1 A reference genome for the critically endangered woylie,

2 *Bettongia penicillata ogilbyi*

3

4 Emma Peel¹, Luke Silver¹, Parice Brandies¹, Carolyn J Hogg¹, Katherine Belov^{1,*}

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

6 Wales, Australia

7 emma.peel@sydney.edu.au ORCID: 0000-0002-2335-898

8 luke.silver@sydney.edu.au ORCID: 0000-0002-1718-5756

9 parice.brandies@sydney.edu.au ORCID: 0000-0003-1702-2938

10 carolyn.hogg@sydney.edu.au ORCID: 0000-0002-6328-398X

11 *kathy.belov@sydney.edu.au ORCID: 0000-0002-9762-5554

12

13 **Abstract**

14 Biodiversity is declining globally, and Australia has one of the worst extinction records for
 15 mammals. The development of sequencing technologies means that genomic approaches are
 16 now available as important tools for wildlife conservation and management. Despite this,
 17 genome sequences are available for only 5% of threatened Australian species. Here we report
 18 the first reference genome for the woylie (*Bettongia penicillata ogilbyi*), a critically
 19 endangered marsupial from Western Australia, and the first genome within the Potoroidae
 20 family. The woylie reference genome was generated using Pacific Biosciences HiFi long-
 21 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.

27

Research Area: Genetics and Genomics

Classifications: Animal Genetics, Genetics

30

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

45 highly contiguous genomes, which can then be used to answer various biological,
46 evolutionary and conservation questions [10–12].

47 The woylie, or brush-tailed bettong (*Bettongia penicillata ogilbyi*, NCBI:txid881300), is a
48 small marsupial of the Potoroidae family (Figure 1) [13]. Marsupials are one of three lineages
49 of mammals, the others being eutherians (e.g. humans and mice) and monotremes (e.g.
50 platypus and echidna). Since their divergence from eutherian mammals around 156 million
51 years ago, marsupials have evolved into over 300 species, most of which are endemic to
52 Australia [14]. Marsupials differ to other mammals in several ways; the most prominent is
53 their pouch. After a short gestation of only 20 days, woylies give birth to altricial young,
54 which develop within the pouch for 100 days [13]. The complex milk profile of the mother
55 provides nutrient and immune support throughout pouch life [15]. Similar to wallabies and
56 kangaroos, woylies undergo embryonic diapause, whereby embryonic development is
57 suspended by the suckling pouch young, and resumes when the young exits the pouch [13].
58 Woylies are ecosystem engineers, with individuals displacing on average 4.8 tonnes of soil
59 per year [16]. This bioturbation is essential for ecosystem health and function, as it activates
60 and disperses the mycorrhizal fungi comprising a large portion of their diet [17, 18], alters
61 soil nutrient composition and water penetration [19], and aids seed dispersal [20].

62

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

65

66 Historically, woylies inhabited much of central and southern Australia; however, populations
67 have contracted to 1% of their former range owing to habitat loss and fragmentation [4, 21].
68 Between 1999 and 2006, woylie populations significantly declined by more than 90% [22],

69 resulting in their current IUCN listing as critically endangered [5]. The decline is thought to
 70 be caused by a combination of predation by introduced feral cats (*Felis catus*) and foxes
 71 (*Vulpes vulpes*), and an unknown disease linked to parasites such as trypanosomes [23–25].
 72 Currently, there are two remaining indigenous populations in the Upper Warren and
 73 Dryandra regions of Western Australia [22]. In response to the decline, several translocations
 74 and reintroductions have been conducted within Western Australia (WA), South Australia
 75 (SA) and New South Wales (NSW), guided by genetic assessments of diversity and
 76 population structure using microsatellite [26, 27] and genomic-based methods [28].
 77 In this study, we present the first *de novo* reference genome assembly of the woylie, as well
 78 as four tissue transcriptomes. This is the first genome sequenced within the Potoroidae
 79 family, which will be a valuable tool for investigating disease-induced decline, basic biology,
 80 and to aid conservation.

