(GIGA)byte

DATA RELEASE

The first *Antechinus* reference genome provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies

Parice A. Brandies[1], Simon Tang[1], Robert S. P. Johnson[2], Carolyn J. Hogg[1,†] and Katherine Belov[1,*,†]

1 School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia
2 Zoologica: Veterinary and Zoological Consulting, Millthorpe, New South Wales, Australia

ABSTRACT

Antechinus are a genus of mouse-like marsupials that exhibit a rare reproductive strategy known as semelparity and also naturally develop age-related neuropathologies similar to those in humans. We provide the first annotated antechinus reference genome for the brown antechinus (*Antechinus stuartii*). The reference genome is 3.3 Gb in size with a scaffold N50 of 73Mb and 93.3% complete mammalian BUSCOs. Using bioinformatic methods we assign scaffolds to chromosomes and identify 0.78 Mb of Y-chromosome scaffolds. Comparative genomics revealed interesting expansions in the NMRK2 gene and the protocadherin gamma family, which have previously been associated with aging and age-related dementias respectively. Transcriptome data displayed expression of common Alzheimer's related genes in the antechinus brain and highlight the potential of utilising the antechinus as a future disease model. The valuable genomic resources provided herein will enable future research to explore the genetic basis of semelparity and age-related processes in the antechinus.

Submitted: 21 September 2020
Accepted: 04 November 2020
Published: 05 November 2020

Subjects Genetics and Genomics, Animal Genetics, Evolutionary Biology

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

† Contributed equally.

Published by GigaScience Press.

Preprint submitted at https://doi.org/10.1101/2020.09.21.305722

Gigabyte, 2020, 1–22

CONTEXT

Antechinus are a genus of small, carnivorous, dasyurid marsupials that are distributed throughout Australia and New Guinea, and exhibit a rare reproductive strategy known as semelparity. Semelparous species reproduce only once in a lifetime [1]. Although this reproductive strategy is common among bacteria, plant and invertebrate species [2], it is rarely seen in mammalian species and is restricted to didelphid and dasyurid marsupials [3, 4]. During the annual breeding season, male antechinus undergo an extreme shift in resource allocation from survival to reproduction, resulting in a complete die-off of all males in the weeks following mating [1, 5–7]. Increased levels of plasma corticosteroid assist antechinus males in utilising their energy reserves to maximise reproductive potential during the breeding season [4]. However, elevation of these corticosteroids results in total immune system collapse leading to gastrointestinal haemorrhage, parasite/pathogen invasion and death [6, 8]. It is currently unknown how semelparity is controlled at the genetic level in the antechinus.

The antechinus has also been proposed as a model species for the physiology of dementias associated with aging such as Alzheimer's disease (AD) [3, 9, 10]. Primarily characterised by the formation of amyloid-β plaques and neurofibrillary tangles in the brain, AD is a progressive neurodegenerative disease that is predicted to affect more than 100 million people by 2050 [11]. Traditionally, transgenic mouse models have been utilised to study AD [12–14]; however, mice do not naturally develop β-amyloid plaques and neurofibrillary tangles [15, 16]. Both of these have been found to develop naturally in mature male and female antechinus, particularly after the breeding season [9, 10]. Antechinus also possess a number of characteristics that could make them an ideal model organism including: a small body size, short lifespan, production of large numbers of offspring and the ability to be easily maintained in captivity [6, 17, 18]. Creating a reference genome for the antechinus and understanding whether there is expression of key AD-related genes in the antechinus' brain is a key first step in determining their suitability as a future disease model for AD in humans.

Here we present an annotated reference genome for the brown antechinus (*Antechinus stuartii*; NCBI:txid9283). We use a bioinformatic approach [19] to provide a more complete characterisation of the Y chromosome which is currently poorly annotated in marsupials, due to its heterochromatic, highly repetitive nature and small size [20]. We also call and annotate phased genome-wide SNVs (single nucleotide variants) and structural variants, and use comparative genomics to identify rapidly evolving gene families. Finally, we characterise variation in a variety of genes that have previously been associated with AD and evaluate the expression of these genes in the antechinus transcriptome.

The annotated genome and other genomic resources provided herein provide a powerful foundation for studying semelparity and neurodegeneration as well as showcasing the potential hidden within the genomes of Australia's unique biodiversity.

METHODS

Sample collection

Using a standard Elliot trapping procedure (University of Sydney Animal Ethics: 2018/1438) [21], one male and one female adult brown antechinus were trapped in June 2019 at Lane Cove National Park, NSW (Figure 1). Individuals were euthanased using pentobarbitone (60 mg/mL) and samples were collected immediately after death. Blood samples were collected in RNAprotect® Animal Blood Tubes and stored at 4 ˚C. Tissue samples were either flash frozen in liquid nitrogen (genomic DNA extraction) or placed in RNAlater (transcriptomic RNA extraction) and stored at 4 ˚C overnight before long-term storage at −80 ˚C.

Genome assembly

DNA was extracted from female and male skeletal muscle tissue using the Circulomics Nanobind HMW DNA kit and quantified using a Qubit dsDNA BR (Broad Range) assay and pulse field gel electrophoresis. 10X Genomics linked-read sequencing libraries were prepared at the Ramaciotti Centre for Genomics (Sydney, NSW, Australia) and sequenced on a NovaSeq 6000 S1 flowcell using 150bp PE reads. *De novo* genome assembly was performed for both sexes independently with Supernova v2.1.1 (RRID:SCR_016756) [22] using all reads, obtaining approximately 75× raw coverage and 55× effective (deduplicated) coverage. BBTools v38.73 (RRID:SCR_016968) [23] was used to generate assembly statistics and BUSCO

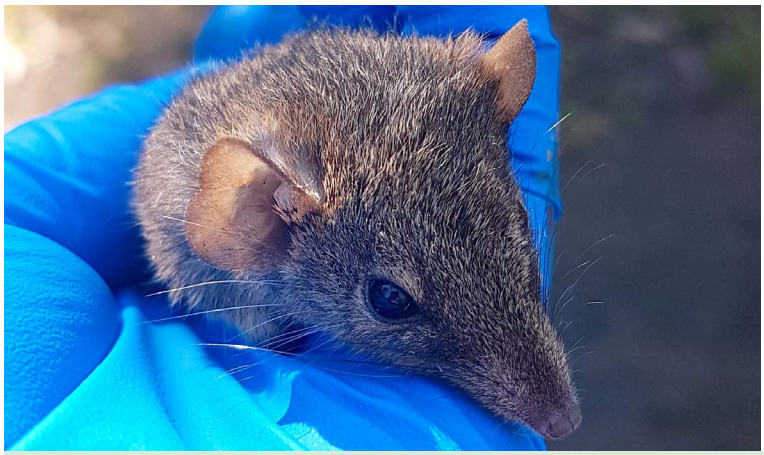

Figure 1. *Antechinus stuartii* individual used for the male reference genome. Image from Carolyn Hogg.

(RRID:SCR_015008) [24] analysis was performed with both v3.0.2 (4,104 mammalian BUSCOs) and v 4.0.6 (9,226 mammalian BUSCOs).

Chromosome assignment and Y chromosome analysis

Putative chromosome assignment of the male assembly was achieved by mapping the male scaffolds to the chromosome-length reference genome of the closely-related Tasmanian devil (*Sarcophilus harrisii*) available on NCBI (RefSeq assembly mSarHar1.11, RRID:SCR_003496) [25] using nucmer v4.0.0beta2 (RRID:SCR_018171) [26] with default parameters and filtering the output using custom bash scripts. Due to the lack of complete Y chromosome sequence in the Tasmanian devil reference genome, additional Y chromosome scaffolds were identified using an AD-ratio (average depth ratio) approach [19] and confirmed through BLAST searches of known marsupial Y genes.

Firstly, both the male and female 10× reads were trimmed to remove the 10× Chromium barcode and low-quality sequence using FastQC v0.11.5 (RRID:SCR_014583) [27] and BBTools (RRID:SCR_016968). Male and female trimmed reads were aligned to the male genome assembly separately using BWA (Burrows-Wheeler Aligner) v0.7.17-r1188 (RRID:SCR_010910) [28] with shorter split hits marked as secondary using the *-M* flag, duplicates were removed using samblaster v0.1.24 (RRID:SCR_000468) [29] with duplicates excluded using the *-e* flag, and alignments with quality scores <20 were removed with samtools v1.10 (RRID:SCR_002105) [30] using the *-q* flag. The output file was converted to bam format, sorted and indexed with samtools and average coverage statistics were generated using Mosdepth v0.2.6 (RRID:SCR_018929) [31] in fast mode. Following a previous study [19], the AD-ratio of each scaffold was calculated for each scaffold whereby a normalized ratio of female reads to male reads should result in a value of ~1 ($0.7 < \text{AD-ratio} < 1.3$) for autosomal scaffolds (as both the male and female should have similar levels of coverage at these regions), a value of ~2 ($1.7 < \text{AD-ratio} < 2.3$) for X chromosome scaffolds (as females should have double the coverage at these regions due to them possessing two X chromosomes) and a value of ~0 ($\text{AD-ratio} \leq 0.3$) for Y chromosomes (as females should have no coverage at these regions due to the lack of a Y chromosome).

In order to improve our confidence in the scaffolds assigned as putatively male using the AD-ratio approach, we used BLAST v2.6.0 (RRID:SCR_004870) [32, 33] to map 20 known

marsupial Y genes and their autosomal or X homologs (if available) from a previous study [34]) against the male antechinus assembly. Scaffolds with an AD-ratio < 0.3 and strong BLAST matches (1×10^{-10}) to marsupial Y genes (but not the respective X chromosome homologs), were deemed as belonging to the Y chromosome.

Transcriptome assembly, annotation and analysis

Total RNA (excluding miRNA) was extracted from blood using the Qiagen RNeasy Protect Animal Blood Kit, and from tissues using the Qiagen RNeasy Mini Kit with quantification performed using the Agilent Bioanalyzer RNA 6000 Nano Kit. TruSeq Stranded mRNA-seq library preparation was performed on male and female spleen, brain, adrenal gland and reproductive tissues (ovary/testis) at the Ramaciotti Centre for Genomics (Sydney, NSW, Australia), and sequenced as 150bp PE reads on a NovaSeq 6000 SP flowcell. RNA-seq reads were quality trimmed and assembled *de novo* to create a global transcriptome assembly using Trinity v2.10.0 (RRID:SCR_013048) [35, 36] with default Trimmomatic (RRID:SCR_011848) [37] and Trinity parameters. Trinity's TrinityStats.pl script was used for general assembly statistics, representation of full-length reconstructed protein-coding genes was examined by Swiss-Prot (RRID:SCR_002380) [38] BLAST searches (RRID:SCR_004870), and completeness was assessed using BUSCO (RRID:SCR_015008) v3 and v4. Trimmed reads were mapped back to the assembly using bowtie2 v2.3.5.1 (RRID:SCR_005476) [39] with a maximum of 20 distinct, valid alignments for each read (using the *-k* flag) to determine read representation. Transcript abundance for each tissue type was estimated using Trinity (RRID:SCR_013048) and Salmon v1.0.0 (RRID:SCR_017036) [40] with default parameters to create a cross-sample TMM normalised matrix of expression values [41, 42]. Finally, the ExN50 statistic was calculated using the normalised expression data. This statistic calculates the N50 for the most highly expressed genes thereby excluding any lowly expressed contigs which are often very short (due to low read coverage preventing assembly of complete transcripts) and hence provides a more useful indicator of transcriptome quality than the standard N50 metric [36].

