Author/s:
Ibeh, N;Feigin, CY;Frankenberg, SR;McCarthy, DJ;Pask, AJ;Romero, IG

Title:
De novo transcriptome assembly and genome annotation of the fat-tailed dunnart (Sminthopsis crassicaudata)

Date:
2024-01-01

Citation:
Ibeh, N., Feigin, C. Y., Frankenberg, S. R., McCarthy, D. J., Pask, A. J. & Romero, I. G. (2024). De novo transcriptome assembly and genome annotation of the fat-tailed dunnart (Sminthopsis crassicaudata). Gigabyte, 2024, https://doi.org/10.46471/gigabyte.118.

Persistent Link:
https://hdl.handle.net/11343/353630

(GIGA)byte

DATA RELEASE

De novo transcriptome assembly and genome annotation of the fat-tailed dunnart (*Sminthopsis crassicaudata*)

Neke Ibeh[1,2,3,4,*], Charles Y. Feigin[1,5], Stephen R. Frankenberg[1], Davis J. McCarthy[2,3], Andrew J. Pask[1] and Irene Gallego Romero[1,2,4,6,*]

1 School of BioSciences, The University of Melbourne, Parkville, VIC, Australia
2 Melbourne Integrative Genomics, The University of Melbourne, Parkville, VIC, Australia
3 Bioinformatics and Cellular Genomics, St Vincent's Institute of Medical Research, Fitzroy, VIC, Australia
4 Human Genomics and Evolution, St Vincent's Institute of Medical Research, Fitzroy, VIC, Australia
5 Department of Environment and Genetics, La Trobe University, Bundoora, VIC, Australia
6 Center for Genomics, Evolution and Medicine, Institute of Genomics, University of Tartu, Riia 23b, 51010, Tartu, Estonia

ABSTRACT

Marsupials exhibit distinctive modes of reproduction and early development that set them apart from their eutherian counterparts and render them invaluable for comparative studies. However, marsupial genomic resources still lag far behind those of eutherian mammals. We present a series of novel genomic resources for the fat-tailed dunnart (*Sminthopsis crassicaudata*), a mouse-like marsupial that, due to its ease of husbandry and *ex-utero* development, is emerging as a laboratory model. We constructed a highly representative multi-tissue *de novo* transcriptome assembly of dunnart RNA-seq reads spanning 12 tissues. The transcriptome includes 2,093,982 assembled transcripts and has a mammalian transcriptome BUSCO completeness score of 93.3%, the highest amongst currently published marsupial transcriptomes. This global transcriptome, along with *ab initio* predictions, supported annotation of the existing dunnart genome, revealing 21,622 protein-coding genes. Altogether, these resources will enable wider use of the dunnart as a model marsupial and deepen our understanding of mammalian genome evolution.

Submitted: 04 December 2023
Accepted: 13 April 2024
Published: 02 May 2024

* Corresponding authors. E-mail: oibeh@student.unimelb.edu.au; irene.gallego@svi.edu.au

Published by GigaScience Press.

Preprint submitted at https://doi.org/10.1101/2023.11.16.567318

Gigabyte, 2024, 1–16

Subjects Genetics and Genomics, Bioinformatics, Evolutionary Biology

DATA DESCRIPTION

Background and context

Marsupials are a strikingly diverse mammalian group predominantly found in Australasia (Australia, Tasmania, New Guinea, and nearby islands), with several species also inhabiting the Americas [1, 2]. While many marsupials exhibit convergent traits with eutherian mammals [3–11], their adaptations to their respective niches encompass highly specialized physiology [12–18], behavior [19–21], and modes of reproduction [22–27], thereby representing a unique component of mammalian diversity. To date, marsupial studies have significantly contributed towards elucidating various aspects of mammalian biology, including reproductive physiology [24–27], sex determination [28–34], X-chromosome inactivation [24, 35–38], age-related obesity [39], postnatal development [15–18, 40–42], and

genome evolution [43–50], among others. Consequently, marsupials represent a critical comparative model system for advancing our understanding of mammalian biology.

Despite the importance of well-developed marsupial models, marsupial genomic resources still lag far behind those of their eutherian counterparts. Currently, there are 753 eutherian reference genome assemblies available through NCBI, but only 23 marsupial reference genomes (with just 9 RefSeq-annotated species). Some of these publicly available whole-genome assemblies include the gray short-tailed opossum (*Monodelphis domestica*) [51], the tammar wallaby (*Macropus eugenii*) [52], the Tasmanian devil (*Sarcophilus harrisii*) [43], the brown antechinus (*Antechinus stuartii*) [53], the koala (*Phascolarctos cinereus*) [54], the numbat (*Myrmecobius fasciatus*) [55], and the eastern quoll (*Dasyurus viverrinus*) [56], with genome assembly recovery of complete single-copy mammalian Benchmarking Universal Single-Copy Orthologs (BUSCOs) ranging from 73.1% to 92.4% [56]. The global transcriptomes generated for some of these species have BUSCO scores ranging from 76.4% to 84% [53, 55].

Recently, the fat-tailed dunnart (*Sminthopsis crassicaudata*, NCBI:txid9301) has emerged as a key laboratory marsupial model for understanding mammalian development and evolution [42, 57–61]. A nocturnal species belonging to the family Dasyuridae, the fat-tailed dunnart has adapted to a wide range of habitats and can be found across south and central mainland Australia [62] (Figure 1A and B). As one of the smallest carnivorous marsupials, adults weigh an average of 15 grams. Fat-tailed dunnarts exhibit some of the shortest known gestation times for mammals (13 days), with much of their development occurring postnatally. Fat-tailed dunnart neonates reside in their mother's pouch, thereby allowing continuous and non-invasive experimental access [63, 64]. The extremely altricial state of the dunnart young, along with very simple husbandry requirements, have facilitated the dunnart's role as a model species for comparative mammalian studies and conservation strategies.

However, the paucity of genomic resources for the fat-tailed dunnart has limited our understanding of this species at the gene level. As such, high-quality genome assembly and genome annotation have become increasingly important for investigations into the dunnart's unique biology. Recently, a draft fat-tailed dunnart genome assembly was released based on sequence data comprising ONT and PacBio long reads as well as Illumina HiSeq short reads [65]. While this scaffold-level assembly is a significant resource, an improved workflow was necessary in order to increase the genome's contiguity and completeness. Moreover, due to the absence of a *de novo* transcriptome, gene annotations had to be lifted over from the Tasmanian devil (*Sarcophilus harrisii*, GCF_902635505.1 - mSarHar1.11) to the dunnart scaffolds, thereby producing an incomplete representation of dunnart gene structure.

To address this knowledge gap, we present a comprehensive *de novo* transcriptome built from RNA-seq data from 24 samples, spanning 12 tissues. This global transcriptome has recovered 93.3% of complete mammalian BUSCOs, indicating its functional completeness. We also report the very first fat-tailed dunnart genome annotation. The genome annotation effort, made possible through the multi-tissue transcriptome assembly and *ab initio* predictions, yielded 21,622 protein-coding genes. Additionally, we provide an improved genome assembly that is 3.23 Gb in size with a scaffold N50 of 72.64 Mb. Annotated genomes and global transcriptomes are of paramount importance for attaching biological meaning to sequencing data. As such, this first-draft annotation and global transcriptome