81

82 **Methods**

83 **Sample collection and sequencing**

84 Spleen, heart, kidney and tongue were opportunistically sampled from a single wild female
 85 woylie (woy01), as well as pouch skin from a second wild female woylie (woy02), both of
 86 which died by vehicle strike at Manjimup, Western Australia (WA) in 2018. In addition,
 87 500 µL of peripheral blood was collected into RNAProtect Animal Blood tubes (Qiagen)
 88 from a third wild male woylie (woy03) from Balban, WA, in 2018 during routine trapping
 89 and health examinations. All samples were collected under the Western Australian
 90 Government Department of Biodiversity, Conservation and Attractions animal ethics 2018-
 91 22F and scientific licence number NSW DPIE SL101204.

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

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

108 Total RNA was extracted from 25 mg of woy01 tongue and heart, woy02 pouch skin, using
109 the RNeasy Plus Mini Kit (Qiagen). In addition, total RNA was extracted from 500 µL of
110 woy03 peripheral blood using the RNa protect Animal Blood Kit (Qiagen). In all extractions,
111 contaminating DNA was removed through on-column digestion using the RNase-free DNase
112 I set (Qiagen). RNA purity was assessed using the NanoDrop 6000, with all samples
113 displaying an A260/280 and A260/230 of 1.9–2.2. RNA concentration and integrity were
114 measured using an RNA Nano 6000 chip (Agilent Technologies), with all samples displaying
115 an RNA integrity number (RIN) greater than 7. Total RNA was submitted to the Ramaciotti
116 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	HiFi reads	woy01 kidney	37
Sequel II			
Illumina	10x linked-reads	woy01 spleen	137
NovaSeq6000			
Illumina	RNA-seq reads	woy01 tongue and heart	23–27 per
NovaSeq6000		woy02 pouch skin	sample
		woy03 blood	

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 *de 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 Fgenes++ 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 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.

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

Metric	Woylie	Koala [60–	Antechinus	Tasmanian	Tammar

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

204

205 Repeat elements comprised 53.05% of the woylie genome, similar to the tammar wallaby
206 (*Notamacropus eugenii*) (52.8%) [58], but higher than antechinus (44.82%) [59] and koala
207 (47.5%) [60]. Repeat families numbering 1184 were identified in the woylie genome, with
208 long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)
209 being the most numerous (Table 3), as in other marsupial genomes [59, 60, 64].
210 Retrotransposon-like elements (RTE) were also identified in the woylie genome, as observed
211 in other marsupial genomes and some mammals such as ruminants [59, 65, 66]. Interestingly,
212 primate-specific ALU repeats were identified in the woylie genome. ALU repeats are a type
213 of SINE that comprise over 10% of the human genome and are involved in genome evolution
214 and disease [67]. These repeat elements contributed only 0.10% to the woylie genome, and
215 were also identified in the antechinus genome (0.04%) [59]. As this may represent an
216 inaccurate repeat annotation, further work is required to confirm the presence of ALU repeats
217 in marsupials.

218

219 Table 3. Repeats identified in the woylie genome.

Repeat type	Number	Sequence masked	
		Length	Percentage (%)
		(bp)	
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			
ERV1	16,900	7,241,092	0.21
ERV2	22,176	7,069,536	0.21
ERV3	22,364	11,265,604	0.33
DNA elements			
hAT-	141,744	23,630,550	0.70
Charlie			
TcMar-	36,287	8,389,535	0.25
Tigger			
Other			
Unclassified	1,017,337	231,493,940	6.82
Satellite	29,228	4,458,406	0.13