Functional annotation of the global transcriptome was performed using Trinotate v3.2.0 (RRID:SCR_018930) [43]. Briefly, TransDECODER v5.5.0 (RRID:SCR_017647) was used to identify candidate coding regions within the Trinity transcripts with default parameters. Blast searches of the TransDECODER peptides and Trinity transcripts were performed against the Swiss-Prot (RRID:SCR_002380) database and the Tasmanian devil reference genome annotations from NCBI (RefSeq assembly mSarHar1.11, RRID:SCR_003496) [25] with an e-value cut-off of 1×10^{-5}. HMMER v3.2.0 (RRID:SCR_005305) [44] was used to identify conserved protein domains with the Pfam (RRID:SCR_004726) [45] database, SignalP v4.1 (RRID:SCR_015644) [46] was used to predict signal peptides and RNAmmer v1.2 (RRID:SCR_017075) [47] was used to detect any ribosomal RNA contamination (all programs were run with default parameters). The results from the above were loaded into a SQLite3 (RRID:SCR_017672) database.

Repeat identification and genome annotation

A custom repeat database was generated with RepeatModeler v2.0.1 (RRID:SCR_015027) [48] and repeats (excluding low complexity regions and simple repeats with the *-nolow* flag) were masked with RepeatMasker (RRID:SCR_012954) v4.0.6 [49]. Genome annotation was performed using Fgenesh++ v7.2.2 (RRID:SCR_018928) [50–52] using optimised gene finding

parameters of the closely related Tasmanian devil (*Sarcophilus harrisii*) with mammalian general pipeline parameters. Transcripts representing the longest protein for each trinity "gene" were extracted from the trinity and trinotate output files for mRNA-based predictions with a custom bash script using seqtk v1.3 (RRID:SCR_018927) and seqkit v0.10.1 (RRID:SCR_018926) [53]. A high-quality non-redundant metazoan protein dataset from NCBI was used for homology-based predictions using the "prot_map" method. *Ab initio* predictions were performed in regions where no genes were predicted by other methods (i.e. mRNA mapping or protein homology). The predicted protein-coding sequences were used in BLAST (RRID:SCR_004870) searches against the Swiss-Prot (RRID:SCR_002380) database with an e-value cut-off of 1×10^{-5} to identify genes with matches to known high quality proteins from other species.

Variant annotation

The male reference genome was altered following the 10× Genomics Long Ranger (RRID:SCR_018925) [54] software recommendations of a maximum 500 fasta sequences as follows: scaffolds <50 kb were extracted and concatenated with gaps of 500 N's and then added to the main genome fasta file as a single scaffold and scaffolds ≥50 kb (428 scaffolds) were listed in the primary_contigs.txt file. A BED file of the assembly gaps was created using faToTwoBit and twoBitinfo (RRID:SCR_005780) [55] to generate the sv_blacklist.bed file. Male and female 10x reads were aligned to the altered male 10x reference genome with whole-genome SNVs, indels and structural variants called and phased using Long Ranger v2.2.2 (RRID:SCR_018925) [54] with the FreeBayes (RRID:SCR_010761) option. Male and female VCF files were merged with bcftools v1.10.1 (RRID:SCR_002105) [30] and variants were annotated using ANNOVAR v20180416 (RRID:SCR_012821) [56, 57] gene-based annotation.

Gene family analysis

Gene ontology (GO) annotation (using the generic GO slim subset) was performed on antechinus proteins based on Swiss-Prot matches using GOnet [58] (RRID:SCR_018977) to identify genes associated with key biological functions.

To identify any rapidly evolving gene families in the antechinus, proteomes from six other target species (Tasmanian devil, koala, opossum, human, mouse and platypus) were downloaded from NCBI (RRID:SCR_003496) [25] and the longest isoform for each gene was extracted using custom bash scripts. Protein sequences from the antechinus Fgenesh++ annotation were also extracted and OrthoFinder v2.4.0 (RRID:SCR_017118) [59, 60] was run with default parameters to identify orthogroups between the 7 target species. CAFE v5 (RRID:SCR_018924) [61, 62] was run on the output data from OrthoFinder (RRID:SCR_017118) using an error model to account for genome assembly error (*-e* flag) and estimating multiple lambda's (gene family evolution rates) for monotremes, marsupials and eutherians (*-y* flag). Significant expansions and contractions within the antechinus branch were examined to identify any interesting patterns.

Alzheimer's genes analysis

Literature searches using the search terms "Alzheimer's" and "gene", and mining the human gene database GeneCards [63] using the keyword "Alzheimer's" were used to identify forty of the most common genes that have previously been associated with

Table 1. Comparison of antechinus genome assembly statistics in comparison with the two current highest-quality marsupial genomes.

Species	Assembly	Genome Size (Gb)	No. Scaffolds ↓	No. Contigs ↓	Scaffold N50 (Mb) ↑	Contig N50 (Mb) ↑	% Genome in Scaffolds > 50 KB ↑	Complete Mammalian BUSCO's v3 (%) ↑	Complete Mammalian BUSCO's v4 (%) ↑
Antechinus (M)	antechinusM_pseudohap2.1 (USYD_AStu_M*)	3.3	30876	106199	72.7	0.08	96.35	93.3	81.3
Antechinus (F)	antechinusF_pseudohap2.1	3.3	31296	107658	58.2	0.08	96.61	92.9	81.6
Koala	phaCin_unsw_v4.1*	3.2	-	1909	-	11.59	99.11	92.3	81.6
Tasmanian Devil	mSarHar1.11*	3.1	106	445	611.3	62.34	99.97	93.8	80.9

Arrows indicate whether higher or lower numbers are considered better quality. *NCBI Assembly ID.

Alzheimer's disease in humans or mice disease models. Human coding sequences (CDS) for the genes of interest were downloaded from Swiss-Prot (RRID:SCR_002380) and were used in BLAST (RRID:SCR_004870) searches against the Fgenesh++ genome annotations to identify the predicted gene sequences within the male antechinus reference genome. The predicted protein sequences were matched against the predicted coding sequences of the global transcriptome using BLAST (RRID:SCR_004870) to identify candidate transcripts and expression of the candidate genes across the sequenced tissues was explored using the TMM-normalised expression matrix. All sequences were used in BLAST (RRID:SCR_004870) searches back to the Human Swiss-Prot (RRID:SCR_002380) proteome to confirm orthology through reciprocal best hits (RBH) and were aligned to human protein sequences with MUSCLE v3.8.425 [64] in order to determine sequence similarity and identity. SNVs associated with the target genes were explored using the ANNOVAR (RRID:SCR_012821) output.

FINDINGS

Genome assembly

The male and female antechinus genome assemblies were both 3.3 Gb in size. Genome contiguity was slightly higher for the male antechinus with a scaffold N50 of 72.7 Mb in comparison with the female scaffold N50 of 58.2 Mb (Table 1). Both male and female genome assemblies showed completeness scores comparable to the two best marsupial reference genomes currently available (the koala: RefSeq phaCin_unsw_v4.1, and the Tasmanian devil: RefSeq mSarHar1.11), with >90% of the 4,104 version 3 mammalian BUSCO's and >80% of the 9,226 version 4 mammalian BUSCO's being complete (Table 1). Male and female assemblies had 90% and 89% of reads mapped as proper pairs and a gap percentage of 2.75% and 2.29% (which is within the normal gap range for 10x genomics assemblies [22]) respectively. The male assembly was chosen to be the reference genome as it showed the highest contiguity and also includes the Y chromosome.

Chromosome assignment and Y chromosome analysis

The *Dasyuridae* family display a high level of karyotypic conservation with all species having almost identical 2*n* =14 karyotypes [65]. Antechinus chromosomes were therefore bioinformatically assigned by alignment of the male antechinus scaffolds to the chromosome-length Tasmanian devil reference assembly (RefSeq mSarHar1.11). This resulted in 94.3% of the genome being assigned to chromosomes with the remaining 5.7% of the genome being unassigned either due to no matches to the Tasmanian devil genome or due to multiple alignments where there was no best match to a single chromosome

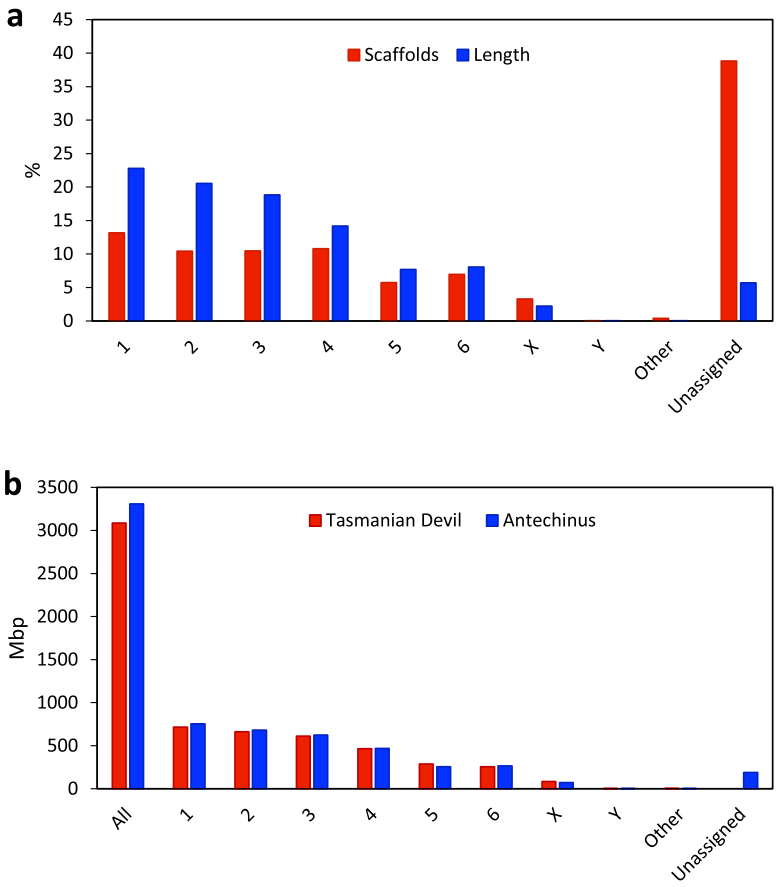

Figure 2. Assignment of antechinus scaffolds to chromosomes by alignment to the Tasmanian devil reference genome. (a) Proportion (%) of scaffolds (blue) and genome length (red) assigned to chromosomes. (b) Comparison of length of sequence assigned to each chromosome from the Tasmanian devil reference genome (blue) and the antechinus genome (red). Other represents scaffolds assigned to "unplaced" Tasmanian devil scaffolds and Unassigned represents scaffolds unable to be assigned due to no matches to the Tasmanian devil genome or due to multiple matches where a best hit to a single chromosome was not identified.