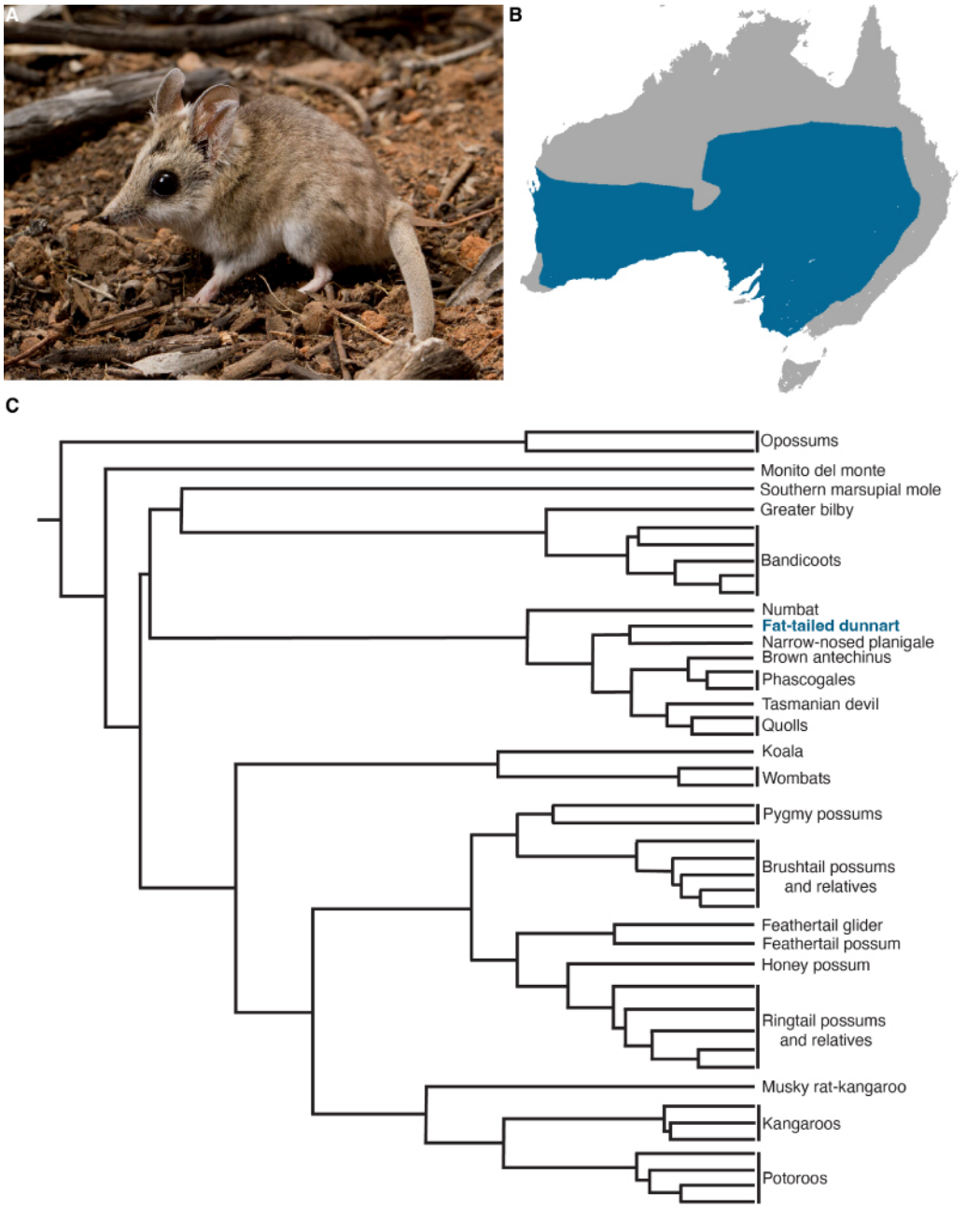

Figure 1. The fat-tailed dunnart (*Sminthopsis crassicaudata*).
(A) Adult fat-tailed dunnart captured in Ned's Corner, Victoria (Photo credit: David Paul, Museums Victoria). (B) The fat-tailed dunnart's range across Australia (CC BY) [66]. (C) Phylogeny of extant marsupial orders (based on [67] and [68]). The fat-tailed dunnart (blue font) is a member of the order Dasyuromorphia.

can serve as tools with which the genomic architecture of the fat-tailed dunnart, an emerging marsupial model species, can be better understood. Most importantly, these comprehensive resources contribute to the growing body of research on marsupial genomics and are therefore invaluable tools for future mammalian studies.

METHODS

Draft genome assembly

Fat-tailed dunnart ONT (~18 Gb, including the new ONT library 20190606 in PRJNA1078592) and Pacific Biosciences CLR (~171 Gb) long reads [65], along with Illumina short reads (~447.5 Gb in 2 × 150 bp format) [50], were combined to produce an improved draft genome assembly. Briefly, *de novo* contigs were first assembled from long reads ≥10 kilobases using Flye v2.9 (RRID:SCR_017016) [69] (parameters: –pacbio-raw, –genome-size 3g –iterations 2 –scaffold). Uncollapsed haplotypes were removed using purge_dups [70] with automatic coverage threshold detection. A second round of scaffolding was then performed using LongStitch v1.0.1 [71] (mode: ntLink-arks with an estimated genome size of 3 Gb). The resulting assembly was then polished in two rounds using Pilon v1.24 (RRID:SCR_014731) [72] (parameters: –vcf –diploid –chunksize 10000000 –fix snps,indels,gaps –minqual 15). To do this, Illumina short reads were first filtered and trimmed with Trimmomatic v0.38 (RRID:SCR_011848) [73] (parameters: SLIDINGWINDOW:5:30, MINLEN:75, AVGQUAL:30). Reads were then aligned against the assembly using BWA-MEM2 [74] (parameter: –M), and the resulting alignments were filtered with Samtools view v1.11 (RRID:SCR_002105) [75] (parameters: –h –b –q 30 –F 3340 –f 3). All of the data used to assemble contigs came from females, with data from two individuals being combined (one female for the PacBio data and one female for the ONT data). Short reads were obtained from an individual of unknown sex and were thus excluded from the contig assembly stage. Benchmark mammal ortholog recovery for the assembly was determined using BUSCO v5.2.2 (RRID:SCR_015008) [76], in genome mode, using the Mammalia_odb v10 database of orthologs (9226 BUSCOs). BUSCO (v5.2.2) genome completeness scores were also computed for the numbat, koala, Tasmanian devil, Brown antechinus, tammar wallaby, gray short-tailed opossum, and the eastern quoll.

Sample collection and sequencing

Adult and fetal fat-tailed dunnart tissues were collected for short-read Illumina sequencing from several individuals housed in a captive colony at the University of Melbourne. The tissues included late pregnancy allantois (n = 3), amnion (n = 3), distal yolk sac without vasculature (n = 2), proximal yolk sac with vasculature (n = 2), endometrium (n = 4), ovary (n = 3), oviduct (n = 2), combined uterus and oviduct (n = 1), testis (n = 1), female liver (n = 1), female eye (n = 1), and prostate gland (n = 1).

RNA samples were pooled in approximately equal proportions for Iso-Seq, namely, allantois, amnion, distal and proximal yolk sacs, endometrium, oviduct, ovary, testis, liver, eye, gastrula-stage conceptus, and late fetus. All RNA samples were extracted using Qiagen RNeasy Mini or Micro kits according to the manufacturer's instructions, with Illumina and Iso-Seq library construction and sequencing outsourced to Azenta Life Sciences (USA). For Illumina sequencing, this included rRNA depletion and strand-specific RNA library preparation, multiplexing, and sequencing on the NovaSeq platform, in a 2 × 150-bp (paired-end) configuration for 23 samples. Iso-Seq (poly-A selected and strand-specific) was performed using a PacBio Sequel II platform (1 sample, mean length of 5,400 bp). RNA Integrity Numbers (RIN) were generated using Bioanalyzer, and are available through Figshare [77].

De novo transcriptome assembly

The raw RNA-seq reads were quality-checked using FastQC v0.11.9 (RRID:SCR_014583) [78]. Quality trimming of the short-read data was carried out using Trimmomatic v0.38 [73] (parameters: SLIDINGWINDOW:4:28, MINLEN:25, AVGQUAL:28). Post-trimming, 464M paired reads remained.

