THE UNIVERSITY OF SYDNEY

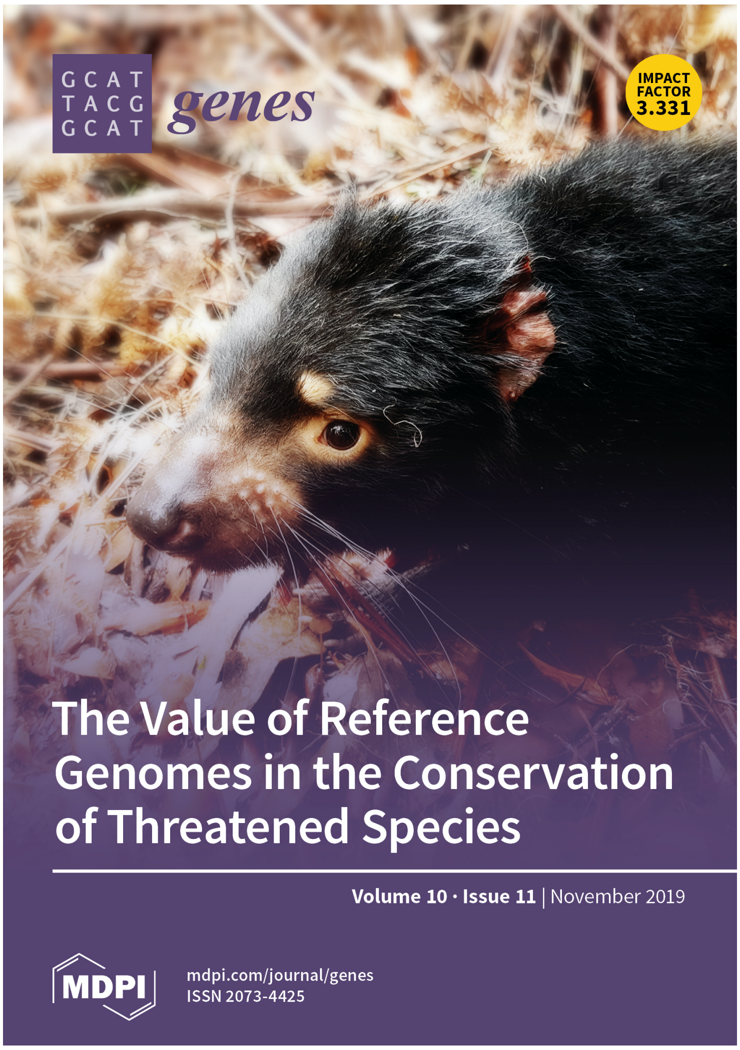
SCHOOL OF LIFE AND ENVIRONMENTAL SCIENCES

FACULTY OF SCIENCE

Conserving Australia's iconic marsupials; one genome at a time

Parice Amber Brandies

Bachelor of Science (Advanced) (Honours I)


GCAT
TACG
GCAT
genes
IMPACT FACTOR 3.331
The Value of Reference Genomes in the Conservation of Threatened Species
Volume 10 · Issue 11 | November 2019
MDPI
mdpi.com/journal/genes
ISSN 2073-4425

A thesis submitted to fulfil requirements for the degree of Doctor of Philosophy

2021

STATEMENT OF ORIGINALITY

This thesis is submitted to the University of Sydney in fulfillment of the requirement for the degree of Doctor of Philosophy. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged throughout. This thesis has not been submitted for any other degree.

Parice Brandies 23/04/2021

ACKNOWLEDGEMENTS

First and foremost, to my wonderful supervisors, Kathy, Carolyn and Catherine, I cannot thank you enough for your guidance, support and encouragement throughout this PhD. You enabled me to follow my interests, provided me with so many incredible opportunities, and opened a whole new world of possibilities for me. I have learnt more than I could ever have imagined these past few years and have grown so much both personally and professionally. Your love for science, your passion for conservation, and your dedication to make a difference in this world constantly inspires me. I will always be so grateful for the knowledge, passion and drive you have all instilled in me.

To the rest of the AWGG lab, especially Kate and Elle, thank you for all of our office chats, all of your insightful advice, and all of your encouragement throughout my PhD. You have all helped make these past few years so enjoyable and I am so lucky to have shared this experience with such a kind, supportive, like-minded group of people who I now call friends. I am so excited to see where each of your paths lead you and know you will all go on to do amazing things.

To my RONIN fam, Nathan, Don, Byron, Aaron, and Darce, thank you not only for enabling me to delve into the world of cloud computing, but for being such big supporters of my research and our lab’s mission. Much of the work presented in this thesis would not have been possible (or at least a lot more difficult) without you and the phenomenal platform you have created. Special thanks to Nathan for introducing me to the wonders of software development and for being that fellow nerdy human that I can always relate to about so many things. Working with you all has been one of the biggest highlights of my PhD and I can’t wait to continue to inspire and help other researchers around the globe with the power of RONIN.