(Figure 2a). The length of assigned antechinus chromosomes was similar to that of the Tasmanian devil as expected (Figure 2b).

The current Tasmanian devil reference genome (RefSeq mSarHar1.11) contains limited Y-chromosome sequence (~130 *kb*) and so only one antechinus scaffold (scaffold 161317, ~73 kb) was assigned as Y chromosome. To identify further putative Y chromosome scaffolds, we implemented an AD-ratio approach (see [19]). Using this approach 3.1 Gb (~95%) of the male genome was assigned as autosomal, 87 Mb (~2. 6%) of the male genome was assigned as X chromosomal and 11.4 Mb (0.3%) of the genome was assigned as Y chromosomal (Figure 3). The results from this approach showed that ~92% of the genome was in agreeance with the chromosome assignment results from mapping the antechinus genome to Tasmanian devil genome with the remaining 8% mainly due to unassigned chromosomes from either method rather than chromosome discrepancies between the two methods (only 0.2% of genome).

In order to identify some high-confidence Y chromosome scaffolds from the putative Y chromosome scaffolds identified with the AD-ratio approach, we aimed to identify scaffolds

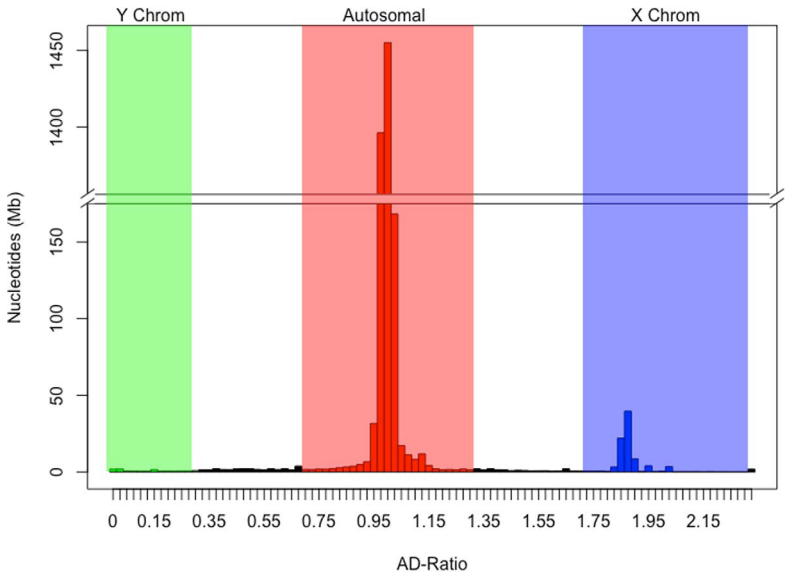

Figure 3. AD-Ratio histogram of antechinus scaffolds. Figure shows the total length of sequence within each 0.025 AD-ratio bin. Scaffolds clustering around an AD-ratio of 0 represent Y-linked sequence (Green), scaffolds clustering around an AD-ratio of 1 represent Autosomal sequence (Red), scaffolds clustering around an AD-ratio of 2 represent X-linked sequence (Blue) and scaffolds between these regions represent unassigned sequence (Black).

containing known Y genes and Y-specific transcripts. Out of 20 known marsupial Y chromosome genes from a previous study [34], 13 showed hits to scaffolds with AD-ratios ≤0.01 indicating a high chance they are putative Y chromosome scaffolds. Furthermore, their autosomal, or X chromosome, homologs mapped to different scaffolds providing additional confidence that the scaffolds identified likely contain the Y homolog. Seven of these Y genes were found to be on scaffold 163451, four were located on scaffold 162475 and one was matched to scaffold 161317 (Figure 4). These scaffolds were deemed Y-chromosome scaffolds and comprise 0.78 Mb of the genome. They represent the largest amount of Y-chromosome sequence characterized in any marsupial species. The remaining gene (ATRY) displayed multiple partial alignment hits to a number of different antechinus scaffolds and could not be reliably annotated to a single scaffold. A number of other genes were also annotated to these scaffolds by Fgenesh++ annotation including an XK-related protein on scaffold 161317, an AMMECR1-like gene on scaffold 163451 and a HMGB3-like protein on scaffold 162475. Identification and annotation of Y chromosome scaffolds in the antechinus will assist with future research wanting to explore male semelparity and key male-specific reproductive genes.

Transcriptome assembly and annotation

The global antechinus transcriptome assembly of 10 tissues (5 male and 5 female) was composed of 1,296,975 transcripts (1,636,859 including predicted splicing isoforms). The average contig length was 773bp and the contig N50 was 1,367bp. Considering only the top 95% most highly expressed transcripts gave an ExN50 (a more useful indicator of transcriptome quality) of 3,020bp which is similar to the average mRNA length in humans (3,392bp) [67]. The assembly showed good overall alignment rates of reads from each of the tissues (>96%) with a high percentage mapped as proper pairs (≥89%). The transcriptome

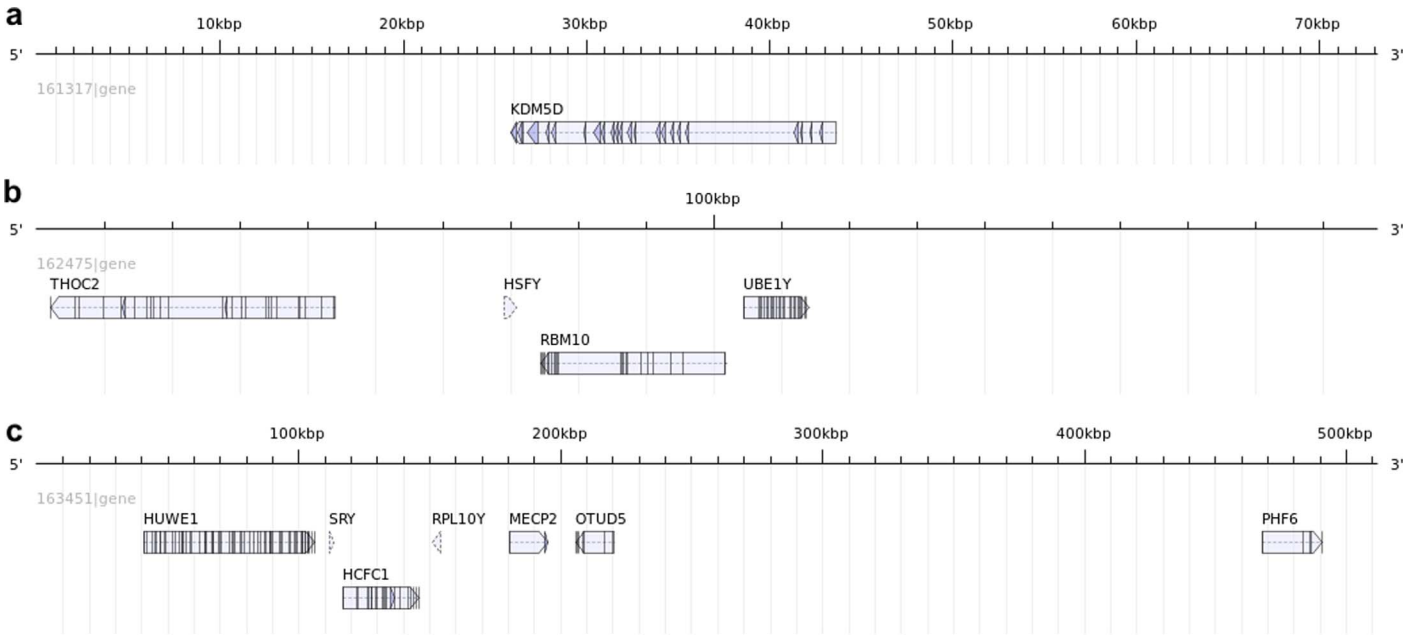

Figure 4. Mapping of known marsupial Y gene homologs on antechinus Y chromosome scaffolds. (a) Scaffold 161317, (b) Scaffold 162475, (c) Scaffold 163451. Figure was created using the AnnotationSketch module from GenomeTools [66].

assembly exhibited similar completeness to the genome with BUSCO analysis identifying 94% and 84% complete BUSCOs for version 3 and version 4 mammalian datasets respectively. TransDecoder predicted 296,706 coding regions within the global transcriptome (including predicted splicing isoforms) of which 181,691 (61%) were complete (contained both a start and stop codon) and 159,121 (54%) had BLAST hits to Swiss-Prot. Taking only the longest complete predicted isoform for each gene resulted in 38,829 mRNA transcripts that were used for genome annotation.

Repeat identification and genome annotation

873 repeat families were identified in the male antechinus genome (Table 2), with 44.82% of the genome being masked as repetitive; a similar repeat content to that of other marsupial and mammalian genomes [68]. A total of 55,827 genes were predicted by Fgenesh++, of which 25,111 had BLAST hits to Swiss-Prot. This number is similar to that of the 26,856 protein-coding genes annotated in the closely related Tasmanian devil reference genome (RefSeq mSarHar1.11). Of these 25,111 gene annotations, 13,189 were predicted based on transcriptome evidence, 1,286 were predicted based on protein evidence and the remaining were predicted *ab initio* based on trained gene finding parameters. BUSCO v3 and v4 completeness scores for the annotation were 78.2% and 67.3% respectively.

Variant annotation

The brown antechinus is predicted to be one of the most common and widespread mammalian species in Eastern Australia where it ranges from southern Queensland to southern New South Wales [69, 70]. The large population size and range of *A. stuartii* implies that this species would likely exhibit healthy levels of genomic diversity, though

Table 2. Summary of repeat classes identified and masked in the antechinus reference genome.