To generate a global dunnart transcriptome, the trimmed, paired-end RNA-seq reads were used as input to Trinity v2.13.2 (RRID:SCR_013048) [79]. We Applied default *in silico* read normalization and set the minimum assembled contig length to report to 200. Circular consensus reads were incorporated for Iso-seq long-read correction (parameter: –long_reads). Contig assembly was executed using three different k-mer settings: 25, 29, and 32. We chose these values because 25 and 32 are the minimum and maximum permitted values for the Trinity contig assembly step. Assembly statistics were obtained using the Trinity script TrinityStats.pl [79]. A reference-free evaluation of assembly quality was conducted using RSEM-EVAL, a component package of Detonate v1.11 [80]. RSEM-EVAL provides a weighted quality score using a probabilistic model. Although these scores are always negative, when comparing two assemblies, a higher value represents a higher-quality assembly. The completeness of the full-length assemblies was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) [76]. The BUSCO gene sets are comprised of nearly universally distributed single-copy orthologous genes representing various phylogenetic levels. Here, BUSCO v5.2.2 assessment was carried out in transcriptome mode using the Mammalia_odb v10 database of orthologs.

To quantify the RNA-seq read representation of the assembly, all reads were mapped back to the global transcriptome assembly using Bowtie2 v2.4.5 (RRID:SCR_005476) [81], setting a maximum of 20 distinct alignments for each read (parameter: –k 20). Transcript abundance was quantified using RSEM v1.3.3 [82], with Bowtie2 read alignments. Prior to annotation, transcript redundancy in the global transcriptome was reduced using CD-HIT v4.8.1 [83] with a homology threshold of 1 (parameter: –c 1) to avoid filtering out true isoforms.

Transcriptome functional annotation

Functional annotation of the assembled transcripts was conducted using the Trinotate v3.2.2 [79] analysis protocol. First, Transdecoder v.5.5.0 [79] was used to identify all open reading frames (with a minimum length of 100 amino acids) and predict coding regions within transcripts. Sequence and domain homologies were captured by running BLAST+ v2.13.0 [84] (parameters: –max_target_seqs 1 –outfmt 6 –evalue 1e-5) against a combined protein database consisting of the UniProt/Swiss-Prot non-redundant protein sequences (RRID:SCR_002380) [85] from human (UP000005640), house mouse (UP000000589), Tasmanian devil (UP000007648), koala (UP000515140), tammar wallaby (txid9315), gray short-tailed opossum (UP000002280), and numbat (txid55782). Functional domains were identified by running a HMMer v3.3.2 [86] search against the PFAM v35.0 [87] database using the predicted protein sequences. Signal peptides and transmembrane domains were predicted using the SignalP v6.0 [88] and DeepTMHMM v1.0.24 [89] software tools, respectively.

Genome annotation

Annotation of the dunnart draft genome was conducted using a combination of *ab initio* gene prediction algorithms and homology-based methods (Figure 2). First, genome repeats

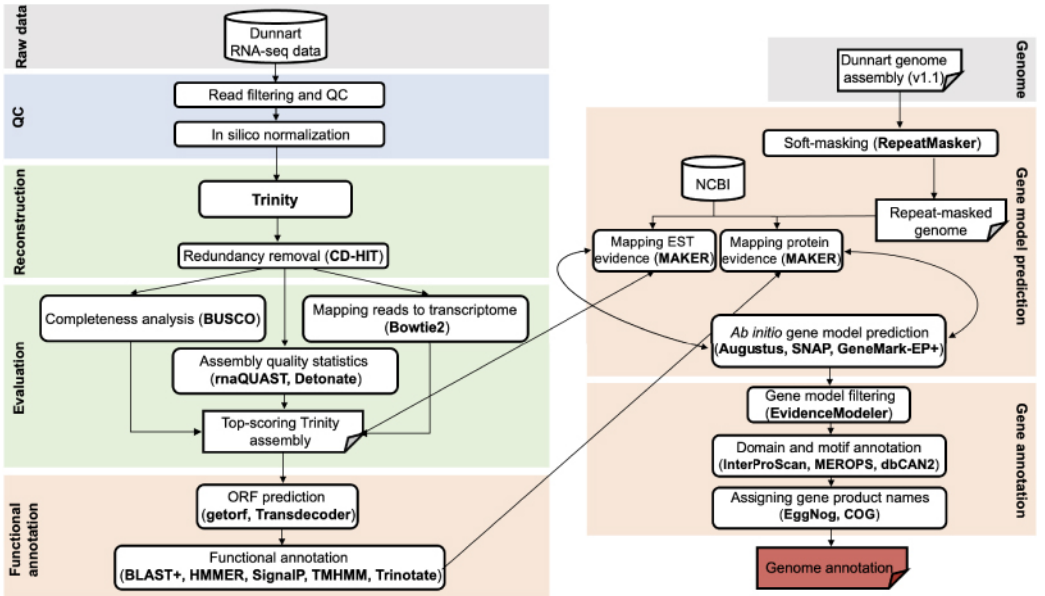

Figure 2. Schematic illustrating the *de novo* transcriptome generation and genome annotation workflow for the fat-tailed dunnart.

were masked using RepeatMasker v4.0.6 (RRID:SCR_012954) [90], with complex repeats being hard-masked while simple repeats were soft-masked. Preliminary gene models were constructed with MAKER2 (RRID:SCR_00530) [91] by aligning the assembled transcriptome and homologous protein sequences to the masked genome using minimap2 v2.26 [92] and DIAMOND v2.1.8 [93], respectively. Both cDNA (parameter: –model est2genome) and protein (parameter: –model protein2genome) alignments were polished with Exonerate v2.4.0 [94], producing high-quality alignments with precise intron/exon positions.

These preliminary gene models were then used to train the *ab initio* gene predictors SNAP (RRID:SCR_002127) [95], Augustus v3.4.0 (RRID:SCR_008417) [96], and GeneMark-EP+ v4.71 [97], all of which generated a statistical model representing the observed intron/exon structure in the genome. The gene model prediction process was iteratively run with MAKER2 (3 total rounds of prediction and re-training), thereby optimizing the performance of the *ab initio* gene predictors. For each round, prediction quality was evaluated using BUSCO scores. Consensus gene models were identified using EVidenceModeler v2.0.0 [98], with input weights set to 2 for high-quality *ab initio* predictions and to 1 for all other *ab initio* predictions and transcript/protein alignments. Gene models that lacked mRNA and protein homology support were excluded from the final annotation file. Lastly, gene names and putative protein functions were assigned using the aforementioned Trinotate output, as well as curated orthologous group and product names from InterProScan v5.60 (RRID:SCR_005829) [99], EggNOG v5.0 (RRID:SCR_002456) [100], MEROPS v12.4 [101], dbCAN3 v3.0.6 [102], and EuKaryotic Orthologous Groups (KOGs) [103].

RESULTS

To generate a genome-level annotation for the fat-tailed dunnart, we began by producing an improved draft genome assembly. We employed a hybrid approach, which integrated the

Table 1. Fat-tailed dunnart genome assembly statistics compared to the numbat, koala, Tasmanian devil, brown antechinus, tammar wallaby, gray short-tailed opossum, and eastern quoll reference genomes currently available on NCBI.