DNA

Small RNA   615   53,205   0.00

220

221 Fgenesh++ predicted 41,868 genes in the woylie genome, of which 24,655 had BLAST hits
 222 to eukaryote genes in the NCBI non-redundant database. Of these 24,655 genes, 15,904 were
 223 supported by mRNA evidence, and 1,309 by protein evidence. This is higher than the number
 224 of protein-coding genes annotated by NCBI in the koala genome (20,103) [60] and
 225 Tasmanian devil (20,053) [63] (Table 4). The higher number of genes annotated in the woylie
 226 genome is likely due to incomplete RNA-seq evidence, and hence gene models, used for gene
 227 prediction by Fgenesh++. In addition, fragmentation of the genome causes fragmentation of
 228 gene sequences, which can result in an overinflated gene count [68, 69]. Statistics for protein-
 229 coding genes annotated within the woylie genome also reflected deficiencies in gene models,
 230 as mean gene and exon length, and mean exon number per gene differed to the NCBI
 231 annotation of the koala and devil genome (Table 4). The NCBI annotation pipeline uses
 232 mRNA and protein evidence from multiple public scientific databases for gene prediction,
 233 resulting in an annotated gene number that more closely resembles humans (~20,000) [70].

234

235 Table 4. Statistics for protein-coding genes annotated in the woylie genome compared to
 236 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)	8,559	7,261	5,936
Mean exon number per gene	3.58	11.11	10.36

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 7,739 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).

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.

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 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 *.

278

279 **Data Validation and Quality Control**

280 BUSCO was used to assess functional completeness by searching for complete single-copy
281 gene orthologs within the genome assembly. Fgenesh++ predicted proteins and the global
282 transcriptome assembly. The genome contained 86.5% of complete mammalian BUSCOv4
283 genes, comparable to other marsupial genomes [58–60, 63, 87]. Fgenesh++-predicted
284 proteins and the global transcriptome also displayed a high level of completeness, with 80.4%
285 and 80.8% of complete mammalian BUSCOv4 identified, respectively.

286 High mapping rates were observed for both the genome and global transcriptome assemblies,
287 indicating high sequencing accuracy and low contaminating DNA. 99.8% of HiFi reads and
288 88.4% of 10x Chromium Illumina reads mapped to the genome assembly. Similarly, 80.25%
289 (blood), 70.96% (pouch skin), 65.30% (tongue) and 60.43% (heart) of RNA-seq reads
290 mapped to the global transcriptome assembly. The lower mapping rate for heart and tongue
291 against the global transcriptome is not unexpected, as reads which map to unannotated
292 transcripts are lost [88]. Alignment of reads from heart and tongue to the genome was higher,
293 with 77.79% and 81.70% of reads mapped, respectively.

294

295 **Reuse Potential**

296 Genomes are valuable tools for wildlife conservation and management [6, 89, 90]. In
297 marsupials, Tasmanian devils [63] and koalas [60] are two examples where genomes have
298 been used to investigate genetic diversity, population structure, adaptation and disease [91–
299 93]. The woylie reference genome is the first genome available for the Potoroidae family of
300 marsupials. Not only will this resource facilitate basic biological research of bettongs and
301 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].

325

326 **Declarations**

327 **List of abbreviations**

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

334

335 **Ethical Approval**

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

339

340 **Consent for publication**

341 Not applicable.

342

343 **Competing interests**

344 The authors declare that they have no competing interests.

345

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 (2019): 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.