Repeat Class	Count	Masked (bp)	Masked (%)
DNA			
CMC-EnSpm	267774	30028201	0.91%
Ginger-1	13763	1594788	0.05%
PIF-Harbinger	763	204495	0.01%
TcMar-Tc1	7165	1616661	0.05%
TcMar-Tc2	3098	1745523	0.05%
TcMar-Tigger	22186	4059186	0.12%
hAT	744	142335	0.00%
hAT-Ac	2400	291924	0.01%
hAT-Charlie	143304	24400026	0.74%
hAT-Tip100	36557	6236166	0.19%
LINE	6840	2038840	0.06%
CR1	301533	59092138	1.79%
Dong-R4	12719	4935572	0.15%
L1	1117136	608623645	18.40%
L2	770053	168785105	5.10%
RTE-BovB	98681	30352289	0.92%
RTE-RTE	64120	17729186	0.54%
LTR			
ERV1	19808	9033177	0.27%
ERVK	56462	49884792	1.51%
ERVL	2556	1297101	0.04%
Gypsy	4842	1375235	0.04%
SINE			
5S-Deu-L2	4816	270426	0.01%
Alu	6938	1367052	0.04%
MIR	1445092	212663300	6.43%
Other			
Unknown	1070813	233112108	7.05%
Satellite	52562	11605904	0.35%
snRNA	382	28484	0.00%
Total	5533107	1482513659	44.82%

there is currently a lack of genome-wide variation information for any antechinus species. Using the linked-read datasets we identify a total of 9,307,342 SNVs and 2,362,144 indels in the male and 16,291,736 SNVs and 3,818,750 indels in the female; with 5,474,811 SNVs (~27%) and 1,079,862 indels (~21%) being genotyped in both individuals. >90% of these variants passed all of the 10X Genomics filters and >99% were phased. Approximately half of the variants were found to be associated with an annotated gene (located within a gene or within 1kb upstream or downstream of a gene) of which 91% were intronic and 2% were exonic (Figure 5a). Within the exonic variants, 58% were nonsynonymous (result in alteration of the protein sequence) and 39% were synonymous (Figure 5b). These results demonstrate considerable genome-wide diversity from just two individuals from the same population. For comparison, just 1,624,852 SNPs (single nucleotide polymorphisms) were identified across 25 individuals of the closely related and endangered Tasmanian devil [71]. Despite the success of *A. stuartii*, other antechinus species, such as the newly-classified and endangered black-tailed dusky antechinus (*A. arktos*), appear in much lower numbers and so may exhibit much lower genome-wide diversity [72]. Most antechinus species diverged in the Pilocene (~5 *mya*) with the brown antechinus and its close relatives separating more recently in the Pleistocene (~2.5 *mya*) [73]. Humans and chimpanzees are predicted to have diverged 7–8 mya [74] but still share 99% of their DNA [75]. The genetic similarity of human

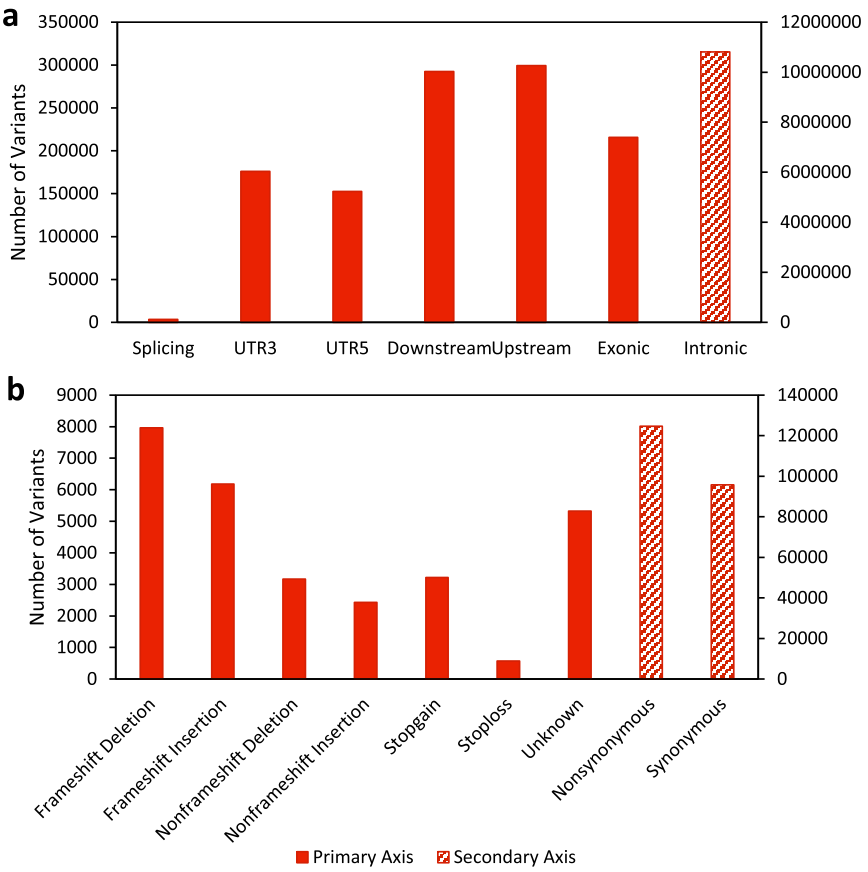

Figure 5. Functional annotation of antechinus variants. (a) Total number of variants annotated to various gene regions including: Splicing (within a splice site of a gene), UTR3 (3′ untranslated region), UTR5 (5′ untranslated region), Downstream (within 1kb downstream of a gene), Upsteam (within 1kb upstream of a gene), Exonic (within the coding sequence of a gene) and Intronic (within an intron of a gene). (b) Total number of exonic variants resulting in specific consequences to the protein sequence including: Frameshift Deletion (deletion of one or more nucleotides that results in a frameshift of the coding sequence), Frameshift Insertion (insertion of one or more nucleotides that results in a frameshift of the coding sequence), Nonframeshift Deletion (deletion of one or more nucleotides that does not result in a frameshift of the coding sequence), Nonframeshift Insertion (insertion of one or more nucleotides that does not result in a frameshift of the coding sequence), Stopgain (variation which results in a stop codon being created within the protein sequence), Stoploss (variation which results in a stop codon being lost from the protein sequence), Unknown (variation with an unknown consequence, perhaps due to complex gene structure), Nonsynonymous (a single nucleotide change that does not result in an amino acid change) and Synonymous (a single nucleotide change that results in an amino acid change). Striped bars indicate variant types that are plotted on the secondary Y-axis.

and chimpanzees (which diverged earlier than the antechinus clades) suggests that the annotated antechinus genome and genome-wide variation provided will be a valuable tool to assist with population monitoring and conservation of all species in the antechinus genus.

In addition to single nucleotide variants, large structural variants can have a pronounced impact on phenotype and account for a significant amount of the diversity seen between individuals [76, 77]. A few interchromosomal and intrachromosomal rearrangements have been identified in the Dasyuridae family using previous G-banding techniques [78]; however, advancements in sequencing technologies, such as the linked-read approach utilized in the current study, allow for more fine-scale characterisation of structural variants in a cost-effective and reliable manner [79]. Using the linked-read datasets, 700

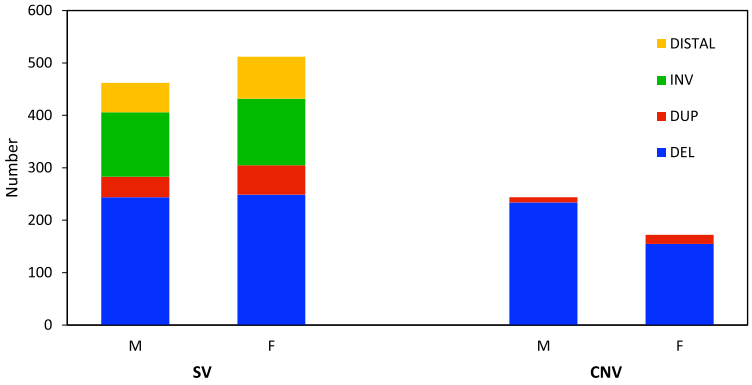

Figure 6. Breakdown of high-quality large structural variants (SVs) and copy number variants (CNVs) in the antechinus. Figure shows both male (M) and female (F) deletions (blue), tandem duplications (red), inversions (green) and distal structural variants (i.e. across two scaffolds, yellow).

large, high-quality structural variants were called in the male and 681 were called in the female of which 35% and 25% were copy number variants (CNVs) respectively (Figure 6). Within the intrachromosomal structural variants, 240 in the male, and 191 in the female were found to contain genes, together encompassing 2,401 genes in total. These findings demonstrate the importance of applying new structural variant identification techniques to explore functional diversity and should be applied more broadly to other Dasyurid species, particularly endangered species such as the Tasmanian devil.

Gene family analysis

GO analysis of the antechinus genome annotations based on matches to Swiss-Prot revealed 2,578 of the genes are involved in response to stress, 1,760 are involved in immune system processes and 1,035 are involved in reproduction. Future studies could use these annotations to design a targeted approach for monitoring the expression of key genes across the breeding season to better understand the interplay between stress, immunity and reproduction in this semelparous species.

To identify any interesting patterns of gene family evolution in the antechinus, proteomes across 7 target species (antechinus, Tasmanian devil, koala, opossum, human, mouse and platypus) were compared and 80.5% of genes were assigned to 19,173 orthogroups of which 12,233 orthogroups had all species present and 9,212 were single-copy orthologs. CAFE identified 282 gene families to be significantly fast evolving. Of these fast-evolving gene families, a number of significant expansions ($<1 \times 10^{-15}$) and contractions were found on the antechinus branch. Many of these expansions and contractions were found in large, complex gene families including olfactory receptors and immune genes which are notoriously difficult to annotate using automated gene annotation methods, particularly in fragmented assemblies, and so require further investigation and manual curation for confirmation. Two other particularly interesting expansions occurred within the protocadherin gamma (Pcdh-γ) gene family (Orthogroup OG0000022) and the NRMK2 gene in the antechinus (Orthogroup OG0000350).

Protocadherins (Pcdhs) belong to the cadherin superfamily and are organised into 3 main gene clusters: α, β and γ [80]. Pcdhs, like all cadherins, are primarily responsible for mediating cell-cell adhesion [81]. Antechinus displayed similar numbers of putative

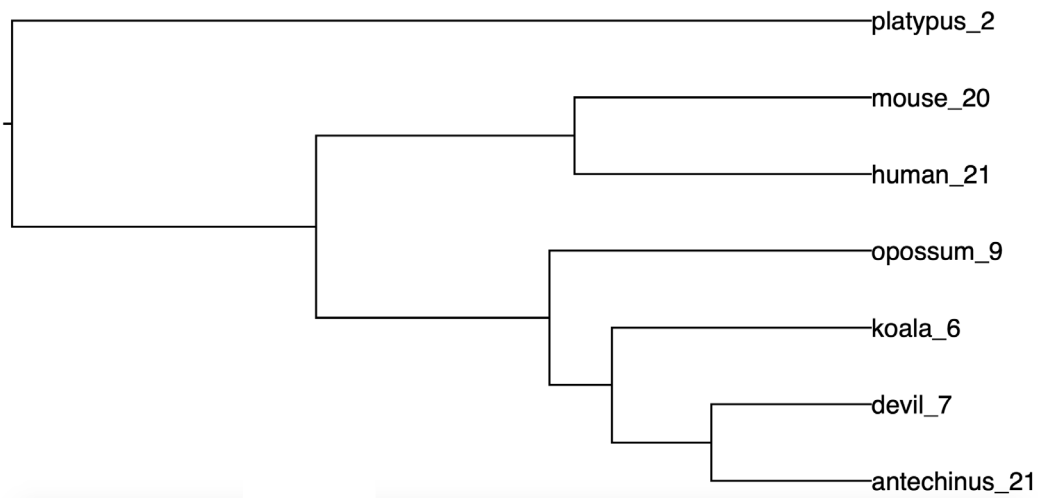

Figure 7. Gene tree showing numbers of Pcdh-γ genes across 7 species.