	Fat-tailed dunnart (this study)	Fat-tailed dunnart [65]	Numbat [55]	Koala [52]	Tasmanian devil [43]	Brown antechinus [53]	Tammar wallaby [52]	Gray short-tailed opossum [51]	Eastern quoll [56]
Genome size (Gb)	3.23	2.84	3.42	3.19	3.17	3.31	3.07	3.59	3.14
Number of scaffolds	1,848	719	112,299	-	105	30,796	314	13	76
Number of contigs	2,569	1,154	219,447	1,906	444	106,199	829	2,268	507
Scaffold N50 (Mb)	72.64	28.02	0.22	-	611.3	72.7	489.7	538.3	628.5
Scaffold L50	15	23	3,890	-	3	14	3	3	3
Contig N50 (Mb)	11.19	10.93	0.037	11.58	62.3	0.078	15.3	3.9	13.8
Contig L50	81	78	24,796	85	14	12,151	60	244	72
GC (%)	36.18	36.25	36.30	39.05	36.04	36.20	38.80	38.00	36.19
Complete Mammalian BUSCOs (v5.2.2, %)	94.2	89.9	73.2	92.4	90.3	90.4	81.8	92.0	92.2

ONT and PacBio long-read data with Illumina paired-end short reads [50]. This resulted in a 3.23 Gb genome that contains 1,848 scaffolds and has a scaffold N50 of 72.64 Mb. The GC content of this draft genome is 36.2% (Table 1). The recovery of complete, single-copy mammalian BUSCOs was 94.2%. Together, these metrics are indicative of a high-quality genome assembly, with marked improvements over the existing dunnart draft genome and notably higher completeness and contiguity compared to other marsupial reference genomes currently available on NCBI (Table 1).

A *de novo* reconstruction of the dunnart transcriptome was conducted using a set of 24 RNA-seq samples originating from the liver, testis, prostate, ovary, oviduct, uterus, eye, whole neonate, allantois, amnion, distal yolk sac, proximal yolk sac, and endometrium. To ensure that the most representative assembly was obtained, we sought to identify the optimal k-mer length for the Trinity contig assembly step, considering k values of 25, 29, and 32 (Table 2). Given that reference-free transcriptome assembly relies on grouping overlapping sequences of read fragments of a predetermined size (i.e., the k-mer), identifying the optimal fragment size might yield a more accurate assembly. To assess this fragment size effect, we computed multiple assembly quality metrics, including the BUSCO completeness score (transcriptome mode) and the Detonate RSEM-EVAL score for each Trinity run. The RSEM-EVAL score represents the sum of three main factors: likelihood estimates of the read representation within the assembly, the assembly prior, which assumes that each contig is generated independently, and the BIC (Bayesian Information Criterion) penalty [80]. When comparing two assemblies, a higher RSEM-EVAL score is indicative of a more complete transcriptome assembly. In our comparison, the Trinity run with a k-mer setting of 29 produced the top-scoring assembly; thus, all subsequent analysis was carried out using this assembly.

This transcriptome assembly was composed of 2,093,982 assembled transcripts (including splicing isoforms), with a GC content of 40.2% and a mean transcript length of 830 bp (Table 2). The transcript N50 was 1,489 bp, and considering only the top 90% most highly expressed transcripts (a more accurate proxy for transcriptome quality [104]) produced an E90N50 of 3,430 bp. Sample reads that were mapped back to the assembly had a very high overall alignment rate (98%), with a high percentage mapped as proper pairs (94%). In addition, the global transcriptome had a 93.3% recovery of complete mammalian BUSCOs (Mammalia_odb v10 [76]). These values are in line with, or higher than, those reported from all other available marsupial transcriptome datasets (Table 3). Specifically,

Table 2. Summary of the *de novo* transcriptome assembly statistics for the Trinity k-mer optimization.

	Trinity-k25	Trinity-k29	Trinity-k32
Total # of assembled Trinity transcripts	2,588,090	2,093,982	1,960,023
Mean transcript length (bp)	731	830	922
Transcript N50 (bp); E90N50 (bp)	1,193; 2,990	1,489; 3,430	1,260; 3,154
GC content (%)	41.7	40.2	40.9
Percentage of mapped RNA-seq PE reads (%)	95	98	97
Total BUSCO score (transcriptome mode)	C:92.1%, n:9226	C:93.3%, n:9226	C:93.0%, n:9226
Detonate RSEM-EVAL score	$-6{,}136.0 \times 10^7$	$-5{,}759.0 \times 10^7$	$-6{,}110.0 \times 10^7$

Table 3. Summary of global transcriptomes from marsupial species.

	Numbat [55]	Tasmanian devil [105]	Brown antechinus [53]
Total # of assembled transcripts	2,119,791	470,729	1,636,859
Mean transcript length (bp)	824	–	773
Transcript N50 (bp)	1,393	687	1,367
Percentage of mapped RNA-seq PE reads (%)	–	95	96
Total BUSCO score (transcriptome mode)	76.4% (v5.2.2)	–	84% (v4.0.6)

the global transcriptome assembly for the brown antechinus yielded 1,636,859 transcripts, with a mean length of 773 bp, a transcript N50 of 1,367 bp, a 96% alignment rate, and 84% complete BUSCOs [53]. The numbat global transcriptome contained 2,119,791 transcripts, a mean transcript length of 824 bp, a transcript N50 of 1,393 bp, and a BUSCO completeness score of 76.4% [55]. The Tasmanian devil transcriptome assembly consisted of 470,729 transcripts with an N50 of 687 bp and a 95% alignment rate of sample reads to the assembly [105].

Using a multi-pronged annotation approach of transcript and protein-level alignment, as well as *ab initio* gene prediction (Figure 2), we obtained 58,271 putative gene models for the fat-tailed dunnart draft genome (Table 4). Of these gene models, 21,622 were protein-coding (BLAST hits to UniProt/Swiss-Prot), which is in line with the reported gene numbers for the numbat (21,465) [55], the koala (27,669) [106], the Tasmanian devil (19,241) [43], the brown antechinus (25,111) [53], the tammar wallaby (15,290) [52], and the gray short-tailed opossum (21,384) [51] (Table 5). Furthermore, we predicted the putative function of the fat-tailed dunnart proteins using several curated protein databases (Table 4, Figure 3). We used InterProScan to identify conserved domains and assign Gene Ontology (GO) terms.

A total of 24,366 transcripts were assigned InterProScan terms, and 13,507 unique genes were assigned GO terms. The most common GO terms were intracellular anatomical structure (17,995 genes), organelle (17,140 genes), protein binding (17,071), cytoplasm (15,355 genes), and regulation of cellular processes (14,193 genes, Figure 3A). Notably, in another marsupial species, the woylie, cellular processes were also the most common GO term under the Biological Processes (BP) category [107]. Our GO annotations totaled 289,985, with a mean annotation level of 7.15 and a standard deviation of 2.7 (Figure 3B). Running an HMMer search against the PFAM database yielded 16,308 domains, while dbCAN3 and MEROPS analyses resulted in 212 and 1,053 predictions, respectively. Altogether, these results highlight valuable avenues through which we can deepen our understanding of marsupial biology at the gene level.

