **Chapman AD**, Numbers of Living Species in Australia and the World. 2nd ed., Canberra, Australia: Australian Government, Department of the Environment and Energy, 2009; ISBN:9780642568618.

4 **Woinarski JCZ**, **Burbidge AA**, **Harrison PL**, Ongoing unraveling of a continental fauna: decline and extinction of Australian mammals since European settlement. ***Proc. Natl. Acad. Sci. USA***, 2015; **112**(15): 4531–4540. doi:10.1073/pnas.1417301112.

5 **IUCN: The IUCN Red List of Threatened Species**. 2021; https://www.iucnredlist.org. Accessed 18 Aug 2021.

6 **Brandies PA**, **Peel E**, **Hogg CJ et al.** The value of reference genomes in the conservation of threatened species. ***Genes***, 2019; **10**(11): 846.

7 **Lewin H**, **Robinson G**, **Kress WJ et al.** Earth BioGenome Project: sequencing life for the future of life. ***Proc. Natl. Acad. Sci. USA***, 2018; **115**(17): 4325–4333. doi:10.1073/pnas.1720115115.

8 **Genome 10K Community of Scientists**. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. ***J. Heredity.***, 2009; **100**(6): 659–674.

9 **English AC**, **Richards S**, **Han Y et al.** Mind the gap: upgrading Genomes with Pacific Biosciences RS long-read sequencing technology. ***PLoS One***, 2012; **7**(11): e47768, doi:10.1371/journal.pone.0047768.

10 **Rhie A**, **McCarthy SA**, **Fedrigo O et al.** Towards complete and error-free genome assemblies of all vertebrate species. ***Nature***, 2021; **592**(7856): 737–746. doi:10.1038/s41586-021-03451-0.

11 **Whibley A**, **Kelley JL**, **Narum SR**, The changing face of genome assemblies: Guidance on achieving high-quality reference genomes. ***Mol. Ecol. Resour.***, 2021; **21**(3): 641–652. doi:10.1111/1755-0998.13312.

12 **Hotaling S**, **Kelley JL**, The rising tide of high-quality genomic resources. ***Mol. Ecol. Resour.***, 2019; **19**: 567–569.

13 **Claridge A**, **Seebeck J**, **Rose R**, Bettongs, potoroos and the musky rat-kangaroo. Collingwood, Victoria: 2007; ISBN:9780643095083.

14 **Bininda-Emonds ORP**, **Cardillo M**, **Jones KE et al.** The delayed rise of present-day mammals. ***Nature***, 2007; **445**: 507–512.

15 **Stannard HJ**, **Miller RD**, **Old JM**, Marsupial and monotreme milk—a review of its nutrient and immune properties. ***PeerJ***, 2020; **8**: e9335. doi:10.7717/peerj.9335.

16 **Garkaklis MJ**, **Bradley JS**, **Wooller RD**, Digging and soil turnover by a mycophagous marsupial. ***J. Arid Environ.***, 2004; **56**(3): 569–578. doi:10.1016/S0140-1963(03)00061-2.

17 **Zosky KL**, **Wayne AF**, **Byrant KA et al.** Diet of the critically endangered woylie (*Bettongia penicillata ogilbyi*) in south-western Australia. ***Aust. J. Zool.***, 2017; **65**: 302–312.

18 **Dundas SJ**, **Hopkins AJM**, **Ruthrof KX et al.** Digging mammals contribute to rhizosphere fungal community composition and seedling growth. ***Biodivers. Conserv.***, 2018; **27**: 3071–3086.

19 **Garkaklis MJ**, **Bradley JS**, **Wooller RD**, The effects of Woylie (Bettongia penicillata) foraging on soil water repellency and water infiltration in heavy textured soils in southwestern Australia. ***Aust. J. Ecol.***, 1998; **23**(5): 492–496. doi:10.1111/j.1442-9993.1998.tb00757.x.

20 **Palmer BJ**, **Beca G**, **Erickson TE et al.** New evidence of seed dispersal identified in Australian mammals. ***Wildl. Res.***, 2021; doi:10.1071/WR21015.