- 369   3.   Chapman AD. Numbers of living species in Australia and the world. 2 ed. Canberra,
370   Australia: Australian Government, Department of the Environment and Energy; 2009.
371   ISBN: 978 0 642 56861 8
- 372   4.   Woinarski JCZ, Burbidge AA, Harrison PL. Ongoing unraveling of a continental
373   fauna: Decline and extinction of Australian mammals since European settlement. *Proc*
374   *Natl Acad Sci U S A*. 2015;112 15:4531. doi:10.1073/pnas.1417301112.
- 375   5.   IUCN: The IUCN Red List of Threatened Species. <https://www.iucnredlist.org>
376   (2021). Accessed 18 Aug 2021.
- 377   6.   Brandies PA, Peel E, Hogg CJ, Belov K. The value of reference genomes in the
378   conservation of threatened species. *Genes*. 2019;10 11:846.
- 379   7.   Lewin H, Robinson G, Kress WJ, et al. Earth BioGenome Project: Sequencing life for
380   the future of life. *Proc Natl Acad Sci U S A*. 2018;115 17:4325-33.
381   doi:10.1073/pnas.1720115115.
- 382   8.   Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-
383   genome sequence for 10 000 vertebrate species. *J Heredity*. 2009;100 6:659-74.
- 384   9.   English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap:
385   Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing
386   Technology. *PLOS One*. 2012;7 11:e47768. doi:10.1371/journal.pone.0047768.
- 387   10.   Rhie A, McCarthy SA, Fedrigo O, et al. Towards complete and error-free genome
388   assemblies of all vertebrate species. *Nature*. 2021;592 7856:737-46.
389   doi:10.1038/s41586-021-03451-0.
- 390   11.   Whibley A, Kelley JL, Narum SR. The changing face of genome assemblies:
391   Guidance on achieving high-quality reference genomes. *Mol Ecol Resources*. 2021;21
392   3:641-52. doi:10.1111/1755-0998.13312.

- 393 12. Hotaling S, Kelley JL. The rising tide of high-quality genomic resources. Mol Ecol
394 Resources. 2019;19:567-9.
- 395 13. Claridge A, Seebeck J, Rose R. Bettongs, potoroos and the musky rat-kangaroo.
396 2007. Collingwood, Victoria. ISBN: 9780643095083
- 397 14. Bininda-Emonds ORP, Cardillo M, Jones KE, et al. The delayed rise of present-day
398 mammals. Nature. 2007;445:507-12.
- 399 15. Stannard HJ, Miller RD, Old JM. Marsupial and monotreme milk—a review of its
400 nutrient and immune properties. PeerJ. 2020;8:e9335. doi:10.7717/peerj.9335.
- 401 16. Garkaklis MJ, Bradley JS, Wooller RD. Digging and soil turnover by a mycophagous
402 marsupial. J Arid Environ. 2004;56 3:569-78. doi:10.1016/S0140-1963(03)00061-2.
- 403 17. Zosky KL, Wayne AF, Byrant KA, et al. Diet of the critically endangered woylie
404 (*Bettongia penicillata ogilbyi*) in south-western Australia. Aust J Zool. 2017;65:302-
405 12.
- 406 18. Dundas SJ, Hopkins AJM, Ruthrof KX, Tay NE, Burgess TI, Hardy GESJ, et al.
407 Digging mammals contribute to rhizosphere fungal community composition and
408 seedling growth. Biodiversity Conservation. 2018;27:3071-86.
- 409 19. Garkaklis MJ, Bradley JS and Wooller RD. The effects of Woylie (*Bettongia*
410 *penicillata*) foraging on soil water repellency and water infiltration in heavy textured
411 soils in southwestern Australia. Aust J Ecology. 1998;23 5:492-6. doi:10.1111/j.1442-
412 9993.1998.tb00757.x.
- 413 20. Palmer BJ, Beca G, Erickson TE, et al. New evidence of seed dispersal identified in
414 Australian mammals. Wildlife Res. 2021:-. doi:10.1071/WR21015.
- 415 21. Groom C. Justification for continued conservation efforts following the delisting of a
416 threatened species: a case study of the woylie, *Bettongia penicillata ogilbyi*
417 (Marsupialia: Potoroidae). Wildlife Res. 2010;37:183-93.