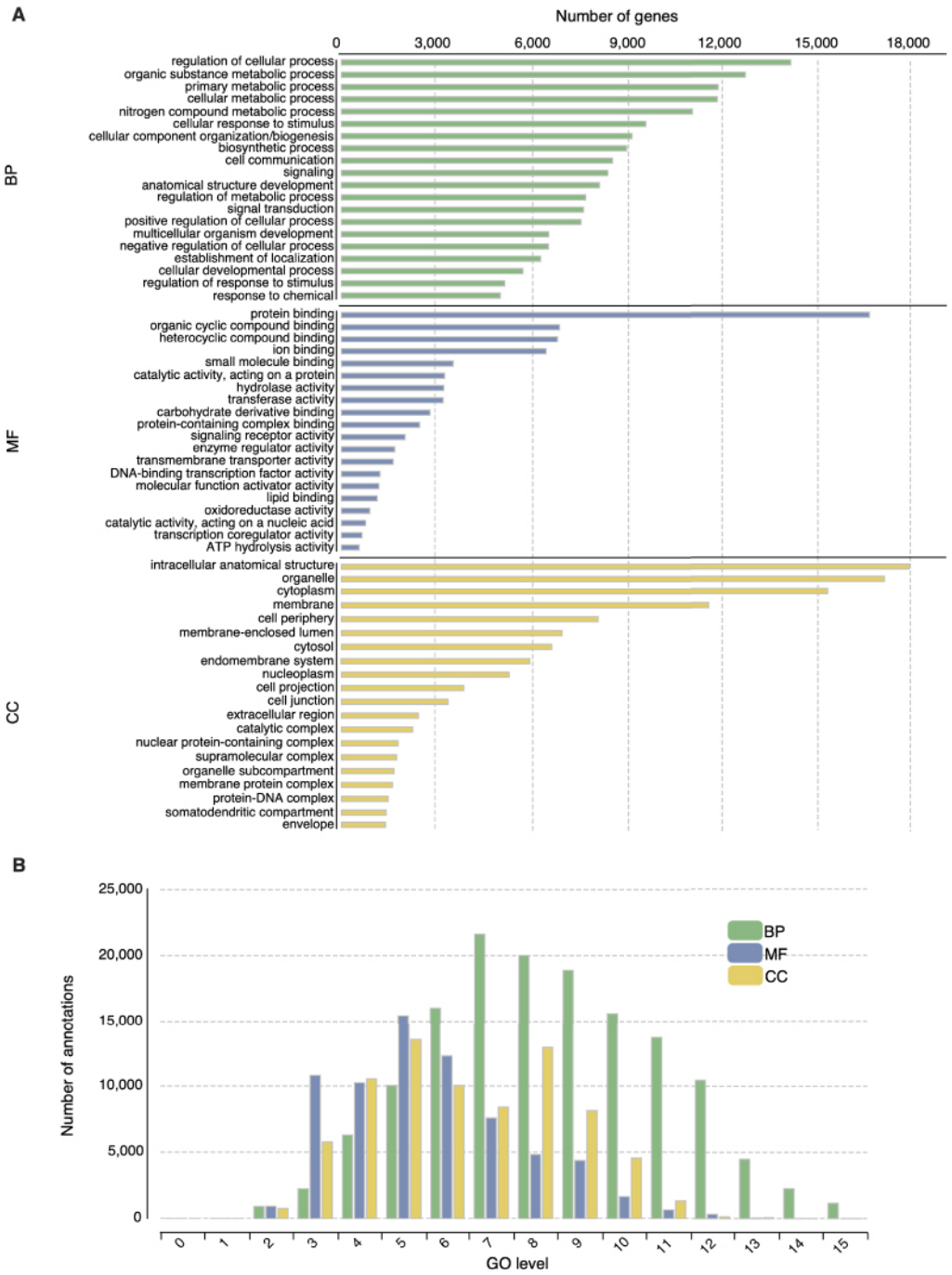

Figure 3. Gene ontology (GO) analysis of the fat-tailed dunnart putative genes.
(A) GO distribution by category (at level 3) for the fat-tailed dunnart gene set. The ontology categories are BP (Biological Process), MF (Molecular Function), and CC (Cellular Component). The top 20 terms are listed for each category. (B) Distribution of sequence annotations for each GO level.

Table 4. Fat-tailed dunnart gene and feature statistics.

	Fat-tailed dunnart
General	
Protein-coding genes	21,622
Predicted gene models	58,271
Transcript level	
mRNA	50,091
tRNA	8,180
Multiple exon transcripts	44,109
Single exon transcripts	5,982
Total exons	246,391
Average exon length	146.1
Functional level	
InterProScan terms	24,366
EggNOG terms	29,372
PFAM domains	16,308
dbCAN3 (CAZymes)	212
MEROPS (proteases)	1,053
GO	13,507

Table 5. Fat-tailed dunnart gene counts compared to the numbat, koala, Tasmanian devil, brown antechinus, tammar wallaby, gray short-tailed opossum, and eastern quoll.

	Number of putative genes	Number of protein-coding genes
Fat-tailed dunnart (this study)	58,271	21,622
Numbat [55]	77,806	21,465
Koala [54]	52,384	27,669
Tasmanian devil [43]	40,469	19,241
Brown antechinus [53]	55,827	25,111
Tammar wallaby [52]	122,304	15,290
Gray short-tailed opossum [51]	43,478	21,384
Eastern quoll [56]	29,622	14,293

CONCLUSION

The increased availability of genomic resources for marsupial species is critical for fostering a deeper understanding of the evolutionary history of both eutherians and marsupials. In this study, we report an enhanced fat-tailed dunnart genome assembly measuring 3.23 Gb in length. The assembly is organized into 1,848 scaffolds, with a scaffold N50 value of 72.64 Mb. We generated a global *de novo* transcriptome assembly of the fat-tailed dunnart using RNA-seq short-read and long-read data, which were sampled from a diverse range of dunnart tissues. The transcriptome reconstruction consisted of 2,093,982 assembled transcripts, with a mean transcript length of 830 bp. The transcriptome BUSCO completeness score of 93.3% is the highest amongst all other published marsupial transcriptome BUSCOs (i.e., numbat and brown antechinus). The high overall alignment rate of reads from each of the tissues to the transcriptome (98%) further underscores that the *de novo* transcriptome is a highly accurate representation of the input reads. The dunnart draft genome annotation revealed 21,622 protein-coding genes, in line with previously reported marsupial gene counts. Overall, these resources provide novel insights into the unique genomic architecture of the fat-tailed dunnart and will therefore serve as valuable tools for future comparative mammalian studies.

AVAILABILITY OF SOURCE CODE AND REQUIREMENTS

- Project name: Dunnart Genome Annotation
- Project home page: https://gitlab.svi.edu.au/igr-lab/dunnart_genome_annotation
- Operating system(s): Platform independent

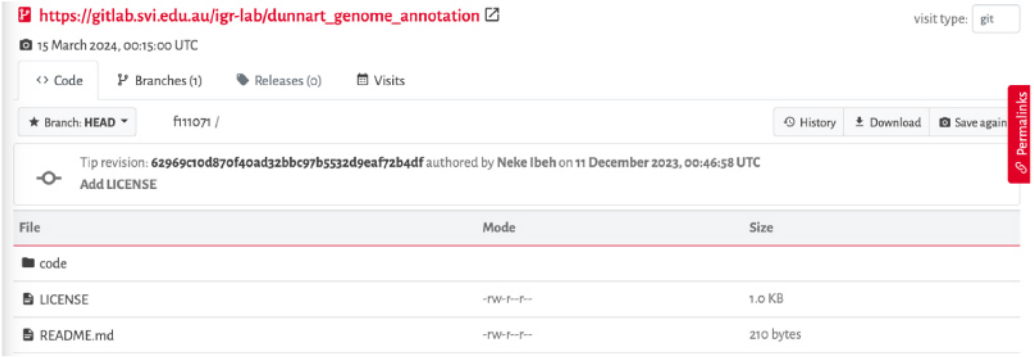

Figure 4. Software Heritage archive of the code [109]. https://archive.softwareheritage.org/browse/embed/swh:1:dir:f111071e06210e948b7a9e64154e62c4004ef5d2;origin=https://gitlab.svi.edu.au/igr-lab/dunnart_genome_annotation;visit=swh:1:snp:3fd382b053a5ca917cb19f1c27cd11c55880b26b;anchor=swh:1:rev:62969c10d870f40ad32bbc97b5532d9eaf72b4df/

- Programming language: Shell, Python, Perl
- License: MIT.

DATA AVAILABILITY

The fat-tailed dunnart transcriptome, draft genome, and genome annotation are available through Figshare [108]. The scripts for reproducing the genome annotation workflow have been made available in gitlab and archived in Software Heritage (Figure 4) [109]. All raw sequencing reads have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive under the accession numbers PRJNA1078592 (genomic reads) and PRJNA1028148 (RNA-Seq).