To all of my incredible friends, Jazz, Mara, Sid, Ruth, Riley, Satoshi, Naomi, words can’t describe how grateful I am to have you in my life. You have all been the greatest support system throughout my entire PhD, being there for me when I needed someone to talk to, someone to celebrate my achievements with and someone to do something fun with when I needed a break. Our many food outings and long conversations are what kept me so sane and happy, and I couldn’t have done this without you. In particular, a big thank you to Jazz and Mara for our weekly catchups; spending time with you both is always the highlight of my week and I feel so lucky to have the best friends anyone could ever ask for! A special thanks to one of my greatest mentors and dear friend, Robert, for encouraging me to pursue my passion for genetics and explore a career in research. I wouldn’t be writing this thesis if it weren’t for you.

Finally, to my amazing family, where do I even begin? To my beautiful Mum, thank you for being my TV buddy every night to unwind with, for our many walks together to get me away from my computer, and for your unconditional love and light that you have always brought to my life. To my extraordinary Dad, thank you for our regular movie dates to forget about reality together, for all our long conversations that we share over delicious meals, and for always making me feel so loved and supported. To my brother and partner in crime, Pierce, thank you for never ceasing to make me laugh every day, for our frequent D&Ms about life, and for always supporting me and my aspirations without question. Of course, a special thanks to my tiniest of companions, Bella and Lily, who were right by my side (literally) every day and night, and have been the best thesis-writing, working-from-home buddies you could ever wish for! To my grandparents and everyone else that has been there for me throughout this PhD, thank you for always believing in me, caring for me, and encouraging me. I am so fortunate to have had such a wonderful support system throughout my life and you have all shaped me to be the person I am today. None of this would have been possible without each and every one of you.

TABLE OF CONTENTS

PUBLICATIONS

Johnson, RN, O'Meally, D, Chen, Z, Etherington, GJ, Ho, SYW, Nash, WJ, Grueber, CE, Cheng, Y, Whittington, CM, Dennison, S, Peel, E, Haerty, W, O'Neill, RJ, Colgan, D, Russell, TL, Alquezar-Planas, DE, Attenbrow, V, Bragg, JG, **Brandies, PA**, Chong, AY-Y, Deakin, JE, Di Palma, F, Duda, Z, Eldridge, MDB, Ewart, KM, Hogg, CJ, Frankham, GJ, Georges, A, Gillett, AK, Govendir, M, Greenwood, AD, Hayakawa, T, Helgen, KM, Hobbs, M, Holleley, CE, Heider, TN, Jones, EA, King, A, Madden, D, Graves, JaM, Morris, KM, Neaves, LE, Patel, HR, Polkinghorne, A, Renfree, MB, Robin, C, Salinas, R, Tsangaras, K, Waters, PD, Waters, SA, Wright, B, Wilkins, MR, Timms, P & Belov, K 2018, 'Adaptation and conservation insights from the koala genome', *Nature Genetics,* vol. 50, pp. 1102-1111.

Brandies, PA, Grueber, CE, Hogg, CJ & Belov, K 2019, 'MHC Genes and Mate Choice', in Choe, J (ed.), *Encyclopedia of Animal Behaviour,* 2 edn, Elsevier, Massachusetts, USA.

Brandies, PA, Peel, E, Hogg, CJ & Belov, K 2019, 'The value of reference genomes in the conservation of threatened species', *Genes,* vol. 10, no. 11, pp. 846.

Brandies, PA, Wright, BR, Hogg, CJ, Grueber, CE & Belov, K 2020, 'Characterisation of reproductive gene diversity in the endangered Tasmanian devil', *Molecular Ecology Resources,* vol. 00, pp. 1-12.

Brandies, PA, Tang, S, Johnson, RSP, Hogg, CJ & Belov, K 2020, 'The first *Antechinus* reference genome provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies', *Gigabyte,* vol. 1, no. 7, pp. 1-22.

Brandies, PA & Hogg, CJ 2021, 'Ten simple rules for getting started with command-line bioinformatics', *PLOS Computational Biology,* vol. 17, no. 2, pp. e1008645.

PRESENTATIONS

CONFERENCE PRESENTATIONS

February, 2019 ‘Koala MHC: Perks of a PacBio Genome.’
Poster Presentation: Lorne Genome Conference 2019, Lorne, Victoria

October, 2020 ‘The first reference genome for the *Antechinus* genus provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies’
Oral Presentation: Biodiversity Genomics Conference 2020, Online

November, 2020 ‘Using the processing power of Intel and RONIN to save endangered species’
Oral Presentation: HPC on AWS Global Conference 2020, Online

March, 2021 ‘The first reference genome for the *Antechinus* genus provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies’
Oral Presentation: The 15th International Symposium on Primatology and Wildlife Science, Online