Pcdh-γ genes as humans and mouse (20–21 genes) in comparison to the other marsupials which showed only 6–9 genes in this family, and the platypus only 2 (Figure 7). Pcdh-γ genes specifically have been implicated in neuronal processes [80] and have previously been associated with Alzheimer's disease [82]. These genes are most highly expressed in the brain in humans and also showed highest levels of expression in the brain and adrenal gland in the antechinus. It is possible that the expansion of Pcdh-γ genes in the antechinus may be linked to the neuropathological changes that occur in mature antechinus. The α and β Pcdhs were also identified as fast evolving across the 7 target species investigated, with marsupials having lower numbers of genes than eutherians, though there were no large differences in the antechinus branch for these clusters.

The antechinus was also found to contain a significant expansion of the NMRK2 gene which appears to be single copy in each of the other species. The NMRK2 gene (Nicotinamide Riboside Kinase 2) is involved in the production of NAD+ (Nicotinamide Adenine Dinucleotide), an essential co-enzyme for various metabolic pathways [83, 84]. The antechinus contains 11 full-length copies of this gene in its genome (Figure 8). Furthermore, genes encoding the subunits of the NADH dehydrogenase enzyme which is responsible for conversion of NADH to NAD+, were among the most highly expressed genes within the antechinus transcriptome across a variety of tissue types. Declining levels of NAD+ have been associated with aging, suggesting that NAD+ may be a key promoter of longevity [84]. NAD+ has also been associated with Alzheimer's disease whereby increased levels of the molecule may be a protective factor of the disease [85]. The antechinus collected in the current study were collected just prior to the annual breeding season and were therefore mature adults. However, the observed neuropathologies in antechinus species are found to be most prominent in post-breeding individuals and so the data presented here will provide a useful comparison for future studies that explore the development of these pathologies and associated genetic changes across the breeding season. Further investigations into the unique expansion of NMRK2 genes in the antechinus may provide crucial insights into aging and age-related dementias in humans.

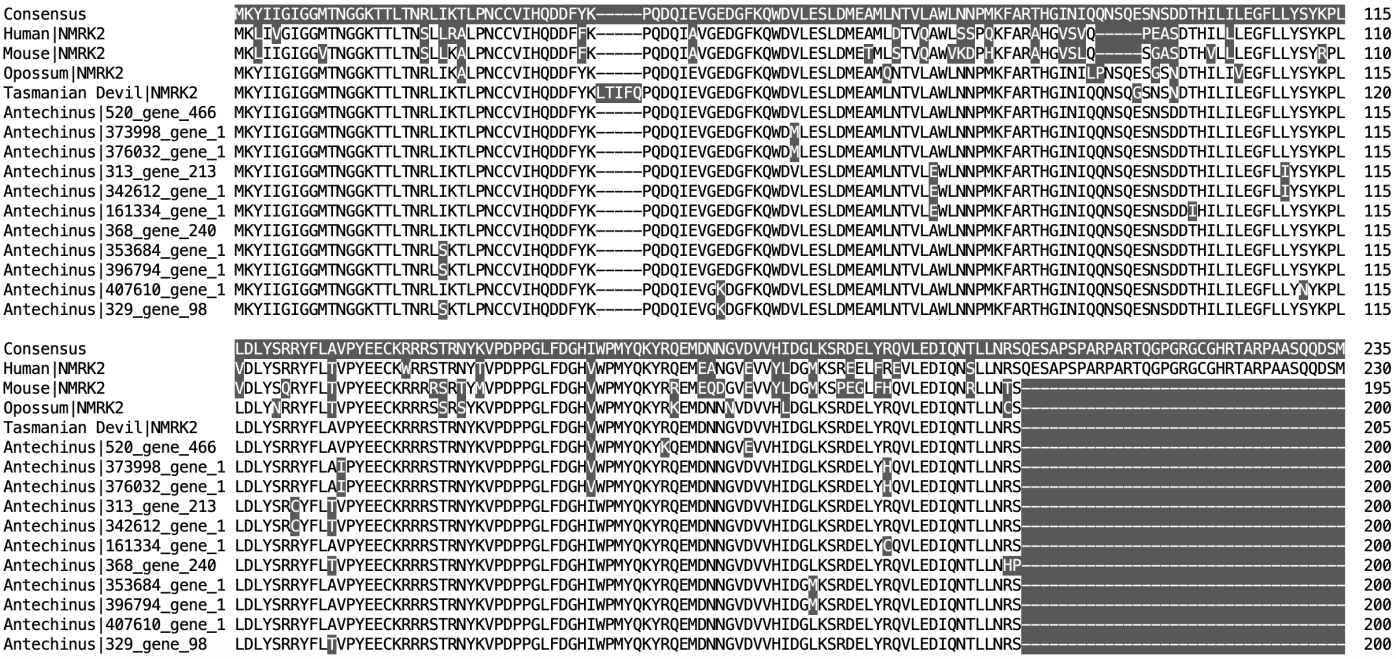

Figure 8. Protein sequence alignment showing expansion of NMRK2 genes in the antechinus. Single copy genes in the human, mouse, gray short-tailed opossum and Tasmanian devil are shown for comparison.

Alzheimer's genes analysis

To investigate further the potential of antechinus being a disease model for AD [3, 9], we analysed expression and identified variation in genes that have previously been associated with AD. Of the 40 target Alzheimer's-associated genes, 39 were annotated in the male antechinus reference genome and all 40 were found to be expressed in the global transcriptome (Table 3). The CD2AP gene was not annotated by Fgenesh++ so was not included in downstream analysis. All of the annotated antechinus proteins except PLD3 were found to be orthologous to the human proteins using a RBH strategy (Table 3). Although the human PLD4 gene was the best BLAST hit for the putative antechinus PLD3 gene, the percentage identity was higher for the human PLD3 gene and the respective antechinus transcript was annotated as PLD3, and therefore this gene was included in further analysis as a putative PLD3 gene. 33 proteins showed >30% similarity to humans [86] (Table 3). Of the seven antechinus gene annotations that showed poor similarity to humans, three (SORL1, CLNK and SLC24A4) were found to have homologous protein-coding transcripts in the global transcriptome suggesting the genome annotations were poor for these genes (likely due to gaps in the reference genome) (Table 3). The remaining four genes (CD33, ZCWPW1, ABCA7 and CR1) did not have homologous genome annotations or transcripts in the antechinus (large gaps were displayed in all sequences compared to the human genes) and were therefore excluded from downstream analysis. Six of the target genes, including APP, PICALM, KAT8, APOE, INPP5D and MAPT were within the top 90% most highly expressed genes of the global transcriptome and were all found to be expressed in the brain. Of these genes, APP (amyloid precursor protein) showed the highest level of expression in antechinus brain tissue. APP is the precursor for

Table 3. Summary of Alzheimer's related genes explored in the Antechinus.

Gene	Gene ID*	Evidence**	Trans ID†	Protein Length (Tran) (bp)	Human Protein Length (bp)	RBH‡	% Ident (Tran)	% Sim (Tran)
APP	76_gene_264	TRINITY_DN490_c2_g1_i21.p1		716	770	Y	86.4	89.9
PSEN1	3_gene_296	*Ab Initio* (PSEN1)	TRINITY_DN960_c7_g2_i1.p1	192 (471)	467	Y	33.97 (88.09)	35.26 (90.95)
CLU	310_gene_647	TRINITY_DN135507_c1_g1_i17.p1		474	449	Y	24.49	39.3
CASS4	3_gene_1296	TRINITY_DN11493_c2_g1_i11.p1		835	786	Y	52.01	63.71
PTK2B	3_gene_1535	Ab Initio (PTK2B)	TRINITY_DN1539_c3_g1_i7.p1	797 (1010)	1009	Y	73.34 (92.57)	76.11 (96.23)
FERMT2	3_gene_6	*Ab Initio* (FERMT2)	TRINITY_DN7191_c0_g1_i2.p1	691 (449)	680	Y	96.96 (60.93)	97.68 (61.94)
MEF2C	0_gene_1343	TRINITY_DN99999960_c0_g1_i3.p1		473	473	Y	99.15	99.58
BIN1	2_gene_709	TRINITY_DN1425_c0_g1_i26.p1		567	593	Y	83.31	88.31
PSEN2	120_gene_116	TRINITY_DN4085_c2_g1_i5.p1		456	448	Y	80.83	85.19
ADAM10	143_gene_1431	TRINITY_DN1482_c5_g1_i3.p1		748	748	Y	93.98	96.12
APH1B	143_gene_1624	TRINITY_DN38091_c0_g1_i11.p1		258	257	Y	84.51	88.68
PICALM	145_gene_551	PROTMAP (PICALM)	TRINITY_DN1843_c1_g1_i11.p1	686 (582)	652	Y	70.93 (87.42)	80.23 (87.88)
DSG2	226_gene_142	TRINITY_DN143_c0_g1_i3.p1		1128	1118	Y	92.59	93.59
ABI3	266_gene_901	TRINITY_DN872_c0_g1_i4.p1		281	366	Y	61.61	72.77
UNC5C	267_gene_1483	*Ab Initio* (UNC5C)	TRINITY_DN20949_c0_g1_i25.p1	852 (932)	931	Y	53.01 (94.41)	60.38 (96.56)
KAT8	96_gene_480	TRINITY_DN613_c1_g1_i45.p1		313	458	Y	79.75	82.04
EPHA1	333_gene_132	TRINITY_DN2610_c0_g2_i6.p1		979	976	Y	63.1	64.19
ECHDC3	333_gene_809	TRINITY_DN23306_c0_g1_i7.p1		228	303	Y	80.82	86.73
CNTNAP2	333_gene_95	*Ab Initio* (CNTNAP2)	TRINITY_DN4057_c0_g2_i4.p1	329 (1325)	1331	Y	60.73 (88.73)	66.01 (91.66)
SORL1	334_gene_344	*Ab Initio* (SORL1)	TRINITY_DN433_c10_g1_i1.p1	1335 (2158)	2214	Y	19.31 (85.37)	20.89 (91.1)
ADAMTS4	335_gene_787	TRINITY_DN799_c4_g1_i2.p1		834	837	Y	37.45	39.57
SCIMP	336_gene_864	TRINITY_DN635_c2_g2_i1.p1		126	145	Y	44.52	57.53
ALPK2	359_gene_112	*Ab Initio* (ALPK2)	TRINITY_DN101181_c0_g1_i5.p1	2237 (1670)	2170	Y	39.21 (34.39)	49.52 (43.65)
CD33	135589_gene_1	*Ab Initio* (CD33)	TRINITY_DN1602_c0_g1_i37.p1	135 (154)	364	Y	19.78 (20.88)	24.73 (26.37)
HESX1	366_gene_560	TRINITY_DN20272_c0_g1_i1.p1		189	185	Y	65.61	70.37
APOE	368_gene_218	TRINITY_DN19355_c0_g1_i12.p1		301	317	Y	42.81	58.41
CELF1	401_gene_24	TRINITY_DN2651_c0_g1_i21.p1		486	486	Y	98.56	98.97
ZCWPW1	427_gene_269	TRINITY_DN2266_c1_g1_i50.p1		255	648	Y	23.9	28.59
MS4A1	432_gene_744	TRINITY_DN3467_c2_g1_i2.p1		287	297	Y	54.85	67.89
CD2AP	NA	NA	TRINITY_DN1647_c3_g1_i14.p1	641 (635)	639	Y	73.58 (74.53)	82.95 (84.01)
AKAP9	499_gene_50	TRINITY_DN250_c13_g1_i6.p1		3783	3907	Y	66.57	75.06
CLNK	535_gene_122	*Ab Initio* (CLNK)	TRINITY_DN108659_c0_g1_i21.p1	677 (342)	428	Y	13.98 (28.26)	24.25 (37.31)
TREM2	608_gene_42	*Ab Initio* (TREM2)	TRINITY_DN33032_c0_g1_i3.p1	261 (287)	230	Y	43.77 (40)	53.96 (49.66)
ABCA7	614_gene_160	TRINITY_DN1943_c1_g1_i15.p1		716	2146	Y	19.83	23.39
CR1	561032_gene_3/ 560671_gene_3	*Ab Initio* (CR1)	TRINITY_DN3772_c0_g1_i39.p1	511 (366)	2039	Y	12.64/12.64 (8.2)	15.63/15.63 (11.96)
SLC24A4	3_gene_564	*Ab Initio* (SLC24A4)	TRINITY_DN8568_c0_g1_i2.p1	304 (543)	622	Y	19.35 (78.69)	23.77 (82.85)
NME8	366_gene_413	TRINITY_DN1228_c0_g1_i1.p1		158	588	Y	65.69	71.64
INPP5D	336_gene_1122	*Ab Initio* (INPP5D)	TRINITY_DN3238_c0_g1_i8.p1	1068 (1209)	1189	Y	39.29 (77.33)	53.57 (84.25)
PLD3	432_gene_623	TRINITY_DN4411_c0_g1_i31.p1		520	490	N (PLD4)	32.96	37.94
MAPT	266_gene_1071	Ab Initio (MAPT)	TRINITY_DN1333_c2_g1_i5.p1	754 (418)	758	Y	41.48 (41.78)	47.42 (43.54)