LIST OF ABBREVIATIONS

BLAST: Basic Local Alignment Search Tool; bp: base pair; BUSCO: Benchmarking Universal Single-Copy Orthologs; CDS: coding sequences; Gb: Gigabase; Kb: Kilobase; Mb: Megabase; NCBI: National Center for Biotechnology Information; ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences; PE: paired-end; RNA-seq: RNA sequencing.

DECLARATIONS

Ethics statement

All sample collection was approved by the University of Melbourne Animal Ethics Committee (Project ID 10206).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SF, AJP, CYF, and IGR conceived the project. SF collected and prepared the samples. NI assembled and annotated the global transcriptome. CYF assembled the draft genome. NI annotated the draft genome. NI drafted the manuscript with input from all authors. All authors read and approved the final version of the manuscript.

Funding

This work was supported by the Australian Research Council Discovery Project DP210102645 to AJP. IGR was partially supported by the European Union through the Horizon 2020 Research and Innovation Program under Grant No. 810645 and the European Union through the European Regional Development Fund Project No. MOBEC008.

Acknowledgements

NI was supported by the University of Melbourne Research Training Program Scholarship, the Rowden White Scholarship, the Dame Margaret Blackwood Soroptimist Scholarship, and the St. Vincent's Institute Top-up Scholarship. St. Vincent's Institute acknowledges the infrastructure support it receives from the National Health and Medical Research Council Independent Research Institutes Infrastructure Support Program and from the Victorian Government through its Operational Infrastructure Support Program.

REFERENCES

1 **Jackson SM**, **Jackson S**, **Groves C**. Taxonomy of Australian Mammals. Csiro Publishing, 2015. ISBN: 9781486300136.

2 **Wilson DE**, **Reeder DM**. Mammal Species of the World: A Taxonomic and Geographic Reference. JHU Press, 2005. ISBN-10: 0801882214.

3 **Archer M**, **Beck R**, **Gott M et al.** Australia's first fossil marsupial mole (Notoryctemorphia) resolves controversies about their evolution and palaeoenvironmental origins. ***Proc. R. Soc. B: Biol. Sci.***, 2010; **278**(1711): 1498–1506.

4 **Diogo R**, **Bello-Hellegouarch G**, **Kohlsdorf T et al.** Comparative myology and evolution of marsupials and other vertebrates, with notes on complexity, Bauplan, and "scala naturae". ***Anat. Rec.***, 2016; **299**(9): 1224–1255.

5 **Stein BR**. Comparative limb myology of two opossums, Didelphis and Chironectes. ***J. Morphol.***, 1981; **169**(1): 113–140.

6 **Schmitz J**, **Ohme M**, **Suryobroto B et al.** The colugo (*Cynocephalus variegatus*, Dermoptera): the primates' gliding sister? ***Mol. Biol. Evol.***, 2002; **19**(12): 2308–2312.

7 **Casanovas-Vilar I**, **Garcia-Porta J**, **Fortuny J et al.** Oldest skeleton of a fossil flying squirrel casts new light on the phylogeny of the group. ***eLife***, 2018; 7: e39270. doi:10.7554/eLife.39270.

8 **Henneberg M**, **Lambert K**, **Leigh C**. Fingerprint homoplasy: koalas and humans. ***Nat. Sci.***, 1997; **1**: 4, https://hdl.handle.net/2440/5433.

9 **McGhee GR**. Convergent Evolution: Limited Forms Most Beautiful. MIT Press, 2011. doi:10.7551/mitpress/9780262016421.001.0001.

10 **Freeman C**. 4. A Tasmanian Wolf. In: Paper Tiger. The Netherlands: Brill, 2010; pp. 117–149, doi:10.1163/ej.9789004181656.i-288.39.

11 **Feigin CY**, **Newton AH**, **Pask AJ**. Widespread cis-regulatory convergence between the extinct Tasmanian tiger and gray wolf. ***Genome Res.***, 2019; **29**(10): 1648–1658.

12 **Geiser F**, **Körtner G**, **Schmidt I**. Leptin increases energy expenditure of a marsupial by inhibition of daily torpor. ***Am. J. Physiol.***, 1998; **275**(5): R1627–R1632.

13 **Hing S**, **Narayan E**, **Thompson RCA et al.** A review of factors influencing the stress response in Australian marsupials. ***Conserv. Physiol.***, 2014; **2**(1): cou027.

14 **Karlen SJ**, **Krubitzer L**. The functional and anatomical organization of marsupial neocortex: evidence for parallel evolution across mammals. ***Prog. Neurobiol.***, 2007; **82**(3): 122–141.

15 **Krause WJ**, **Cutts JH**, **Leeson CR**. Postnatal development of the epidermis in a marsupial, *Didelphis virginiana*. ***J. Anat.***, 1978; **125**(Pt 1): 85–99.

16 **Cutts JH**, **Leeson CR**, **Krause WJ**. The postnatal development of the liver in a marsupial, *Didelphis virginiana*. 1. Light microscopy. ***J. Anat.***, 1973; **115**(Pt 3): 327–346.

17 **Fadem BH**, **Harder JD**. Evidence for high levels of androgen in peripheral plasma during postnatal development in a marsupial: the gray short-tailed opossum (*Monodeiphis Domestica*). ***Biol. Reprod.***, 1992; **46**(1): 105–108.

18 **Runciman SI**, **Baudinette RV**, **Gannon BJ**. Postnatal development of the lung parenchyma in a marsupial: the tammar wallaby. ***Anat. Rec.***, 1996; **244**(2): 193–206.

19 **Goldingay RL**. The behavioural ecology of the gliding marsupial, *Petaurus australis*. PhD thesis, University of Wollongong, 1989; http://ro.uow.edu.au/theses/1077.

20 **Menário Costa W**, **King WJ**, **Bonnet T et al.** Early-life behavior, survival, and maternal personality in a wild marsupial. ***Behav. Ecol.***, 2023; **34**(6): arad070.

21 **Russell EM**. Social behaviour and social organization of marsupials. ***Mamm. Rev.***, 1984; **14**(3): 101–154.

22 **Renfree MB**. Monotreme and marsupial reproduction. ***Reprod. Fertil. Dev.***, 1995; **7**(5): 1003–1020.

23 **Sharman GB**. Reproductive physiology of marsupials. ***Science***, 1970; **167**(3922): 1221–1228.

24 **Harder JD**, **Jackson LM**. Chemical communication and reproduction in the gray short-tailed opossum (*Monodelphis domestica*). ***Vitam. Horm.***, 2010; **83**: 373–399.

25 **Bergallo HG**, **Cerqueira R**. Reproduction and growth of the opossum *Monodelphis domestica* (Mammalia: Didelphidae) in northeastern Brazil. ***J. Zool.***, 1994; **232**(4): 551–563.

26 **Chen Y**, **Yu H**, **Pask AJ et al.** Hormone-responsive genes in the SHH and WNT/*β*-catenin signaling pathways influence urethral closure and phallus growth. ***Biol. Reprod.***, 2018; **99**(4): 806–816.

27 **Coveney D**, **Shaw G**, **Hutson JM et al.** Effect of an anti-androgen on testicular descent and inguinal closure in a marsupial, the tammar wallaby (*Macropus eugenii*). ***Reproduction***, 2002; **124**(6): 865–874.

28 **Moore HDM**, **Thurstan SM**. Sexual differentiation in the grey short-tailed opossum, *Monodelphis domestica*, and the effect of oestradiol benzoate on development in the male. ***J. Zool.***, 1990; **221**(4): 639–658.

29 **Renfree MB**, **Pask AJ**, **Shaw G**. Sex down under: the differentiation of sexual dimorphisms during marsupial development. ***Reprod. Fertil. Dev.***, 2001; **13**(7–8): 679–690.