- 418 22. Wayne AF, Maxwell MA, Ward CG, et al. Importace of getting the numbers right:
419 quantifying the rapid and substantial decline of an abundant marsupial, *Bettongia*
420 *penicillata*. Wildlife Res. 2013;40:169-83.
- 421 23. Thompson CK, Wayne AF, Godfrey SS, et al. Temporal and spatial dynamics o
422 trypanosomes infecting the brush-tailed bettong (*Bettongia penicillata*): a cautionary
423 note of disease-induced population decline. Parasites Vectors. 2014;7:169.
- 424 24. Botero A, Thompson CK, Peacock CS, et al. Trypanosomes genetic diversity,
425 polyparasitism and the population decline of the critically endangered Australian
426 marsupial, the brush tailed bettong or woylie (*Bettongia penicillata*). Int J Parasitol
427 Parasites Wildlife. 2013;2:77-89.
- 428 25. Cooper C, Keatley S, Northover A, et al. Next generation sequencing reveals
429 widespread trypanosome diversity and polyparasitism in marsupials from Western
430 Australia. Int J Parasitol Parasites Wildlife. 2018;7:58-67.
- 431 26. Pacioni C, Wayne AF, Spencer PBS. Genetic outcomes from the translocations of the
432 critically endangered woylie. Curr Zool. 2013;59 3:294-310.
- 433 27. Pacioni C. The population and epidemiological dynamics associated with recent
434 decline of woylies (*Bettongia penicillata*) in Australia. Murdoch University, 2010.
- 435 28. Farquharson KA, McLennan EA, Wayne A, et al. Metapopulation management of a
436 critically endangered marsupial in the age of genomics. Global Ecol Conservation.
437 2021;vol 31, e01869.
- 438 29. PACBIO. SMRT Link software installation (v9.0). Pacific Biosciences, CA; 2020.
439 https://www.pacb.com/wp-content/uploads/SMRT_Link_Installation_v90.pdf.
- 440 30. Andrews S. (2010). FastQC. A quality control analysis tool for high throughput
441 sequencing data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- 442 31. Pacific Biosciences (2020). pbipa (version 1.1.2).
443 <https://github.com/PacificBiosciences/pbipa>.
- 444 32. Bushnell B. (2014). BBMap. <https://sourceforge.net/projects/bbmap/>.
- 445 33. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and
446 Annotation Completeness. In: Kollmar M, editor. Gene Prediction: Methods and
447 Protocols. New York, NY: Springer New York; 2019. p. 227-45.
- 448 34. Guan D, McCarthy SA, Wood J, et al. Identifying and removing haplotypic
449 duplication in primary genome assemblies. Bioinformatics. 2020;36 9:2896-8.
450 doi:10.1093/bioinformatics/btaa025.
- 451 35. Guan DF. (2020). purge_dups (version 1.0.1). https://github.com/dfguan/purge_dups.
- 452 36. Marks P, Garcia S, Barrio AM, et al. Resolving the full spectrum of human genome
453 variation using Linked-Reads. Genome Res. 2019;29 4:635-45.
454 doi:10.1101/gr.234443.118.
- 455 37. Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive
456 Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE.
457 2014;9 11:e112963. doi:10.1371/journal.pone.0112963.
- 458 38. Smit A, Hubley R, Green P (2008–2015). RepeatModeler Open-1.0.
459 <http://www.repeatmasker.org>.
- 460 39. Smit A, Hubley R, Green P (2013–2015). RepeatMasker Open-4.0.
461 <http://www.repeatmasker.org>.
- 462 40. Simao FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly
463 and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31
464 19:3210-2.