INVITED PRESENTATIONS

August, 2018	‘Examining selection at reproductive genes in the Tasmanian devil.’ Oral Presentation: Save the Tasmanian Devil Team Meeting, Ross, Tasmania
September, 2018	‘Koala MHC: Perks of a PacBio Genome.’ Oral Presentation: PacBio User Group Meeting, UNSW, New South Wales
September, 2019	‘RONIN – Conservation Genomics in the Cloud.’ Oral Presentation: ResBaz Sydney 2019, UNSW, New South Wales
February, 2021	‘Top 10 RONIN features for researchers.’ Oral Presentation: Internet2 NET+ I2 Online Event
March, 2021	‘A beginner’s guide to genomics in the cloud with RONIN and AWS.’ Oral Presentation: AWS Research Webinar Series

AUTHORSHIP ATTRIBUTION STATEMENTS

This thesis is composed of six chapters, four of which include manuscripts that have been published in academic journals and have benefitted from the peer review process. Authorship attribution statements are provided below for the published papers that form Chapters 1-4 of this thesis. The candidate is the principal author for each of these papers and the inclusion of these published manuscripts follows the University of Sydney's Thesis and Examination of Higher Degrees by Research Policy 2015 guidelines for a thesis with publications. In the published manuscripts, table and figure numbering has been updated to reflect the chapter. Chapter 5 is formatted as a stand-alone manuscript but forms part of a larger consortium project that is currently ongoing, consequently this chapter is unpublished. Chapter 6 provides a summary of the research undertaken and highlights the main conclusions and implications of the thesis. Supplementary materials are provided after each relevant chapter. PDF versions of published chapters and any additional publications that were contributed to during the course of this degree are presented in the appendices.

Chapter 1 of this thesis includes the publication: Brandies, PA, Peel, E, Hogg, CJ & Belov, K 2019, 'The value of reference genomes in the conservation of threatened species', *Genes*, vol. 10, no. 11, pp. 846.

Katherine Belov and Carolyn J. Hogg conceived the project. Parice A. Brandies, the candidate, wrote the manuscript with assistance and revisions from Emma Peel, Carolyn J. Hogg and Katherine Belov.

Chapter 2 of this thesis is published as Brandies, PA & Hogg, CJ 2021, 'Ten simple rules for getting started with command-line bioinformatics', *PLOS Computational Biology*, vol. 17, no. 2, pp. e1008645.

Parice A. Brandies, the candidate, wrote the manuscript with Carolyn J. Hogg providing feedback and revisions.

Chapter 3 of this thesis is published as Brandies, PA, Wright, BR, Hogg, CJ, Grueber, CE & Belov, K 2020, 'Characterisation of reproductive gene diversity in the endangered Tasmanian devil', *Molecular Ecology Resources*, vol. 00, pp. 1-12.

Katherine Belov, Catherine E. Grueber and Carolyn J. Hogg contributed to the design of the study and sourced funding. Parice A. Brandies, the candidate, compiled the list of target genes, performed bioinformatic gene characterisation and SNP prediction and wrote the manuscript. Belinda Wright performed alignments of resequencing data to the reference genome and assisted Parice A. Brandies with the analysis of SNP genotypes. All authors revised the manuscript.

Chapter 4 of this thesis is published as Brandies, PA, Tang, S, Johnson, RSP, Hogg, CJ & Belov, K 2020, 'The first *Antechinus* reference genome provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies', *Gigabyte*, vol. 1, no. 7, pp. 1-22.

Katherine Belov, Carolyn J. Hogg and the candidate, Parice A. Brandies, conceived and designed the project. Carolyn J. Hogg and Parice A. Brandies collected the samples with assistance from Robert S.P. Johnson. Parice A. Brandies prepared the samples, created the reference genome, performed downstream analysis and wrote the manuscript. Simon Tang assisted with downstream analysis. All authors revised the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Parice Brandies 23/04/2021

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Kathy Belov 10/05/2021

ABSTRACT

In the midst of a global sixth mass extinction event, conservation initiatives are now more crucial than ever before. Australia houses the largest and most diverse range of marsupial species in the world; however, the number that are threatened (vulnerable, endangered, or critically endangered) is growing every year. Genetic management of threatened populations is vital in species recovery, yet incorporation of genetic data in conservation management is currently limited. International and national consortia such as the Earth BioGenomes Project, the Genome 10K project and the Oz Mammals Genomics initiative are currently producing reference genomes for a large variety of species across the phylogenetic tree, though there is currently a gap between the creation of these genomic resources and their downstream applications, particularly in conservation management of threatened species. One of the major drivers of this gap is due to the bioinformatic expertise and resources that are required to analyse large next generation sequencing datasets and to translate the findings into conservation contexts. This PhD employs a variety of bioinformatic and sequencing approaches to develop reference genomes and other genomic resources to answer key biological questions and provide direct management applications for the conservation of threatened Australian Marsupials. The value of genomic data for conservation is demonstrated for a range of species under varying scenarios including: i) using existing genomic datasets for the endangered Tasmanian devil to answer new conservation questions relating to reproduction, ii) creating a reference genome for a common species, the brown antechinus, to act as a model species for its threatened congeneric counterparts and iii) generating and uniting a suite of genomic resources to assist in the management of the vulnerable greater bilby. In addition, I provide ten simple rules for getting started with command-line bioinformatics to assist those wanting to utilise genomic data for wildlife conservation. Bridging the research-implementation gap is essential for harnessing the power of genomic resources for the conservation of threatened species. The findings from this PhD provide crucial steps into bridging this gap.