*ID corresponding to the Fgenesh++ genome annotation. **Evidence for the genome prediction – Transcriptome evidence = TRINITY ID, Protein evidence = PROTMAP Gene ID, Ab Initio Predictions = Top BLAST hit. †For genes without transcriptome evidence the annotations were used in BLAST searches against the predicted protein sequences from the global antechinus transcriptome to identify candidate transcripts. Values associated with these proteins are provided in brackets in the following tables to distinguish them from the genome annotations. ‡Reciprocal Best Hit of antechinus and human genes was a match.

the amyloid beta (Aβ) proteins that form amyloid plaques in the brain and is predicted to contribute to early-onset AD in humans [87]. The MAPT gene was also most highly expressed in antechinus brain tissue and is responsible for the creation of tau proteins which form the neurofibrillary tangles associated with AD [88]. APOE (apolipoprotein E) is

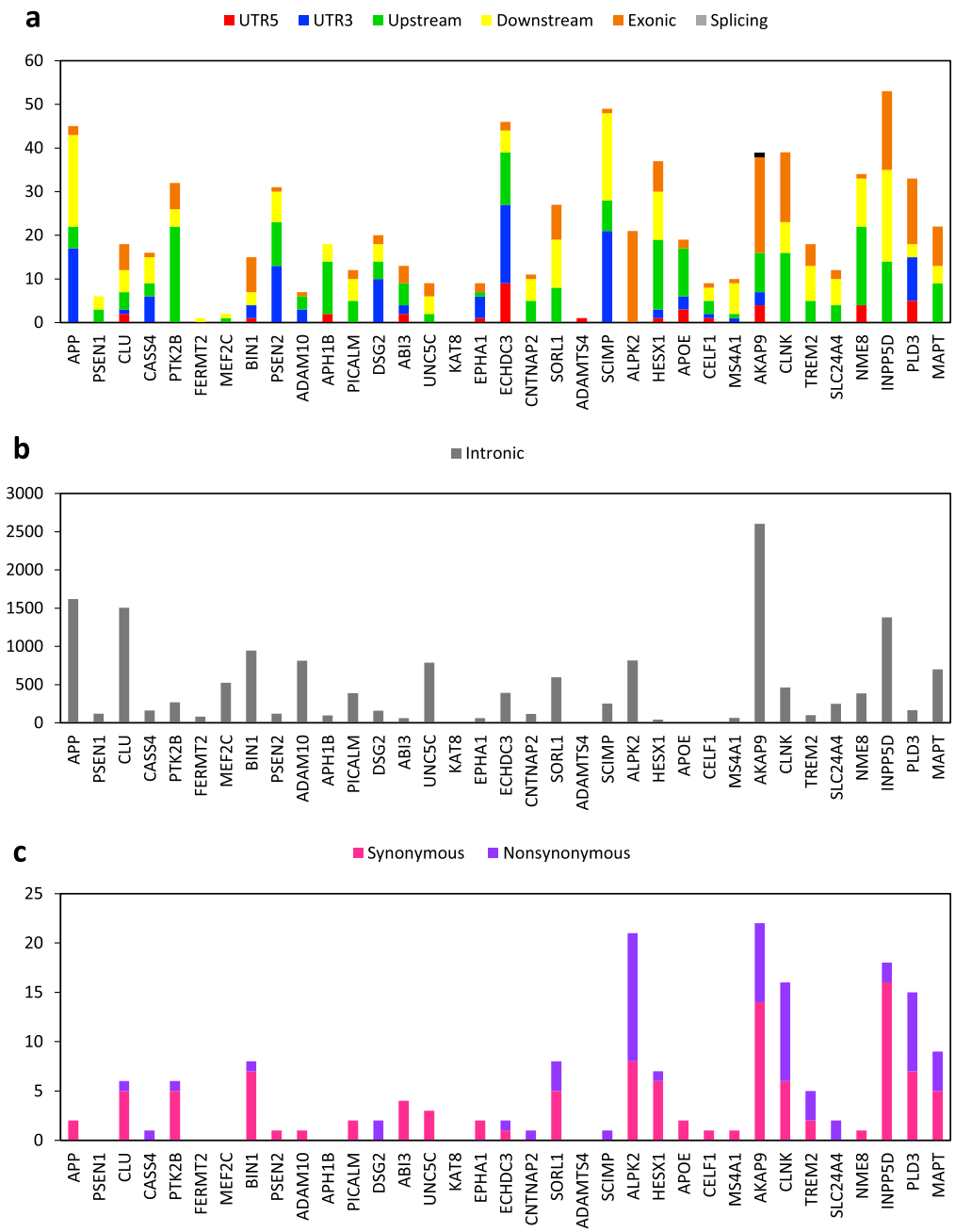

Figure 9. Number of each type of SNV associated with the target Alzheimers-related genes in the antechinus. (a) Numbers of SNVs present in the 5′ UTR, 3′ UTR, 1kb upstream region, 1kb downstream region, exons, and splice sites of each gene. (b) Numbers of intronic SNVs present in each gene. (c) Number of synonymous and nonsynonymous SNVs present in each gene.

the most common risk-factor gene associated with late-onset AD [89] and was highly expressed across a range of antechinus tissues including the brain. PICALM is another common gene which has been associated with an increased risk of developing late-onset AD [90]. PICALM is predicted to help flush Aβ proteins out of the brain and so increased

expression of the PICALM gene in the brain is predicted to reduce AD risk [91]. This gene was found to be quite lowly expressed in antechinus brain tissue when compared with other tissues such as the spleen or in the blood suggesting that it may be contributing to the development of Aβ plaques observed in the antechinus. Finally, KAT8 and INPP5D have been linked to AD through genome-wide association studies [92, 93] and may also be candidates for downstream research. Our finding of expression of some of the most common AD-associated genes in the antechinus brain confirm the potential for this species to be utilized as an AD disease model.

A large variety of genetic variants have been associated with AD in humans, primarily due to their impact on gene expression [92, 94–98]. We utilised the annotated genome-wide SNV data to determine whether antechinus also exhibit variation at Alzheimer's-associated genes. A total of 16,761 high-quality SNVs (which passed all of the 10× Genomics filters) were associated with the 40 target genes with majority of these being intronic (Figure 9). A total of 81 phased nonsynonymous SNVs were identified across 20 of the target genes, of which 24 were genotyped in both the male and female (Figure 9c). While the phenotypic effects of these putatively functional variants are currently unknown, mutations in these genes are commonly associated with AD neuropathologies in humans [92, 94–98] and may also be associated with the age-related development of neuropathologies observed in mature antechinus brains [3].

CONCLUSIONS AND IMPLICATIONS

Here we present the first annotated reference genome within the antechinus genus for a common species, the brown antechinus. The reference genome assembly exhibits completeness comparable to the two current most high-quality marsupial assemblies available (Tasmanian devil and koala), and contains the largest amount of Y-chromosome sequence identified in a marsupial species. Characterisation and annotation of phased, genome-wide variants (including large structural variants) demonstrates considerable diversity within the brown antechinus and provides a resource of gene regions that may have functional implications both in this antechinus and closely related species. Gene ontology analysis of the annotated antechinus proteins identified genes involved in a wide range of biological processes such as immunity, reproduction and stress demonstrating the value of this reference genome in supporting future work investigating the genetic interplay of such processes in this semelparous species. A comparative analysis revealed a number of fast-evolving gene families in the antechinus, most notably within the protocadherin gamma family and NMRK2 gene which have previously been associated with aging and/or aging-related dementias. Target gene analysis revealed high levels of expression of some of the most common genes associated with Alzheimer's disease in the brain, as well as a number of associated variants that may be involved in the Alzheimer's-like neuropathological changes that occur in antechinus species. Future research will be able to use the antechinus genome as a springboard to study age-related neurodegeneration, as well as a model for extreme life history trade-offs like semelparity.

DATA AVAILABILITY

The male antechinus reference genome assembly and all raw sequencing reads including the male and female whole genome 10× genomics reads and the 10 tissue transcriptome RNA-seq reads are available from NCBI under the BioProject accession [PRJNA664282].

All other data sets supporting the results of this article are available in the *GigaScience* GigaDB repository [99].