21 **Groom C**, Justifiation for continued conservation efforts following the delisting of a threatened species: a case study of the woylie, *Bettongia penicillata ogilbyi* (Marsupialia: Potoroidae). ***Wildl. Res.***, 2010; **37**: 183–193.

22 **Wayne AF**, **Maxwell MA**, **Ward CG et al.** Importace of getting the numbers right: quantifying the rapid and substantial decline of an abundant marsupial, *Bettongia penicillata*. ***Wildl. Res.***, 2013; **40**: 169–183.

23 **Thompson CK**, **Wayne AF**, **Godfrey SS et al.** Temporal and spatial dynamics o trypanosomes infecting the brush-tailed bettong (*Bettongia penicillata*): a cautionary note of disease-induced population decline. ***Parasites Vectors***, 2014; **7**: 169.

24 **Botero A**, **Thompson CK**, **Peacock CS et al.** Trypanosomes genetic diversity, polyparasitism and the population decline of the critically endangered Australian marsupial, the brush tailed bettong or woylie (*Bettongia penicillata*). ***Int. J. Parasitol. Parasites Wildl.***, 2013; **2**: 77–89.

25 **Cooper C**, **Keatley S**, **Northover A et al.** Next generation sequencing reveals widespread trypanosome diversity and polyparasitism in marsupials from Western Australia. ***Int. J. Parasitol. Parasites Wildl.***, 2018; **7**: 58–67.

26 **Pacioni C**, **Wayne AF**, **Spencer PBS**, Genetic outcomes from the translocations of the critically endangered woylie. ***Curr. Zool.***, 2013; **59**(3): 294–310.

27 **Pacioni C**, The population and epidemiological dynamics associated with recent decline of woylies (*Bettongia penicillata*) in Australia. Murdoch University, 2010.

28 **Farquharson KA**, **McLennan EA**, **Wayne A et al.** Metapopulation management of a critically endangered marsupial in the age of genomics. ***Glob. Ecol. Conserv.***, 2021; **31**: e01869.

29 **PACBIO**. SMRT link software installation (v9.0). Pacific Biosciences, CA; 2020. https://www.pacb.com/wp-content/uploads/SMRT_Link_Installation_v90.pdf.

30 **Andrews S**, FastQC. A quality control analysis tool for high throughput sequencing data. 2010; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

31 **Pacific Biosciences**. pbipa (version 1.1.2). 2020; https://github.com/PacificBiosciences/pbipa.

32 **Bushnell B.**, BBMap. 2014; https://sourceforge.net/projects/bbmap/.

33 **Seppey M**, **Manni M**, **Zdobnov EM**, BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M (ed.), Gene Prediction: Methods and Protocols. New York, NY: Springer, 2019; pp. 227–245.

34 **Guan D**, **McCarthy SA**, **Wood J et al.** Identifying and removing haplotypic duplication in primary genome assemblies. ***Bioinformatics***, 2020; **36**(9): 2896–2898. doi:10.1093/bioinformatics/btaa025.

35 **Guan DF**, purge_dups (version 1.0.1). 2020; https://github.com/dfguan/purge_dups.

36 **Marks P**, **Garcia S**, **Barrio AM et al.** Resolving the full spectrum of human genome variation using linked-reads. ***Genome Res.***, 2019; **29**(4): 635–645. doi:10.1101/gr.234443.118.

37 **Walker BJ**, **Abeel T**, **Shea T et al.** Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. ***PLoS One***, 2014; **9**(11): e112963. doi:10.1371/journal.pone.0112963.

38 **Smit A**, **Hubley R**, **Green P**, RepeatModeler Open-1.0. 2008–2015; http://www.repeatmasker.org.

39 **Smit A**, **Hubley R**, **Green P**, RepeatMasker Open-4.0. 2013–2015; http://www.repeatmasker.org.

40 **Simao FA**, **Waterhouse RM**, **Ioannidis P et al.** BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. ***Bioinformatics***, 2015; **31**(19): 3210–3212.