ABBREVIATIONS

10KP	10,000 plants project
AD	Alzheimer's disease
AD-ratio	average depth ratio
ANNOVAR	annotate variation
APT	advanced package tool
A_R	allelic richness
AWC	Australian Wildlife Conservancy
B10K	Bird 10K Project
Bat1K	Bat 1K Project
BED	browser extensible data
BLAST	basic local alignment search tool
bp	base pairs
BQSR	base quality score recalibration
BUSCO	benchmarking universal single-copy orthologs
BWA	Burrows-Wheeler aligner
CAFE	computational analysis of gene family evolution
CDS	coding domain sequence
CEGMA	core eukaryotic genes mapping approach
CGSG	Conservation Genetics Specialist Group
CI	confidence interval
CNV	copy number variant
ConGRESS	Conservation Genetic Resources for Effective Species Survival
CPU	central processing unit
DArTseq	diversity arrays technologies sequencing
ddRAD	double digest restriction-site associated DNA sequencing
DFTD	devil facial tumour disease
DNA	deoxyribonucleic acid
EBP	Earth Biogenome Project
F_{IS}	inbreeding coefficient
F_{ST}	pairwise fixation index
G10K	Genome 10K Project
GATK	genomic analysis toolkit
Gb	gigabase pairs
GEO BON	Group on Earth Observations Biodiversity Observation Network
GIGA	Global Invertebrate Genomics Alliance
GO	gene ontology
GWAS	genome wide association study
H_E	expected heterozygosity
HMW	high molecular weight
H_O	observed heterozygosity

HPC	high performance computer
I5K	Insect 5K project
indel	insertion or deletion
IT	information technology
IUCN	International Union for Conservation of Nature
kb	kilobase pairs
MAF	minor allele frequency
Mb	megabase pairs
MHC	major histocompatibility complex
miRNA	microRNA
NCBI	National Center for Biotechnology Information
NK	natural killer
OMG	Oz Mammal Genomics initiative
PCoA	principal coordinate analysis
PCR	polymerase chain reaction
PE	paired-end
QC	quality control
RADseq	restriction site-associated DNA sequencing
RAM	random access memory
RBH	reciprocal best hit
RNA	ribonucleic acid
ROH	runs of homozygosity
RRS	reduced representation sequencing
SE	standard error
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SSC	Species Survival Commission
STDP	Save the Tasmanian Devil Program
SV	structural variant
TLR	toll-like receptors
UTR3	3 prime untranslated region
UTR5	5 prime untranslated region
VCF	variant call format
VGP	Vertebrate Genomes Project
VM	virtual machine
WGR	whole genome resequencing
YUM	yellowdog updater, modified
ZAA	Zoo and Aquarium Association Australasia

CHAPTER 1

Thesis Introduction and Literature Review

THESIS INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION TO THESIS

In 2021, we entered the United Nations Decade on Restoration which aims to "prevent, halt and reverse the degradation of ecosystems on every continent and in every ocean" (United Nations Environment Programme, 2021). With more than 35,500 species now threatened with extinction and more plant and animal species being listed as endangered each year (IUCN, 2020), conservation initiatives are vital in preserving our Earth's biodiversity. Australia is one of seventeen "megadiverse" countries that comprises a large proportion of the Earth's biological diversity (Mittermeier, 1997) and houses a multitude of unique and endemic species (Chapman, 2009). Yet, Australia is known for having the worst record of mammal extinctions in the world (IUCN, 2020; Johnson & Isaac, 2009; Johnson, Isaac & Fisher, 2006; Short & Smith, 1994) and currently has more than 1,700 species (comprising both plants and animals) listed as threatened under the Australian Environment Protection and Biodiversity Conservation Act (EPBC Act) (Commonwealth of Australia, 1999). Effective management of threatened populations is imperative to species recovery and the conservation of Australia's unique biodiversity.