DECLARATIONS

ABBREVIATIONS

AD: Alzheimer's disease; RNA: ribonucleic acid; miRNA: microRNA; DNA: deoxyribonucleic acid; SNV: single nucleotide variant; HMW: high molecular weight; bp: base pairs; kb: kilobase pairs; Mb: megabase pairs; Gb: gigabase pairs; PE: paired-end; BUSCO: Benchmarking Universal Single-Copy Orthologs; AD-ratio: average depth ratio; BLAST: Basic Local Alignment Search Tool; NCBI: National Center for Biotechnology Information; BED: Browser Extensible Data; VCF: Variant Call Format; GO: Gene Ontology; CDS: coding domain sequence; ANNOVAR: Annotate Variation; CAFE: computational analysis of gene family evolution; CNV: copy number variant; SV: structural variant; SNP: single nucleotide polymorphism; RBH: reciprocal best hit.

ETHICS STATEMENT

All samples were collected in accordance with the *Animal Research Act 1985*, *Animal Research Regulation 2010*, the *Australian code for the care and use of animals for scientific purposes 8th edition 2013* (the Code) and the *Biodiversity Conservation Act 2016*. University of Sydney Animal Ethics Committee number: 2018/1438 and NSW Scientific License number SL101204.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This project was supported by the Australasian Wildlife Genomics Group at The University of Sydney.

AUTHORS' CONTRIBUTIONS

P.B., K.B. and C.H. conceived and designed the project. K.B. and C.H. provided funding. P.B., C.H. and R.S.P.J. collected the samples, P.B prepared the samples, and P.B. and S.T. analysed the data. P.B drafted the manuscript. S.T, C.H, R.S.P.J. and K.B modified the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We thank Emma Peel for her assistance with DNA and RNA extractions and advice on sample collection. We thank Peter Banks and Mathew Crowther for their guidance on trapping procedures and the Desert Ecology Research Group at the University of Sydney for providing us with the necessary trapping equipment. This research was supported by the Sydney Informatics Hub and the Australian BioCommons which is enabled by NCRIS. Compute resources were provided through a University of Sydney partnership with RONIN and AWS (Amazon Web Services) with the support of Intel.

REFERENCES

1 **Braithwaite RW**, **Lee AK**, A mammalian example of semelparity. ***Am. Nat.***, 1979; **113**(1): 151–155.

2 **Cole LC**, The population consequences of life history phenomena. ***Q. Rev. Biol.***, 1954; **29**(2): 103–137.

3 **Naylor R**, **Richardson S**, **McAllan B**, Boom and bust: a review of the physiology of the marsupial genus Antechinus. ***J. Comp. Physiol. B Biochem. Syst. Environ. Physiol.***, 2008; **178**(5): 545–562.

4 **Lee AK**, **Cockburn A**, Evolutionary Ecology of Marsupials. Cambridge University Press 1985.

5 **Promislow DEL**, **Harvey PH**, Living fast and dying young: A comparative analysis of life-history variation among mammals. ***J. Zool.***, 1990; **220**(3): 417–437, doi:10.1111/j.1469-7998.1990.tb04316.x.

6 **Bradley A**, **McDonald I**, **Lee A**, Stress and mortality in a small marsupial (Antechinus stuartii, Macleay). ***Gen. Comp. Endocrinol.***, 1980; **40**(2): 188–200.

7 **P Woolley**, Reproduction in Antechinus spp. and other dasyurid marsupials. ***Symp. Zool. Soc. Lond.***, 1966; 281–294.

8 **Lee AK**, **Bradley AJ**, **Braithwaite RW**, Corticosteroid levels and male mortality in Antechinus stuartii. In: The Biology of Marsupials. Springer 1977; pp. 209–220.

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

10 **McAllan B**, **Hobbs S**, **Norris D**, Effects of stress on the neuroanatomy of a marsupial. ***J. Exp. Zool. A Comp. Exp. Biol.***, 2006; **305A**: 154.

11 **Ulep MG**, **Saraon SK**, **McLea S**, Alzheimer disease. ***J. Nurse. Pract.***, 2018; **14**(3): 129–135, doi:10.1016/j.nurpra.2017.10.014.

12 **Götz J**, **Streffer J**, **David D**, **Schild A**, **Hoerndli F**, **Pennanen L et al.** Transgenic animal models of Alzheimer's disease and related disorders: histopathology, behavior and therapy. ***Mol. Psychiatry***, 2004; **9**(7): 664–683.

13 **Schwab C**, **Hosokawa M**, **McGeer PL**, Transgenic mice overexpressing amyloid beta protein are an incomplete model of Alzheimer disease. ***Exp. Neurol.***, 2004; **188**(1): 52–64.

14 **Elder GA**, **Gama Sosa MA**, **De Gasperi R**, Transgenic mouse models of Alzheimer's disease. ***Mt. Sinair. J. Med.***, 2010; 77(1): 69–81.

15 **Reardon S**, Frustrated Alzheimer's researchers seek better lab mice. ***Nature***, 2018; **563**(7731): 611–613.

16 **King A**, The search for better animal models of Alzheimer's disease. ***Nature***, 2018; **559**(7715): S13.

17 **Holleley CE**, **Dickman CR**, **Crowther MS**, **Oldroyd BP**, Size breeds success: multiple paternity, multivariate selection and male semelparity in a small marsupial, Antechinus stuartii. ***Mol. Ecol.***, 2006; **15**(11): 3439–3448, doi:10.1111/j.1365-294X.2006.03001.x.

18 **Wood D**, An ecological study of Antechinus stuartii (Marsupialia) in a south-east Queensland rain forest. ***Aust. J. Zool.***, 1970; **18**(2): 185–207.

19 **Bidon T**, **Schreck N**, **Hailer F**, **Nilsson MA**, **Janke A**, Genome-wide search identifies 1.9 Mb from the polar bear Y chromosome for evolutionary analyses. ***Genome Biol. Evol.***, 2015; **7**(7): 2010–2022.

20 **Toder R**, **Wakefield M**, **Graves J**, The minimal mammalian Y chromosome–the marsupial Y as a model system. ***Cytogenet. Genome Res.***, 2000; **91**(1–4): 285–292.

21 **Tasker EM**, **Dickman CR**, A review of Elliott trapping methods for small mammals in Australia. ***Aust. Mammal.***, 2001; **23**(2): 77–87.

22 **Weisenfeld NI**, **Kumar V**, **Shah P**, **Church DM**, **Jaffe DB**, Direct determination of diploid genome sequences. ***Genome Res.***, 2017; **27**(5): 757–767.

23 **Bushnell B**, BBTools. sourceforge.net/projects/bbmap/ (2014).

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

25 **O'Leary NA**, **Wright MW**, **Brister JR**, **Ciufo S**, **Haddad D**, **McVeigh R et al.** Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. ***Nucleic Acids Res.***, 2015; **44**(D1): D733–D745.

26 **Kurtz S**, **Phillippy A**, **Delcher AL**, **Smoot M**, **Shumway M**, **Antonescu C et al.** Versatile and open software for comparing large genomes. ***Genome Biol.***, 2004; **5**(2): R12.

27 **Andrews S**, FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010). Accessed 29th April 2020.

28 **Li H**, **Durbin R**, Fast and accurate short read alignment with Burrows–Wheeler transform. ***Bioinformatics***, 2009; **25**(14): 1754–1760, doi:10.1093/bioinformatics/btp324.

29 **Faust GG**, **Hall IM**, SAMBLASTER: fast duplicate marking and structural variant read extraction. ***Bioinformatics***, 2014; **30**(17): 2503–2505.

30 **Li H**, **Handsaker B**, **Wysoker A**, **Fennell T**, **Ruan J**, **Homer N et al.** The sequence alignment/map format and SAMtools. ***Bioinformatics***, 2009; **25**(16): 2078–2079, doi:10.1093/bioinformatics/btp352.

31 **Pedersen BS**, **Quinlan AR**, Mosdepth: quick coverage calculation for genomes and exomes. ***Bioinformatics***, 2017; **34**(5): 867–868.

32 **Altschul SF**, **Gish W**, **Miller W**, **Myers EW**, **Lipman DJ**, Basic local alignment search tool. ***J. Mol. Biol.***, 1990; **215**(3): 403–410, doi:10.1016/S0022-2836(05)80360-2.

33 **Camacho C**, **Coulouris G**, **Avagyan V**, **Ma N**, **Papadopoulos J**, **Bealer K et al.** BLAST+: architecture and applications. ***BMC Bioinformatics***, 2009; **10**(1): 421.

34 **Cortez D**, **Marin R**, **Toledo-Flores D**, **Froidevaux L**, **Liechti A**, **Waters PD et al.** Origins and functional evolution of Y chromosomes across mammals. ***Nature***, 2014; **508**(7497): 488.

35 **Grabherr MG**, **Haas BJ**, **Yassour M**, **Levin JZ**, **Thompson DA**, **Amit I et al.** Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. ***Nat. Biotechnol.***, 2011; **29**(7): 644.

36 **Haas BJ**, **Papanicolaou A**, **Yassour M**, **Grabherr M**, **Blood PD**, **Bowden J 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.

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

38 **Consortium U**, UniProt: a worldwide hub of protein knowledge. ***Nucleic Acids Res.***, 2018; **47**(D1): D506–D515.

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

40 **Patro R**, **Duggal G**, **Love MI**, **Irizarry RA**, **Kingsford C**, Salmon provides fast and bias-aware quantification of transcript expression. ***Nat. Meth.***, 2017; **14**(4): 417.

41 **Dillies M-A**, **Rau A**, **Aubert J**, **Hennequet-Antier C**, **Jeanmougin M**, **Servant N et al.** A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. ***Brief Bioinform.***, 2013; **14**(6): 671–683.

42 **Robinson MD**, **Oshlack A**, A scaling normalization method for differential expression analysis of RNA-seq data. ***Genome Biol.***, 2010; **11**(3): 1–9.

43 **Bryant DM**, **Johnson K**, **DiTommaso T**, **Tickle T**, **Couger MB**, **Payzin-Dogru D et al.** A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. ***Cell Rep.***, 2017; **18**(3): 762–776.

44 **Eddy SR**, HMMER. http://hmmer.org (2018). Accessed 11th May 2020.

45 **El-Gebali S**, **Mistry J**, **Bateman A**, **Eddy SR**, **Luciani A**, **Potter SC et al.** The Pfam protein families database in 2019. ***Nucleic Acids Res.***, 2019; **47**(D1): D427–D432.

46 **Nielsen H**, Predicting secretory proteins with SignalP. ***Methods Mol. Biol.***, 2017; **1611**: 59–73.

47 **Lagesen K**, **Hallin P**, **Rødland EA**, **Stærfeldt H-H**, **Rognes T**, **Ussery DW**, RNAmmer: consistent and rapid annotation of ribosomal RNA genes. ***Nucleic Acids Res.***, 2007; **35**(9): 3100–3108.