- 465 41. Solovyev V, Kosarev P, Seledsov I, et al. Automatic annotation of eukaryotic genes,
466 pseudogenes and promoters. *Genome Biol.* 2006;7 Suppl 1:S10. doi:10.1186/gb-
467 2006-7-s1-s10.
- 468 42. National Center for Biotechnology Information (NCBI). Ref-seq non-redundant
469 proteins database. <https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/>.
470 Accessed 2 Aug 2020.
- 471 43. Card D. (2018). Genestats (version 1.0).
472 <https://gist.github.com/darencard/fcb32168c243b92734e85c5f8b59a1c3>.
- 473 44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
474 sequence data. *Bioinformatics.* 2014;30 15:2114-20.
- 475 45. Kim D, Paggi JM, Park C, et al. Graph-based genome alignment and genotyping with
476 HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37 8:907-15.
477 doi:10.1038/s41587-019-0201-4.
- 478 46. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools.
479 *Gigascience.* 2021 Feb 16;10:giab008. doi: 10.1093/gigascience/giab008.
- 480 47. Kovaka S, Zimin AV, Pertea GM, et al. Transcriptome assembly from long-read
481 RNA-seq alignments with StringTie2. *Genome Biol.* 2019;20 1:278.
482 doi:10.1186/s13059-019-1910-1.
- 483 48. Kuo RI, Cheng Y, Zhang R, et al. Illuminating the dark side of the human
484 transcriptome with long read transcript sequencing. *BMC Genomics.* 2020;21 1:751.
485 doi:10.1186/s12864-020-07123-7.
- 486 49. GenomeRIK. (2019). Tama. <https://github.com/GenomeRIK/tama/>.
- 487 50. Kang Y-J, Yang D-C, Kong L, et al. CPC2: a fast and accurate coding potential
488 calculator based on sequence intrinsic features. *Nucleic Acids Res.* 2017;45 W1:W12-
489 W6. doi:10.1093/nar/gkx428.

- 490 51. Center for Biotechnology, Peking University (2017). Coding Potential Calculator 2.
491 <http://cpc2.gao-lab.org/>.
- 492 52. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence
493 reconstruction from RNA-seq using the Trinity platform for reference generation and
494 analysis. Nat Protocols. 2013;8 8:1494-512.
- 495 53. US National Library of Medicine, National Center for Biotechnology Information.
496 Basic Local Alignment Search Tool (BLAST).
497 <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.
- 498 54. UniProt Consortium. (2021). UniProt KB. <https://www.uniprot.org/uniprot/>.
- 499 55. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie2. Nat Methods.
500 2012;9 4:357-9.
- 501 56. Patro R, Duggal G, Love MI, et al. Salmon provides fast and bias-aware
502 quantification of transcript expression. Nat Methods. 2017;14 4:417-9.
503 doi:10.1038/nmeth.4197.
- 504 57. Mi H, Ebert D, Muruganujan A, et al. PANTHER version 16: a revised family
505 classification, tree-based classification tool, enhancer regions and extensive API.
506 Nucleic Acids Res. 2021;49 D1:D394-D403. doi:10.1093/nar/gkaa1106.
- 507 58. Renfree MB, Papenfuss AT, Deakin JE, et al. Genome sequence of an Australian
508 kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian
509 reproduction and development. Genome Biol. 2011;12:2-25.
- 510 59. Brandies PA, Tang S, Johnson RSP, et al. The first Antechinus reference genome
511 provides a resource for investigating the genetic basis of semelparity and age-related
512 neuropathologies. Gigabyte. 2020;2020:0. doi:10.46471/gigabyte.7.
- 513 60. Johnson RN, O'Meally D, Chen Z, et al. Adaptation and conservation insights from
514 the koala genome. Nat Genet. 2018;50 8:1102-11. doi:10.1038/s41588-018-0153-5.