Incorporating genomic data into conservation efforts is key in ensuring populations have the best chance of long-term survival (Ballou et al., 2010; Frankham, Ballou & Briscoe, 2010; Frankham et al., 2017; Hohenlohe, Funk & Rajora, 2021). Recent advances in sequencing technology coupled with significant reductions in sequencing costs are facilitating the development of genomic resources for a diverse array of threatened species worldwide. In particular, numerous national and international consortia have recently been established with the aim to sequence a large proportion of the world's biodiversity (refer to section 1.2 for more detail). Conservation managers are eager to use genomic data as a tool to assist in the management of threatened species (Taylor, Dussex & van Heezik, 2017); however, the application of genomic data in species recovery is presently still lacking. A number of reviews have discussed this conservation genomics research-implementation gap and explored the potential barriers driving this gap (Britt et al., 2018; Galla et al., 2016; Hohenlohe, Funk & Rajora, 2021; Knight et al., 2008; Shafer et al., 2015; Taylor, Dussex & van Heezik, 2017). These studies highlighted that a lack of understanding of how genomic data may benefit species conservation, and a lack of clear examples showing how genomic data can be incorporated into threatened species management, are some of the major barriers.

Encouraging communication and collaboration between geneticists and industry is a crucial step in closing the gap between genomics and conservation (Britt et al., 2018; Galla et al., 2016; Knight et al., 2008; Taylor, Dussex & van Heezik, 2017). A great example of this is the Devil Tools & Tech program which amalgamates conservation research and management practice to help save Australia's endangered Tasmanian devil (Hogg et al., 2017b) (see section 1.2 for further details). However, for such projects to be successful, conservation managers must first understand the value of genomic data in species conservation, while researchers must identify what genomic resources are required to assist conservation efforts, generate the required resources, and provide clear explanations of how such resources can be implemented into conservation management (Hogg et al., 2017b). A major challenge that many researchers are facing with the greater availability of high-throughput sequencing technologies and the rise of the genomics era, is the bioinformatic expertise required to work with these large datasets (Marx, 2013). Without previous experience in bioinformatics or a strong background in IT (information technologies), conservation geneticists and other life scientists may not feel confident in tackling the creation and analysis of these genomic resources themselves, adding additional barriers that further widen the research-implementation gap.

This thesis aims to take crucial steps towards closing the research-implementation gap by equipping researchers and conservation managers with the knowledge and tools needed to employ genomic resources in species conservation efforts. The following literature review provides further context for this thesis and addresses one of the major drivers of the research-implementation gap by clearly demonstrating the value of reference genomes and associated genomic datasets in the conservation of threatened species.

1.2 THE VALUE OF REFERENCE GENOMES IN THE CONSERVATION OF THREATENED SPECIES

This section comprises the published review:

Brandies, PA, Peel, E, Hogg, CJ & Belov, K 2019, 'The value of reference genomes in the conservation of threatened species', *Genes*, vol. 10, no. 11, pp. 846.

Abstract

Conservation initiatives are now more crucial than ever – over a million plant and animal species are at risk of extinction over the coming decades. Genetic management of threatened species held in insurance programs is recommended, however few are taking advantage of the full range of genomic technologies available today. Less than 1% of the 13,505 species currently listed as threated by the IUCN (International Union for Conservation of Nature) have a published genome. While there has been much discussion in the literature about the importance of genomics for conservation, there are limited examples of how having a reference genome has changed conservation management practice. The Tasmanian devil (*Sarcophilus harrisii*), is an endangered Australian marsupial, threatened by an infectious clonal cancer devil facial tumour disease (DFTD). Populations have declined by 80% since the disease was first recorded in 1996. A reference genome for this species was published in 2012 and has been crucial for understanding DFTD and the management of the species in the wild. Here we use the Tasmanian devil as an example of how a reference genome has influenced management actions in the conservation of a species.

Introduction

We are currently in the midst of a global sixth mass extinction event with biodiversity rapidly declining around the world (Ceballos et al., 2015), and extinction rates are accelerating (IUCN, 2019). Australia has the worst mammal extinction rate of any country, with 25 mammals declared extinct since European settlement and almost 20% of current mammalian species listed as vulnerable (IUCN, 2019; Johnson & Isaac, 2009; Johnson, Isaac & Fisher, 2006; Short & Smith, 1994). This significant decline is concerning as Australia is one of seventeen “megadiverse” countries that comprises a large proportion of the Earth's biological diversity (Mittermeier, 1997). Megadiverse countries have at least 5,000 endemic plant species and have marine ecosystems within their borders (Mittermeier, 1997). In Australia, 87% of mammals, 93% of Australian reptiles and 94% of Australian frogs are endemic (Chapman, 2009). Therefore, conservation initiatives that protect and maintain Australia's biodiversity are now more crucial than ever.