48 **Smit A**, **Hubley R**, **Green P**, RepeatModeler Open-1.0. http://www.repeatmasker.org (2008–2015). Accessed 19th December 2019.

49 **Smit A**, **Hubley R**, **Green P**, RepeatMasker Open-4.0. http://www.repeatmasker.org (2013–2015). Accessed 19th December 2019.

50 **Salamov AA**, **Solovyev VV**, Ab initio gene finding in Drosophila genomic DNA. ***Genome Res.***, 2000; **10**(4): 516–522.

51 **Solovyev V**, **Kosarev P**, **Seledsov I**, **Vorobyev D**, Automatic annotation of eukaryotic genes, pseudogenes and promoters. ***Genome Biol.***, 2006; **7**(S1): S10.

52 **Solovyev VV**, Finding genes by computer: probabilistic and discriminative approaches. In: Tao JYX, Zhang MQ (eds), Current Topics in Computational Molecular Biology. 2002; pp. 201–248.

53 **Shen W**, **Le S**, **Li Y**, **Hu F**, SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. ***PLoS ONE***, 2016; **11**(10): e0163962.

54 **Zheng GX**, **Lau BT**, **Schnall-Levin M**, **Jarosz M**, **Bell JM**, **Hindson CM et al.** Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. ***Nat. Biotechnol.***, 2016; **34**(3): 303.

55 **Kent WJ**, **Sugnet CW**, **Furey TS**, **Roskin KM**, **Pringle TH**, **Zahler AM et al.** The human genome browser at UCSC. ***Genome Res.***, 2002; **12**(6): 996–1006.

56 **Wang K**, **Li M**, **Hakonarson H**, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. ***Nucleic Acids Res.***, 2010; **38**(16): e164-e.

57 **Yang H**, **Wang K**, Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. ***Nat. Protoc.***, 2015; **10**(10): 1556–1566, doi:10.1038/nprot.2015.105.

58 **Pomaznoy M**, **Ha B**, **Peters B**, GOnet: a tool for interactive Gene Ontology analysis. ***BMC Bioinformatics***, 2018; **19**(1): 470.

59 **Emms DM**, **Kelly S**, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. ***Genome Biol.***, 2015; **16**(1): 157.

60 **Emms DM**, **Kelly S**, OrthoFinder: phylogenetic orthology inference for comparative genomics. ***Genome Biol.***, 2019; **20**(1): 1–14.

61 **De Bie T**, **Cristianini N**, **Demuth JP**, **Hahn MW**, CAFE: a computational tool for the study of gene family evolution. ***Bioinformatics***, 2006; **22**(10): 1269–1271.

62 **Hahn MW**, **De Bie T**, **Stajich JE**, **Nguyen C**, **Cristianini N**, Estimating the tempo and mode of gene family evolution from comparative genomic data. ***Genome Res.***, 2005; **15**(8): 1153–1160.

63 **Stelzer G**, **Rosen N**, **Plaschkes I**, **Zimmerman S**, **Twik M**, **Fishilevich S et al.** The GeneCards suite: from gene data mining to disease genome sequence analyses. ***Curr. Protoc. Bioinformatics***, 2016; **54**(1): 1.30.1–1..3.

64 **Edgar RC**, MUSCLE: multiple sequence alignment with high accuracy and high throughput. ***Nucleic Acids Res.***, 2004; **32**(5): 1792–1797.

65 **Deakin JE**, Chromosome evolution in marsupials. ***Genes***, 2018; **9**(2): 72.

66 **Gremme G**, **Steinbiss S**, **Kurtz S**, GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. ***IEEE/ACM Trans. Comput. Biol. Bioinform.***, 2013; **10**(3): 645–656.

67 **Piovesan A**, **Caracausi M**, **Antonaros F**, **Pelleri MC**, **Vitale L**, GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. ***Database***, 2016; **2016**.

68 **Margulies EH**, **Maduro VV**, **Thomas PJ**, **Tomkins JP**, **Amemiya CT**, **Luo M et al.** Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. ***Proc. Natl Acad. Sci. USA***, 2005; **102**(9): 3354–3359.

69 **Van Dyck S**, **Crowther M**, Reassessment of northern representatives of the Antechinus stuartii complex (Marsupialia: Dasyuridae): A subtropicus sp. nov. and A. adustus new status. ***Mem. Queensl. Mus.***, 2000; **45**(2): 611–635.

70 **Crowther M**, **Braithwaite RW**, Brown antechinus, Antechinus stuartii. In: Van Dyckm RS S (ed.), The mammals of Australia. Sydney, Australia: Reed New Holland 2008.

71 **Wright BR**, **Farquharson KA**, **McLennan EA**, **Belov K**, **Hogg CJ**, **Grueber CE**, A demonstration of conservation genomics for threatened species management. ***Mol. Ecol. Resour.***, 2020; **00**: 1–16.

72 **Gray EL**, **Baker AM**, **Firn J**, Autecology of a new species of carnivorous marsupial, the endangered black-tailed dusky antechinus (Antechinus arktos), compared to a sympatric congener, the brown antechinus (Antechinus stuartii). ***Mammal. Res.***, 2017; **62**(1): 47–63.

73 **Mutton TY**, **Phillips MJ**, **Fuller SJ**, **Bryant LM**, **Baker AM**, Systematics, biogeography and ancestral state of the Australian marsupial genus Antechinus (Dasyuromorphia: Dasyuridae). ***Zool. J. Linn. Soc.***, 2019; **186**(2): 553–568.

74 **Langergraber KE**, **Prüfer K**, **Rowney C**, **Boesch C**, **Crockford C**, **Fawcett K et al.** Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. ***Proc. Natl Acad. Sci. USA***, 2012; **109**(39): 15716–15721.

75 **Mikkelsen T**, **Hillier L**, **Eichler E**, **Zody M**, **Jaffe D**, **Yang S-P et al.** Initial sequence of the chimpanzee genome and comparison with the human genome. ***Nature***, 2005; **437**(7055): 69–87.

76 **Feuk L**, **Carson AR**, **Scherer SW**, Structural variation in the human genome. ***Nat. Rev. Genet.***, 2006; **7**(2): 85–97.

77 **Mahmoud M**, **Gobet N**, **Cruz-Dávalos DI**, **Mounier N**, **Dessimoz C**, **Sedlazeck FJ**, Structural variant calling: the long and the short of it. ***Genome Biol.***, 2019; **20**(1): 246.

78 **Deakin JE**, **Kruger-Andrzejewska M**, Marsupials as models for understanding the role of chromosome rearrangements in evolution and disease. ***Chromosoma***, 2016; **125**(4): 633–644.

79 **P Balachandran**, **Beck CR**, Structural variant identification and characterization. ***Chromosome Res.***, 2020; **28**: 31–47.

80 **Hayashi S**, **Takeichi M**, Emerging roles of protocadherins: from self-avoidance to enhancement of motility. ***J. Cell Sci.***, 2015; **128**(8): 1455–1464.

81 **Chen WV**, **Maniatis T**, Clustered protocadherins. ***Development***, 2013; **140**(16): 3297–3302.

82 **Li Y**, **Chen Z**, **Gao Y**, **Pan G**, **Zheng H**, **Zhang Y et al.** Synaptic adhesion molecule Pcdh-γC5 mediates synaptic dysfunction in Alzheimer's disease. ***J. Neurosci.***, 2017; **37**(38): 9259–9268.

83 **Yang Y**, **Sauve AA**, NAD+ metabolism: Bioenergetics, signaling and manipulation for therapy. ***Biochim. Biophys. Acta Proteins Proteom.***, 2016; **1864**(12): 1787–1800.

84 **Johnson S**, **Imai S-i**, NAD+ biosynthesis, aging, and disease. ***F1000Research***, 2018; **7**: 132.

85 **Hou Y**, **Lautrup S**, **Cordonnier S**, **Wang Y**, **Croteau DL**, **Zavala E et al.** NAD+ supplementation normalizes key Alzheimer's features and DNA damage responses in a new AD mouse model with introduced DNA repair deficiency. ***Proc. Natl Acad. Sci. USA***, 2018; **115**(8): E1876–E1885.

86 **Kuzniar A**, **van Ham RC**, **Pongor S**, **Leunissen JA**, The quest for orthologs: finding the corresponding gene across genomes. ***Trends Genet.***, 2008; **24**(11): 539–551.

87 **O'Brien RJ**, **Wong PC**, Amyloid precursor protein processing and Alzheimer's disease. ***Annu. Rev. Neurosci.***, 2011; **34**: 185–204.

88 **Iqbal K**, **Liu F**, **Gong C-X**, **Grundke-Iqbal I**, Tau in Alzheimer disease and related tauopathies. ***Curr. Alzheimer. Res.***, 2010; 7(8): 656–664.

89 **Liu C-C**, **Kanekiyo T**, **Xu H**, **Bu G**, Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. ***Nat. Rev. Neurol.***, 2013; **9**(2): 106–118.

90 **Xu W**, **Tan L**, **Yu J-T**, The role of PICALM in Alzheimer's disease. ***Mol. Neurobiol.***, 2015; **52**(1): 399–413.

91 **Zhao Z**, **Sagare AP**, **Ma Q**, **Halliday MR**, **Kong P**, **Kisler K et al.** Central role for PICALM in amyloid-β blood-brain barrier transcytosis and clearance. ***Nat. Neurosci.***, 2015; **18**(7): 978–987.

92 **Tábuas-Pereira M**, **Santana I**, **Guerreiro R**, **Brás J**, Alzheimer's disease genetics: Review of Novel Loci associated with disease. ***Curr. Genet. Med. Rep.***, 2020; **8**(1): 1–16.

93 **Lambert J-C**, **Ibrahim-Verbaas CA**, **Harold D**, **Naj AC**, **Sims R**, **Bellenguez C et al.** Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. ***Nat. Genet.***, 2013; **45**(12): 1452–1458.

94 **Cuyvers E**, **Sleegers K**, Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. ***Lancet Neurol.***, 15; **2016**(8): 857–868.

95 **Mendez MF**, Early-onset Alzheimer disease and its variants. ***Continuum (Minneapolis, Minn)***, 2019; **25**(1): 34.

96 **Sun Q**, **Xie N**, **Tang B**, **Li R**, **Shen Y**, Alzheimer's disease: from genetic variants to the distinct pathological mechanisms. ***Front. Mol. Neurosci.***, 2017; **10**: 319.

97 **Sims R**, **Hill M**, **Williams J**, The multiplex model of the genetics of Alzheimer's disease. ***Nat. Neurosci.***, 2020; **23**(3): 311–322.

98 **Rosenthal SL**, **Kamboh MI**, Late-onset Alzheimer's disease genes and the potentially implicated pathways. ***Curr. Genet. Med. Rep.***, 2014; **2**(2): 85–101.

99 **Brandies PA**, **Tang S**, **Johnson RS**, **Hogg C**, **Belov K**, Supporting data for "The first Antechinus reference genome provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies". 2020, GigaScience Database; http://doi.org/10.5524/100807.