41 **Solovyev V**, **Kosarev P**, **Seledsov I et al.** Automatic annotation of eukaryotic genes, pseudogenes and promoters. ***Genome Biol.***, 2006; **7**(Suppl. 1): S10, doi:10.1186/gb-2006-7-s1-s10.

42 **National Center for Biotechnology Information (NCBI)**. Ref-seq non-redundant proteins database. https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/. Accessed 2 Aug 2020.

43 **Card D**, Genestats (version 1.0). 2018; https://gist.github.com/darencard/fcb32168c243b92734e85c5f8b59a1c3.

44 **Bolger AM**, **Lohse M**, **Usadel B**, Trimmomatic: a flexible trimmer for Illumina sequence data. ***Bioinformatics***, 2014; **30**(15): 2114–2120.

45 **Kim D**, **Paggi JM**, **Park C et al.** Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. ***Nat. Biotechnol.***, 2019; **37**(8): 907–915. doi:10.1038/s41587-019-0201-4.

46 **Danecek P**, **Bonfield JK**, **Liddle J et al.** Twelve years of SAMtools and BCFtools. ***Gigascience***, 2021; **10**: giab008. doi:10.1093/gigascience/giab008.

47 **Kovaka S**, **Zimin AV**, **Pertea GM et al.** Transcriptome assembly from long-read RNA-seq alignments with StringTie2. ***Genome Biol.***, 2019; **20**(1): 278, doi:10.1186/s13059-019-1910-1.

48 **Kuo RI**, **Cheng Y**, **Zhang R et al.** Illuminating the dark side of the human transcriptome with long read transcript sequencing. ***BMC Genomics***, 2020; **21**(1): 751, doi:10.1186/s12864-020-07123-7.

49 **Genome RIK**, Tama. 2019; https://github.com/GenomeRIK/tama/.

50 **Kang Y-J**, **Yang D-C**, **Kong L et al.** CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. ***Nucleic Acids Res.***, 2017; **45**(W1): W12–W16. doi:10.1093/nar/gkx428.

51 **Center for Biotechnology, Peking University**. Coding potential calculator 2. 2017; http://cpc2.gao-lab.org/.

52 **Haas BJ**, **Papanicolaou A**, **Yassour M et al.** De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. ***Nat. Protocols***, 2013; **8**(8): 1494–1512.

53 **US National Library of Medicine, National Center for Biotechnology Information**. Basic Local Alignment Search Tool (BLAST). https://blast.ncbi.nlm.nih.gov/Blast.cgi.

54 **UniProt Consortium**. UniProt KB. 2021; https://www.uniprot.org/uniprot/.

55 **Langmead B**, **Salzberg SL**, Fast gapped-read alignment with Bowtie2. ***Nat. Methods***, 2012; **9**(4): 357–359.

56 **Patro R**, **Duggal G**, **Love MI et al.** Salmon provides fast and bias-aware quantification of transcript expression. ***Nat. Methods***, 2017; **14**(4): 417–419. doi:10.1038/nmeth.4197.

57 **H Mi**, **Ebert D**, **Muruganujan A et al.** PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. ***Nucleic Acids Res.***, 2021; **49**(D1): D394–D403. doi:10.1093/nar/gkaa1106.

58 **Renfree MB**, **Papenfuss AT**, **Deakin JE et al.** Genome sequence of an Australian kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. ***Genome Biol.***, 2011; **12**: 2–25.

59 **Brandies PA**, **Tang S**, **Johnson RSP et al.** The first Antechinus reference genome provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies. ***Gigabyte***, 2020; **2020**: doi:10.46471/gigabyte.7.

60 **Johnson RN**, **O'Meally D**, **Chen Z et al.** Adaptation and conservation insights from the koala genome. ***Nat. Genet.***, 2018; **50**(8): 1102–1111. doi:10.1038/s41588-018-0153-5.

61 **Dudchenko O**, **Batra SS**, **Omer AD et al.** De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. ***Sci. Adv.***, 2017; **356**: 92–95.

62 **Dudchenko O**, **Shamim MS**, **Batra SS et al.** The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv. 2018; 254797. doi:10.1101/254797.