Only 39% of the 1,890 Australian species (517 animals; 1,373 plants) listed as threatened under the Environment Protection and Biodiversity Conservation Act (EPBC Act) have a recovery plan in place to improve their threat status (Department of the Environment and Energy, 2019). These recovery plans set out management and research actions to slow population decline and promote recovery of threatened species and/communities. This is achieved by providing a framework for key interest groups and government agencies to coordinate their efforts to improve the plight of threatened species (Department of the Environment and Energy, 2019). Management actions range from mitigating threatening processes such as predation, habitat loss or change, in addition to research into basic species biology, ecosystem integration and genetics. The main goal of recovery plans is to maintain the long-term viability of a chosen population/community. Maintaining genetic diversity is an important component of population viability as it assists with mitigating negative effects associated with inbreeding and arms populations with the potential to adapt to future environmental change (Ballou et al., 2010; Frankham, Ballou & Briscoe, 2010; Lacy, 1997). As such, understanding a populations' inherent genetic diversity, in addition to their historical diversity and future potential, is of utmost importance in species conservation. For this reason, more than 80% of the current 200 Australian national vertebrate recovery plans have some form of genetic action listed in the species' recovery plan. Yet, less than 15% of these recovery plans have any form of genetic or genomic data available, either in existence or currently in development. Here we refer to genetic data as information based on specific, limited regions of the genome (e.g., targeted gene sequencing, microsatellite analysis, etc.), whilst genomic data is information based on the whole genome (e.g., whole genome sequencing/resequencing, whole-genome single nucleotide polymorphism (SNP) analysis/reduced representation sequencing etc.).

Advances in sequencing technologies and reduction in sequencing costs have given rise to the era of genomics, whereby holistic genome-wide approaches are rapidly replacing traditional genetic marker approaches in many non-model species (Allendorf, 2017; Johnson & Koepfli, 2014; Supple & Shapiro, 2018). Although recent reviews have highlighted the importance of implementing genomic data into conservation initiatives (Fuentes-Pardo & Ruzzante, 2017; Larsen & Matocq, 2019; Supple & Shapiro, 2018), the application of such powerful advances in sequencing technologies is lacking in the current literature. This limited use in conservation may be due to a number of reasons including: costs, a lack of understanding of the potential of new genomics approaches, lack of expertise in developing and utilising the data, and the absence of a reference genome for the species of interest (or a closely-related species) (Fuentes-Pardo & Ruzzante, 2017; McMahon, Teeling & Höglund, 2014; Supple & Shapiro, 2018). The latter is an important concern, as the generation of a reference genome requires considerable expertise, funds, computational resources and time that are not often accessible by wildlife managers and conservation teams (Fuentes-Pardo & Ruzzante, 2017; Khan et al., 2016).

Of the 13,505 animal species that are listed as threatened (Lower Risk/Conservation Dependent or worse) on the IUCN (International Union for Conservation of Nature) Red List (IUCN, 2019), 108 (< 1%) have published genomes on NCBI (Kitts et al., 2015). This equates to only 6% of the 1,842 animal genomes currently available on NCBI (Kitts et al., 2015). Creating high-quality reference genomes that can provide insights into species evolution and biology is a costly task (~$30,000 for an average eukaryotic genome size of 2.5 Gbp; Lewin et al., 2018), and also requires large collaborative groups to provide expertise from varying fields (e.g. Groenen et al., 2012; Johnson et al., 2018; Li et al., 2010). Fortunately, in recent years, a number of national and international consortia and genome projects have been formed with the aim of creating high-quality reference genomes for species spanning the phylogenetic tree of life including: the Earth Biogenome Project (EBP) (Lewin et al., 2018), the Genome 10K Project (G10K) (Genome 10K Community of Scientists, 2009; Koepfli et al., 2015), the Vertebrate Genomes Project (VGP) (Genome 10K Community of Scientists, 2017), the Bird 10K Project (B10K) (China National GeneBank, 2016), the Bat 1K Project (Bat1K) (Teeling et al., 2018), the Global Invertebrate Genomics Alliance (GIGA) (GIGA Community of Scientists, 2013; Voolstra, Wörheide & Lopez, 2017) and the Oz Mammal Genomics initiative (OMG) (Potter & Eldridge, 2017), to name a few. The goal of many of these consortia is to bring together the required expertise to generate reference genomes of sufficient quality that are publicly available to the science community, thereby providing the vital resources required to better implement genomics into conservation management (Fuentes-Pardo & Ruzzante, 2017; Khan et al., 2016; Supple & Shapiro, 2018). However, just providing the reference genomes, or genomic data, is not enough to improve conservation outcomes. Geneticists need to continually communicate how genomic techniques can be utilised in a cost-effective manner to better assist species conservation (McMahon, Teeling & Höglund, 2014; Ralls et al., 2018). As highlighted by Taylor, Dussex and van Heezik (2017), targeted education and training is also required to teach conservation managers how to interpret and utilise genomic data. To better assist conservation managers, a number of groups and communities have already been established to assist in providing conservation genetics advice for threated species management. These include the IUCN/SSC (Species Survival Commission) Conservation Genetics Specialist Group (CGSG), the Genetic Composition Working Group of GEO BON (Group on Earth Observations Biodiversity Observation Network) and the pan-European COST (Cooperation in Science and Technology) action ConGRESS (Conservation Genetic Resources for Effective Species Survival) (for further information and examples from these groups see Holderegger et al. 2019). Conservationists in their respective countries can get in touch with these groups to obtain the contact details of geneticists who work in their region who may be able to assist them with their management needs.