- 515 61. Dudchenko O, Batra SS, Omer AD, et al. De novo assembly of the *Aedes aegypti*
516 genome using Hi-C yields chromosome-length scaffolds. *Sci Adv.* 2017;356:92-5.
- 517 62. Dudchenko O, Shamim MS, Batra SS, et al. The Juicebox Assembly Tools module
518 facilitates de novo assembly of mammalian genomes with chromosome-length
519 scaffolds for under \$1000. *bioRxiv.* 2018:254797. doi:10.1101/254797.
- 520 63. Murchison EP, Schulz-Trieglaff OB, Ning Z, et al. Genome sequencing and analysis
521 of the Tasmanian devil and its transmissible cancer. *Cell.* 2012;148:780-91.
- 522 64. Margulies EH, Program NCS, Maduro VVB, et al. Comparative sequencing provides
523 insights about the structure and conservation of marsupial and monotreme genomes.
524 *Proc Natl Acad Sci U S A.* 2005;102 9:3354. doi:10.1073/pnas.0408539102.
- 525 65. Gentles AJ, Wakefield MJ, Kohany O, et al. Evolutionary dynamics of transposable
526 elements in the short-tailed opossum *Monodelphis domestica*. *Genome Res.* 2007;17
527 7:992-1004. doi:10.1101/gr.6070707.
- 528 66. Gallus S, Hallström BM, Kumar V, et al. Evolutionary Histories of Transposable
529 Elements in the Genome of the Largest Living Marsupial Carnivore, the Tasmanian
530 Devil. *Mol Biol Evol.* 2015;32 5:1268-83. doi:10.1093/molbev/msv017.
- 531 67. Deininger P. Alu elements: know the SINEs. *Genome Biol.* 2011;12 12:236.
532 doi:10.1186/gb-2011-12-12-236.
- 533 68. Mudge JM, Harrow J. The state of play in higher eukaryote gene annotation. *Nat Rev*
534 *Genet.* 2016;17 12:758-72. doi:10.1038/nrg.2016.119.
- 535 69. Salzberg SL. Next-generation genome annotation: we still struggle to get it right.
536 *Genome Biol.* 2019;20 1:92. doi:10.1186/s13059-019-1715-2.
- 537 70. National Center for Biotechnology Information (2021). The NCBI Eukaryotic
538 Genome Annotation Pipeline.

- 539   https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/. Accessed 19th July
- 540   2021.
- 541   71.   Callewaert L, Michiels CW. Lysozymes in the animal kingdom. J Biosci. 2010;35
- 542   1:127-60. doi:10.1007/s12038-010-0015-5.
- 543   72.   Ryckmann C, Vandal K, Rouleau P, et al. Proinflammatory activities of S100:
- 544   proteins S100A8, S100A9, and S100A8/A9 induce neutrophil chemotaxis and
- 545   adhesion. J Immunol. 2003;170:3233-42.
- 546   73.   Sohnle PG, Hunter MJ, Hahn B, et al. Zinc-Reversible Antimicrobial Activity of
- 547   Recombinant Calprotectin (Migration Inhibitory Factor—Related Proteins 8 and 14).
- 548   J Infect Dis. 2000;182 4:1272-5. doi:10.1086/315810.
- 549   74.   Büchau AS, Hassan M, Kukova G, et al. S100A15, an Antimicrobial Protein of the
- 550   Skin: Regulation by E. coli through Toll-Like Receptor 4. J Invest Dermatol.
- 551   2007;127 11:2596-604. doi:10.1038/sj.jid.5700946.
- 552   75.   Crouch EC. Surfactant protein-D and pulmonary host defense. Respiratory Res.
- 553   2000;1 2:6. doi:10.1186/rr19.
- 554   76.   Wang S, Song R, Wang Z, et al. S100A8/A9 in Inflammation. Front Immunol. 2018;9
- 555   1298 doi:10.3389/fimmu.2018.01298.
- 556   77.   Tyndale-Biscoe CH. Life of marsupials. Collingwood: CSIRO Publishing; 2005.
- 557   78.   Edwards MJ, Hinds LA, Deane EM, et al. A review of complimentary mechanisms
- 558   which protect the developing marsupial pouch young. Develop Comparative
- 559   Immunol. 2012;27:212-20.
- 560   79.   Peel E, Cheng Y, Djordjevic JT, et al. Cathelicidins in the Tasmanian devil
- 561   (*Sarcophilus harrisii*). Sci Rep. 2016;6:e35019.
- 562   80.   Cheng Y, Heasman K, Peck S, et al. Significant decline in anticancer immune
- 563   capacity during puberty in the Tasmanian devil. Sci Rep. 2017;7:e44716.