63 **Murchison EP**, **Schulz-Trieglaff OB**, **Ning Z et al.** Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. ***Cell***, 2012; **148**: 780–791.

64 **Margulies EH**, **Program NCS**, **Maduro VVB et al.** Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. ***Proc. Natl. Acad. Sci. USA***, 2005; **102**(9): 3354–3359. doi:10.1073/pnas.0408539102.

65 **Gentles AJ**, **Wakefield MJ**, **Kohany O et al.** Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. ***Genome Res.***, 2007; **17**(7): 992–1004. doi:10.1101/gr.6070707.

66 **Gallus S**, **Hallström BM**, **Kumar V et al.** Evolutionary histories of transposable elements in the genome of the largest living marsupial carnivore, the Tasmanian devil. ***Mol. Biol. Evol.***, 2015; **32**(5): 1268–1283. doi:10.1093/molbev/msv017.

67 **Deininger P**, Alu elements: know the SINEs. ***Genome Biol.***, 2011; **12**(12): 236. doi:10.1186/gb-2011-12-12-236.

68 **Mudge JM**, **Harrow J**, The state of play in higher eukaryote gene annotation. ***Nat. Rev. Genet.***, 2016; **17**(12): 758–772. doi:10.1038/nrg.2016.119.

69 **Salzberg SL**, Next-generation genome annotation: we still struggle to get it right. ***Genome Biol.***, 2019; **20**(1): 92. doi:10.1186/s13059-019-1715-2.

70 **National Center for Biotechnology Information**. The NCBI Eukaryotic Genome Annotation Pipeline. 2021; https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/. Accessed 19th July 2021.

71 **Callewaert L**, **Michiels CW**, Lysozymes in the animal kingdom. ***J. Biosci.***, 2010; **35**(1): 127–160. doi:10.1007/s12038-010-0015-5.

72 **Ryckmann C**, **Vandal K**, **Rouleau P et al.** Profinlammatory activities of S100: proteins S100A8, S100A9, and S100A8/A9 induce neutrophl chemotaxis and adhesion. ***J. Immunol.***, 2003; **170**: 3233–3242.

73 **Sohnle PG**, **Hunter MJ**, **Hahn B et al.** Zinc-reversible antimicrobial activity of recombinant calprotectin (migration inhibitory factor—related proteins 8 and 14). ***J. Infect. Dis.***, 2000; **182**(4): 1272–1275. doi:10.1086/315810.

74 **Büchau AS**, **Hassan M**, **Kukova G et al.** S100A15, an antimicrobial protein of the skin: regulation by E. coli through toll-like receptor 4. ***J. Invest. Dermatol.***, 2007; **127**(11): 2596–2604. doi:10.1038/sj.jid.5700946.

75 **Crouch EC**, Surfactant protein-D and pulmonary host defense. ***Respir. Res.***, 2000; **1**(2): 93–108. doi:10.1186/rr19.

76 **Wang S**, **Song R**, **Wang Z et al.** S100A8/A9 in inflammation. ***Front. Immunol.***, 2018; **9**: 1298. doi:10.3389/fimmu.2018.01298.

77 **Tyndale-Biscoe CH**, Life of marsupials. Collingwood: CSIRO Publishing, 2005.

78 **Edwards MJ**, **Hinds LA**, **Deane EM et al.** A review of complimentary mechanisms which protect the developing marsupial pouch young. ***Dev. Comp. Immunol.***, 2012; **27**: 212–220.

79 **Peel E**, **Cheng Y**, **Djordjevic JT et al.** Cathelicidins in the Tasmanian devil (*Sarcophilus harrisii*). ***Sci. Rep.***, 2016; **6**: e35019.

80 **Cheng Y**, **Heasman K**, **Peck S et al.** Significant decline in anticancer immune capacity during puberty in the Tasmanian devil. ***Sci. Rep.***, 2017; **7**: e44716.

81 **Hewavisenti RV**, **Morris KM**, **O'Meally D et al.** The identification of immune genes in the milk transcriptome of the Tasmanian devil (*Sarcophilus harrisii*). ***PeerJ***, 2016; **4**: e1569.