While a number of papers have reviewed current genomic techniques and the way they can, or have been, applied to assist in conservation decisions across species (Fuentes-Pardo & Ruzzante, 2017; McMahon, Teeling & Höglund, 2014), questions are still raised as to whether reference genomes are necessary for species conservation. Reference genomes hold the key to investigate a number of paradigms which are essential for species conservation including: demography, inbreeding, hybridisation, disease susceptibility, behavioural ecology and adaptation (Fuentes-Pardo & Ruzzante, 2017; Johnson & Koepfli, 2014; Khan et al., 2016; Larsen & Matocq, 2019; Supple & Shapiro, 2018). Here we demonstrate the value of a reference genome to the conservation effort of an endangered species, the Tasmanian devil (*Sarcophilus harrisii*), and how this information has been applied in real-time management practice (Hogg et al., 2017b).

The Tasmanian devil, an endangered Australian marsupial, is often used in the literature as an example for how genetics/genomics approaches can be used in conservation (Grueber, 2015; Johnson & Koepfli, 2014; Supple & Shapiro, 2018). However, something that is not often discussed is that having a reference genome for this species is one of the key factors that contributed to using genomics in management practice. Although this species has a unique conservation issue, low genetic diversity coupled with an infectious clonal cancer, the methods described herein apply to many other threatened species. Here we show how the reference genome has allowed a range of conservation questions to be answered in a timely, cost-effective manner, and enabled conservation researchers to adapt to the rapid advances in genomic technologies.

The Tasmanian devil and its genome

The Tasmanian devil is the largest extant carnivorous marsupial, native to mainland Tasmania, Australia. Emergence of a transmissible cancer, devil facial tumour disease (DFTD) in the mid-1990's has led to a rapid population decline of up to 80% across their range (Lazenby et al., 2018). In 2003, the Tasmanian and Australian governments responded to the disease threat by establishing the Save the Tasmanian Devil Program (STDP). Since then, researchers, wildlife managers and the zoo industry have worked closely with the STDP to ensure that Tasmanian devils have a sustainable ecological function in the Tasmanian ecosystem and landscape (Hogg et al., 2019a; Hogg et al., 2017b). This work has included a range of activities such as monitoring of wild populations, developing an insurance population, describing and characterising the disease, and developing new genomic tools to understand the disease, and the Tasmanian devil (Hogg et al., 2019a).

Prior to the publication of a reference genome for the Tasmanian devil, traditional genetic approaches such as MHC (major histocompatibility complex) typing and microsatellite analysis were used to explore genetic diversity at specific genes as well as general genetic diversity in the species (Cheng & Belov, 2012; Jones et al., 2003; Siddle et al., 2010). These techniques showed that the Tasmanian devil had low genetic diversity (Andrews et al., 2016; Cheng & Belov, 2012; Jones et al., 2003; Siddle et al., 2010). However, the low rates of polymorphism at most of these markers did not have high enough resolution to assist in answering crucial conservation questions such as determining founder relatedness within the insurance population

(Hogg et al., 2015; Hogg et al., 2019b), identifying high-resolution population substructure (Miller et al., 2011), or better understanding the origin and evolution of DFTD (Murchison et al., 2012). For these questions, further genomic data was required to improve resolution. For other threatened species, where there may be moderate to high genome-wide diversity, microsatellite markers may be highly polymorphic and so these markers have value as a continuing simple and cost-effective genetic management tool.

To overcome this knowledge gap in questions that could not be answered with previously employed methods, the Tasmanian devil genome was sequenced independently by two different research groups in 2011 (Miller et al., 2011; Murchison et al., 2012). Miller et al. (2011) sequenced the nuclear genome of two individuals (originating from extreme northwest and southeast Tasmania), as well as the tumour from one individual, using both Roche and Illumina sequencing platforms. Analysis of genome-wide SNPs confirmed low genetic diversity across the Tasmanian devil genome, as well as enabling the construction of genotyping arrays which revealed new population substructure, and identification of tumour-specific SNPs. However, the low contiguity of this reference genome assembly (148,891 scaffolds, scaffold N50 147 kb) limited the applicability of the data in downstream research. In 2012, a more contiguous, annotated nuclear genome (35,974 scaffolds, scaffold N50 1.85 Mb), and tumour genome, was published by Murchison et al. (2012), resulting in the primary reference genome used today. This higher quality assembly facilitated an enormous effort in downstream genetic and genomic research. It should be noted that as of August 2019, the 2012 Tasmanian devil reference genome paper (Murchison et al., 2012) has been cited over 200 times (Google Scholar Citation Search), highlighting the value of this reference genome to the research community. It is not possible to cover all of the research that has stemmed from the sequencing of the 2012 genome here. Rather, here we present key examples of how having a reference genome has contributed to conservation decisions and outcomes for the Tasmanian devil. We also note that at the time of this publication, an updated Tasmanian devil genome assembly has been released (Patton et al., 2019). This assembly utilised an in vitro proximity ligation technique to further improve scaffolding of the 2012 assembly (10,010 scaffolds, N50 7.75 Mb), however chromosome assignment and annotation has not been performed at this stage.