30 **Pask AJ**, **Harry JL**, **Renfree MB et al.** Absence of SOX3 in the developing marsupial gonad is not consistent with a conserved role in mammalian sex determination. ***Genesis***, 2000; **27**(4): 145–152.

31 **Pask A**, **Renfree MB**, **Marshall Graves JA**. The human sex-reversing ATRX gene has a homologue on the marsupial Y chromosome, ATRY: implications for the evolution of mammalian sex determination. ***Proc. Natl. Acad. Sci. USA***, 2000; **97**(24): 13198–13202.

32 **Scherer G**, **Schmid M**. Genes and mechanisms in vertebrate sex determination. Introduction. ***Exp. Suppl.***, 2001; **91**(3): XI–XII.

33 **Hornecker JL**, **Samollow PB**, **Robinson ES et al.** Meiotic sex chromosome inactivation in the marsupial *Monodelphis domestica*. ***Genesis***, 2007; **45**(11): 696–708.

34 **Foster JW**, **Brennan FE**, **Hampikian GK et al.** Evolution of sex determination and the Y chromosome: SRY-related sequences in marsupials. ***Nature***, 1992; **359**(6395): 531–533.

35 **Ishihara T**, **Hickford D**, **Shaw G et al.** DNA methylation dynamics in the germline of the marsupial tammar wallaby, *Macropus eugenii*. ***DNA Res.***, 2019; **26**(1): 85–94.

36 **Whitworth DJ**, **Pask AJ**. The X factor: X chromosome dosage compensation in the evolutionarily divergent monotremes and marsupials. ***Semin. Cell Dev. Biol.***, 2016; **56**: 117–121.

37 **Wang X**, **Douglas KC**, **Vandeberg JL et al.** Chromosome-wide profiling of X-chromosome inactivation and epigenetic states in fetal brain and placenta of the opossum, *Monodelphis domestica*. ***Genome Res.***, 2014; **24**(1): 70–83.

38 **Das R**, **Anderson N**, **Koran MI et al.** Convergent and divergent evolution of genomic imprinting in the marsupial *Monodelphis domestica*. ***BMC Genom.***, 2012; **13**: 394.

39 **McAllan BM**. Dasyurid marsupials as models for the physiology of ageing in humans. ***Aust. J. Zool.***, 2006; **54**(3): 159–172.

40 **Bartkowska K**, **Tepper B**, **Turlejski K et al.** Postnatal and adult neurogenesis in mammals, including marsupials. ***Cells***, 2022; **11**(17): 2735.

41 **Szdzuy K**, **Zeller U**, **Renfree M et al.** Postnatal lung and metabolic development in two marsupial and four eutherian species. ***J. Anat.***, 2008; **212**(2): 164–179.

42 **Cook LE**, **Newton AH**, **Hipsley CA et al.** Postnatal development in a marsupial model, the fat-tailed dunnart (*Sminthopsis crassicaudata*; Dasyuromorphia: Dasyuridae). ***Commun. Biol.***, 2021; **4**(1): 1028.

43 **Stammnitz MR**, **Gori K**, **Kwon YM et al.** The evolution of two transmissible cancers in Tasmanian devils. ***Science***, 2023; **380**(6642): 283–293.

44 **De Leo AA**. Genome evolution in Australian Marsupials. PhD thesis, University of Melbourne, Department of Zoology, Faculty of Science, 2005; http://hdl.handle.net/11343/341740.

45 **Graves JA**. Mammalian genome evolution: new clues from comparisons of eutherians, marsupials and monotremes. ***Comp. Biochem. Physiol. A***, 1991; **99**(1-2): 5–11.

46 **Deakin JE**, **O'Neill RJ**. Evolution of marsupial genomes. ***Annu. Rev. Anim. Biosci.***, 2020; **8**: 25–45.

47 **Ishihara T**, **Hickford D**, **Fenelon JC et al.** Evolution of the short form of DNMT3A, DNMT3A2, occurred in the common ancestor of mammals. ***Genome Biol. Evol.***, 2022; **14**(7): evac094. doi:10.1093/gbe/evac094.

48 **Deakin JE**. Marsupial genome sequences: providing insight into evolution and disease. ***Scientifica***, 2012; **2012**: 543176.

49 **Janke A**, **Feldmaier-Fuchs G**, **Thomas WK et al.** The marsupial mitochondrial genome and the evolution of placental mammals. ***Genetics***, 1994; **137**(1): 243–256.

50 **Feigin C**, **Frankenberg S**, **Pask A**. A chromosome-scale hybrid genome assembly of the extinct Tasmanian Tiger (*Thylacinus cynocephalus*). ***Genome Biol. Evol.***, 2022; **14**(4): evac048. doi:10.1093/gbe/evac048.

51 **Mikkelsen TS**, **Wakefield MJ**, **Aken B et al.** Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. ***Nature***, 2007; **447**(7141): 167–177.

52 **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**(8): R81.

53 **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**: gigabyte7. doi:10.46471/gigabyte.7.

54 **Johnson RN**, **O'Meally D**, **Chen Z et al.** Adaptation and conservation insights from the koala genome. ***Nat. Genet.***, 2018; **50**(8): 1102–1111.

55 **Peel E**, **Silver L**, **Brandies P et al.** Genome assembly of the numbat (*Myrmecobius fasciatus*), the only termitivorous marsupial. ***GigaByte***, 2022; **2022**: gigabyte47. doi:10.46471/gigabyte.47.

56 **Hartley GA**, **Frankenberg SR**, **Robinson NM et al.** Genome of the endangered eastern quoll (*Dasyurus viverrinus*) reveals signatures of historical decline and pelage color evolution. bioRxiv. 2023; https://doi.org/10.1101/2023.09.06.556354.

57 **Polymeropoulos ET**, **Jastroch M**, **Frappell PB**. Absence of adaptive nonshivering thermogenesis in a marsupial, the fat-tailed dunnart (*Sminthopsis crassicaudata*). ***J. Comp. Physiol. B***, 2012; **182**(3): 393–401.

58 **Suárez R**, **Paolino A**, **Kozulin P et al.** Development of body, head and brain features in the Australian fat-tailed dunnart (*Sminthopsis crassicaudata*; Marsupialia: Dasyuridae); A postnatal model of forebrain formation. ***PLoS One***, 2017; **12**(9): e0184450.

59 **Garrett A**, **Lannigan V**, **Yates NJ et al.** Physiological and anatomical investigation of the auditory brainstem in the Fat-tailed dunnart (*Sminthopsis crassicaudata*). ***PeerJ***, 2019; **7**: e7773.

60 **Noy EB**, **Scott MK**, **Grommen SVH et al.** Molecular cloning and tissue distribution of Crh and Pomc mRNA in the fat-tailed dunnart (*Sminthopsis crassicaudata*), an Australian marsupial. ***Gene***, 2017; **627**: 26–31.

61 **Suárez R**, **Paolino A**, **Kozulin P et al.** Development of body, head and brain features in the Australian fat-tailed dunnart (*Sminthopsis crassicaudata*; Marsupialia: Dasyuridae); A postnatal model of forebrain formation. ***PLoS One***, 2017; **12**(9): e0184450.

62 **Collins LR**. Monotremes and Marsupials. Smithsonian Institution, 1973. ASIN: B0006C8ZKY.

63 **Tyndale-Biscoe CH**, **Janssens PA**. The Developing Marsupial: Models for Biomedical Research. Springer Science & Business Media, 2012. doi:10.1007/978-3-642-88402-3.

64 **Godfrey GK**, **Crowcroft P**. Breeding the Fat-tailed marsupial mouse in captivity. ***Int. Zoo Yearbook***, 1971; **11**(1): 33–38. doi:10.1111/j.1748-1090.1971.tb01839.x.