82 **Morris KM**, **O'Meally D**, **Zaw T et al.** Characterisation of the immune compounds in koala milk using a combined transcriptomic and proteomic approach. ***Sci. Rep.***, 2016; **6**: e35011.

83 **Cheng Y**, **Fox S**, **Pemberton D et al.** The Tasmanian devil microbiome - implications for conservation and management. ***Microbiome***, 2015; **3**(1): 76.

84 **Weiss S**, **Taggart D**, **Smith I et al.** Host reproductive cycle influences the pouch microbiota of wild southern hairy-nosed wombats (*Lasiorhinus latifrons*). ***Animal Microbiome***, 2021; **3**(1): 13. doi:10.1186/s42523-021-00074-8.

85 **Peel E**, **Cheng Y**, **Djordjevic JT et al.** Marsupial and monotreme cathelicidins display antimicrobial activity, including against methicillin-resistant *Staphylococcus aureus*. ***Microbiology***, 2017; **163**: 1457–1465.

86 **Peel E**, **Cheng Y**, **Djordjevic JT et al.** Koala cathelicidin PhciCath5 has antimicrobial activity, including against *Chlamydia pecorum*. ***PLoS One***, 2021; **16**(4): e0249658. doi:10.1371/journal.pone.0249658.

87 **Feigin CY**, **Newton AH**, **Doronina L et al.** Genome of the Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore. ***Nat. Ecol. Evol.***, 2018; **2**(1): 182–192. doi:10.1038/s41559-017-0417-y.

88 **Conesa A**, **Madrigal P**, **Tarazona S et al.** A survey of best practices for RNA-seq data analysis. ***Genome Biol.***, 2016; **17**: 13.

89 **Supple MA**, **Shapiro B**, Conservation of biodiversity in the genomics era. ***Genome Biol.***, 2018; **19**: 131. doi:10.1186/s13059-018-1520-3.

90 **Hohenlohe PA**, **Funk WC**, **Rajora OP**, Population genomics for wildlife conservation and management. ***Mol. Ecol.***, 2021; **30**(1): 62–82. doi:10.1111/mec.15720.

91 **Quigley BL**, **Timms P**, The koala immune response to chlamydial infection and vaccine development—advancing our immunological understanding. ***Animals***, 2021; **11**(2): 380.

92 **Wright BR**, **Farquharson KA**, **McLennan EA et al.** A demonstration of conservation genomics for threatened species management. ***Mol. Ecol. Resour.***, 2020; **20**(6): 1526–1541. doi:10.1111/1755-0998.13211.

93 **McLennan EA**, **Wright BR**, **Belov K et al.** Too much of a good thing? Finding the most informative genetic data set to answer conservation questions. ***Mol. Ecol. Resour.***, 2019; **19**(3): 659–671. doi:10.1111/1755-0998.12997.

94 **Storfer A**, **Kozakiewicz CP**, **Beer MA et al.** Applications of population genomics for understanding and mitigating wildlife disease. In: Hohenlohe PA, Rajora OP (eds), Population Genomics: Wildlife. Cham: Springer International Publishing, 2021; pp. 357–383.

95 **Frankham R**, **Ballou JD**, **Briscoe DA**, Introduction to Conservation Genetics. Cambridge, UK: Cambridge University Press, 2002.

96 **Wayne AF**, **Maxwell MA**, **Ward CG et al.** Sudden and rapid decline of the adundant marsupial *Bettongia penicillata* in Australia. ***ORYX***, 2015; **49**(1): 175–185.

97 **AWGG-Lab**. Woylie (*Bettongia penicillata ogilbyi*). Genomic data repository for the Threatened Species Initiative and the ARC Centre for Innovations in Peptides and Protein Science. 2021; https://awgg-lab.github.io/australasiangenomes/species/Bettongia_penicillata_ogilbyi.html.

98 **Peel E**, **Silver L**, **Brandies PA et al.** Supporting data for "A reference genome for the critically endangered woylie, *Bettongia penicillata ogilbyi*". ***GigaScience Database***. 2021; http://dx.doi.org/10.5524/100951.