- 564 81. Hewavisenti RV, Morris KM, O'Meally D, et al. The identification of immune genes
565 in the milk transcriptome of the Tasmanian devil (*Sarcophilus harrisii*). PeerJ.
566 2016;4:1569.
- 567 82. Morris KM, O'Meally D, Zaw T, et al. Characterisation of the immune compounds in
568 koala milk using a combined transcriptomic and proteomic approach. Sci Rep.
569 2016;6:e35011.
- 570 83. Cheng Y, Fox S, Pemberton D, et al. The Tasmanian devil microbiome - implications
571 for conservation and management. Microbiome. 2015;3 1:76.
- 572 84. Weiss S, Taggart D, Smith I, et al. Host reproductive cycle influences the pouch
573 microbiota of wild southern hairy-nosed wombats (*Lasiorninus latifrons*). Animal
574 Microbiome. 2021;3 1:13. doi:10.1186/s42523-021-00074-8.
- 575 85. Peel E, Cheng Y, Djordjevic JT, et al. Marsupial and monotreme cathelicidins display
576 antimicrobial activity, including against methicillin-resistant *Staphylococcus aureus*.
577 Microbiology. 2017;163:1457-65.
- 578 86. Peel E, Cheng Y, Djordjevic JT, et al. Koala cathelicidin PhciCath5 has antimicrobial
579 activity, including against *Chlamydia pecorum*. PLOS One. 2021;16 4:e0249658.
580 doi:10.1371/journal.pone.0249658.
- 581 87. Feigin CY, Newton AH, Doronina L, et al. Genome of the Tasmanian tiger provides
582 insights into the evolution and demography of an extinct marsupial carnivore. Nat
583 Ecol Evol. 2018;2 1:182-92. doi:10.1038/s41559-017-0417-y.
- 584 88. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data
585 analysis. Genome Biol. 2016;17:13.
- 586 89. Supple MA, Shapiro B. Conservation of biodiversity in the genomics era. Genome
587 Biol. 2018;19 1 doi:10.1186/s13059-018-1520-3.

- 588 90. Hohenlohe PA, Funk WC, Rajora OP. Population genomics for wildlife conservation
589 and management. *Mol Ecol.* 2021;30 1:62-82. doi:10.1111/mec.15720.
- 590 91. Quigley BL, Timms P. The Koala Immune Response to Chlamydial Infection and
591 Vaccine Development—Advancing Our Immunological Understanding. *Animals.*
592 2021;11 2:380.
- 593 92. Wright BR, Farquharson KA, McLennan EA, et al. A demonstration of conservation
594 genomics for threatened species management. *Mol Ecol Resources.* 2020;20 6:1526-
595 41. doi:10.1111/1755-0998.13211.
- 596 93. McLennan EA, Wright BR, Belov K, et al. Too much of a good thing? Finding the
597 most informative genetic data set to answer conservation questions. *Mol Ecol*
598 *Resources.* 2019;19 3:659-71. doi:10.1111/1755-0998.12997.
- 599 94. Storfer A, Kozakiewicz CP, Beer MA, et al. Applications of Population Genomics for
600 Understanding and Mitigating Wildlife Disease. In: Hohenlohe PA and Rajora OP,
601 editors. *Population Genomics: Wildlife.* Cham: Springer International Publishing;
602 2021. p. 357-83.
- 603 95. Frankham R, Ballou JD, Briscoe DA. Introduction to conservation genetics.
604 Cambridge, United Kingdom: Cambridge University Press; 2002.
- 605 96. Wayne AF, Maxwell MA, Ward CG, et al. Sudden and rapid decline of the abundant
606 marsupial *Bettongia penicillata* in Australia. *ORYX.* 2015;49 1:175-85.
- 607 97. AWGG-Lab. (2021). Woylie (*Bettongia penicillata ogilbyi*). Genomic data repository
608 for the Threatened Species Initiative and the ARC Centre for Innovations in Peptides
609 and Protein Science. [https://awgg-](https://awgg-lab.github.io/australasiangenomes/species/Bettongia_penicillata_ogilbyi.html)
610 [lab.github.io/australasiangenomes/species/Bettongia_penicillata_ogilbyi.html](https://awgg-lab.github.io/australasiangenomes/species/Bettongia_penicillata_ogilbyi.html).

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

614

A

Biological processes

B

Immune system process