65 **Cook LE**, **Feigin CY**, **Pask AJ et al.** Cis-regulatory landscapes of the fat-tailed dunnart and mouse provide insights into the drivers of craniofacial heterochrony. bioRxiv. 2023; https://doi.org/10.1101/2023.02.13.528361.

66 **Chermundy**. Fat-tailed Dunnart area. Wikimedia Commons, IUCN Red List of Threatened Species, CC BY-SA 3.0. 2010; https://en.m.wikipedia.org/wiki/File:Fat-tailed_Dunnart_area.png.

67 **Duchêne DA**, **Bragg JG**, **Duchêne S et al.** Analysis of phylogenomic tree space resolves relationships among marsupial families. ***Syst. Biol.***, 2018; **67**(3): 400–412.

68 **Doronina L**, **Feigin CY**, **Schmitz J**. Reunion of Australasian possums by shared SINE insertions. ***Syst. Biol.***, 2022; **71**(5): 1045–1053.

69 **Kolmogorov M**, **Yuan J**, **Lin Y et al.** Assembly of long, error-prone reads using repeat graphs. ***Nat. Biotechnol.***, 2019; **37**(5): 540–546. doi:10.1038/s41587-019-0072-8.

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

71 **Coombe L**, **Li JX**, **Lo T et al.** LongStitch: high-quality genome assembly correction and scaffolding using long reads. ***BMC Bioinform.***, 2021; **22**(1): 534.

72 **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.

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

74 **Vasimuddin M**, **Misra S**, **Li H et al.** Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2019; pp. 314–324.

75 **Li H**, **Handsaker B**, **Wysoker A et al.** The sequence alignment/map format and SAMtools. ***Bioinformatics***, 2009; **25**(16): 2078–2079.

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

77 **Gallego Romero I**, **Ibeh N**, **Feigin C et al.** Supplementary table 1 - RIN for RNA seq samples. The University of Melbourne. Figshare. [Dataset]. 2024; https://doi.org/10.26188/25377487.v1.

78 **Andrews S**. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. 2010; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

79 **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. Protoc.***, 2013; **8**(8): 1494–1512.

80 **Li B**, **Fillmore N**, **Bai Y et al.** Evaluation of de novo transcriptome assemblies from RNA-Seq data. ***Genome Biol.***, 2014; **15**(12): 553.

81 **Langmead B**, **Salzberg SL**. Fast gapped-read alignment with Bowtie 2. ***Nat. Methods***, 2012; **9**(4): 357–359.

82 **Li B**, **Dewey CN**. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. ***BMC Bioinform.***, 2011; **12**: 323.

83 **Huang Y**, **Niu B**, **Gao Y et al.** CD-HIT Suite: a web server for clustering and comparing biological sequences. ***Bioinformatics***, 2010; **26**(5): 680–682.

84 **Camacho C**, **Coulouris G**, **Avagyan V et al.** BLAST+: architecture and applications. ***BMC Bioinform.***, 2009; **10**: 421.

85 **Boutet E**, **Lieberherr D**, **Tognolli M et al.** UniProtKB/Swiss-Prot. ***Methods Mol. Biol.***, 2007; **406**: 89–112.

86 **Finn RD**, **Clements J**, **Eddy SR**. HMMER web server: interactive sequence similarity searching. ***Nucleic Acids Res.***, 2011; **39**(Web Server issue): W29–W37.

87 **Mistry J**, **Chuguransky S**, **Williams L et al.** Pfam: The protein families database in 2021. ***Nucleic Acids Res.***, 2021; **49**(D1): D412–D419.

88 **Teufel F**, **Almagro Armenteros JJ**, **Johansen AR et al.** SignalP 6.0 predicts all five types of signal peptides using protein language models. ***Nat. Biotechnol.***, 2022; **40**(7): 1023–1025.

89 **Hallgren J**, **Tsirigos KD**, **Pedersen MD et al.** DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv. 2022; https://doi.org/10.1101/2022.04.08.487609.

90 **Nishimura D**. RepeatMasker. ***Biotech Softw. Inter. Report***, 2000; **1**(1–2): 36–39. doi:10.1089/152791600319259.

91 **Holt C**, **Yandell M**. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. ***BMC Bioinform.***, 2011; **12**: 491.

92 **Li H**. Minimap2: pairwise alignment for nucleotide sequences. ***Bioinformatics***, 2018; **34**(18): 3094–3100.

93 **Buchfink B**, **Xie C**, **Huson DH**. Fast and sensitive protein alignment using DIAMOND. ***Nat. Methods***, 2015; **12**(1): 59–60.

94 **Slater GSC**, **Birney E**. Automated generation of heuristics for biological sequence comparison. ***BMC Bioinform.***, 2005; **6**: 31.

95 **Korf I**. Gene finding in novel genomes. ***BMC Bioinform.***, 2004; **5**: 59.

96 **Stanke M**, **Diekhans M**, **Baertsch R et al.** Using native and syntenically mapped cDNA alignments to improve de novo gene finding. ***Bioinformatics***, 2008; **24**(5): 637–644.

97 **Brůna T**, **Lomsadze A**, **Borodovsky M**. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. ***NAR Genom. Bioinform.***, 2020; **2**(2): lqaa026.

98 **Haas BJ**, **Salzberg SL**, **Zhu W et al.** Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. ***Genome Biol.***, 2008; **9**(1): R7.

99 **Jones P**, **Binns D**, **Chang HY et al.** InterProScan 5: genome-scale protein function classification. ***Bioinformatics***, 2014; **30**(9): 1236–1240.

100 **Huerta-Cepas J**, **Szklarczyk D**, **Heller D et al.** eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. ***Nucleic Acids Res.***, 2019; **47**(D1): D309–D314.

101 **Rawlings ND**, **Barrett AJ**, **Bateman A**. MEROPS: the peptidase database. ***Nucleic Acids Res.***, 2010; **38**(Database issue): D227–D233.

102 **Zheng J**, **Ge Q**, **Yan Y et al.** dbCAN3: automated carbohydrate-active enzyme and substrate annotation. ***Nucleic Acids Res.***, 2023; **51**(W1): W115–W121.

103 **Tatusov RL**, **Fedorova ND**, **Jackson JD et al.** The COG database: an updated version includes eukaryotes. ***BMC Bioinform.***, 2003; **4**: 41.

104 **Haas B**. Transcriptome contig Nx and ExN50 stats. 2016; https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcriptome-Contig-Nx-and-ExN50-stats.

105 **Murchison EP**, **Schulz-Trieglaff OB**, **Ning Z et al.** Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. ***Cell***, 2012; **148**(4): 780–791.

106 **Blanchard AM**, **Emes RD**, **Greenwood AD et al.** Genome reference assembly for bottlenecked Southern Australian Koalas. ***Genome Biol. Evol.***, 2023; **15**(1): evac176.

107 **Peel E**, **Silver L**, **Brandies P et al.** A reference genome for the critically endangered woylie, *Bettongia penicillata* ogilbyi. ***GigaByte***, 2021; **2021**: gigabyte35. doi:10.46471/gigabyte.35.

108 **Gallego Romero I**. Fat tailed dunnart transcriptome reference files. The University of Melbourne. Figshare. [Dataset]. 2024; https://melbourne.figshare.com/projects/Fat_tailed_dunnart_transcriptome_reference_files/183307.

109 **Gallego Romero I**, **Ibeh N**, **Feigin C et al.** Dunnart_Genome_Annotation (Version 1). [Computer software]. Software Heritage, 2024; https://archive.softwareheritage.org/swh:1:snp:3fd382b053a5ca917cb19f1c27cd11c55880b26b;origin=https://gitlab.svi.edu.au/igr-lab/dunnart_genome_annotation.