Conservation applications as a result of a reference genome

Basic Conservation Management

Microsatellite Analysis

Traditionally population genetic measures to answer basic questions regarding population structure, population size, population dynamics (migration, bottlenecks), kinship, inbreeding etc (Allendorf, 2017; Selkoe & Toonen, 2006) have used microsatellites, or short tandem repeats (Selkoe & Toonen, 2006). Where microsatellite markers have already been developed for the species of interest, or in a closely related species that may carry similar sequences, they provide a cost-effective, quick conservation management tool (Abdul-Muneer, 2014; Selkoe & Toonen, 2006). However, for those species where appropriate microsatellite markers are not currently available, or cross-species microsatellite amplification are not effective, and a reference genome is also not available, considerable time and resources are required to develop species-specific microsatellite markers. For example, prior to sequencing the Tasmanian devil genome, 11 putatively neutral microsatellite markers were developed to assess genetic diversity in Tasmanian devils (Jones et al., 2003). The development of these microsatellites involved creation and screening of a genomic library, sequencing of positive clones, primer design, and PCR optimisation (Jones et al., 2003). Several years later, MHC-linked microsatellite markers were developed in a similar manner as a cheaper and faster method of investigating MHC diversity when compared to traditional MHC typing techniques, such as cloning and sequencing particular MHC regions (Cheng & Belov, 2012). These traditional microsatellite isolation and marker development approaches require considerable laboratory expertise, time and funds (Abdul-Muneer, 2014), that today may be better spent developing more powerful molecular approaches (see Reduced Representation Sequencing section below).

Contrarily, the availability of the Tasmanian devil reference genome enabled 22 additional microsatellite markers to be identified and developed in a much faster, cost-effective manner using bioinformatic methods (Gooley et al., 2017). More importantly each of these microsatellites were known to be in non-coding regions across all of the autosomes, providing a greater representation of neutral genome-wide diversity in comparison to the original 11 putatively neural microsatellites. It has previously been estimated that development of just 10 microsatellite markers without prior genetic data can cost up to $10,000 (Abdelkrim et al., 2009). The availability of a reference genome mitigates the need for traditional microsatellite isolation procedures and therefore significantly reduces costs associated with marker development (<$1000 for primer optimisation and testing). The commercial development of microsatellite-based PCR kits resulted in further reductions to the time and cost associated with microsatellite marker development and use (Gooley et al., 2017). To date, 33 microsatellite markers have successfully been applied to Tasmanian devil conservation to investigate inbreeding (Gooley et al., 2017), reconstruct the pedigree of offspring born in group housing and on Maria Island (Farquharson, Hogg & Grueber, 2019; Gooley et al., 2017; Gooley et al., 2018; McLennan et al., 2018), and investigate mate choice within captivity and the wild (Day et al., 2019) (Table 1.1). These microsatellite markers have also successfully been applied to genotype individuals using non-invasive scat samples (Grueber et al., 2020) which are notoriously known for producing low quantities of low-quality DNA (Taberlet, Waits & Luikart, 1999). Globally, microsatellite markers continue to be an effective tool in conservation decision making by answering population questions (Armstrong et al., 2019; Faria et al., 2016; Grueber et al., 2019b; Shaney et al., 2016; Storfer et al., 2017). They are particularly valuable when using non-invasive samples that are often unsuitable for more complex genomic methods that require high-quality input DNA, such as reduced representation sequencing and other whole-genome sequencing methods (Fuentes-Pardo & Ruzzante, 2017). A reference genome allows for fast, easy and inexpensive development of such markers, improving their utility in the conservation management space.

Reduced Representation Sequencing

While microsatellite analysis is one of the most common population genetics tools, sometimes more statistical power is needed to address specific conservation management questions, particularly in species with low genetic diversity (Fernández et al., 2013; Hogg et al., 2015; Tokarska et al., 2009). For instance, in the Tasmanian devil, microsatellite analysis was unable to accurately estimate the relatedness of founders sourced for the insurance population between 2006 and 2008 (Hogg et al., 2015). Single nucleotide polymorphisms (SNPs) enable greater resolution for addressing some common conservation issues such as resolving parentage and population structure, understanding genetic diversity, and identifying regions of the genome which may be linked to important phenotypes (Andrews et al., 2016).

Table 1.1 Examples of Tasmanian devil conservation questions, actions and outcomes that have been facilitated by the reference genome.

Reference Genome Use	Conservation Questions Addressed	Conservation Actions	Conservation Outcomes
• Microsatellite development • Genome-wide SNP analysis	• Were the founders related? • Does the metapopulation have equal founder representation to ensure maintenance of gene diversity? • Is inbreeding accumulating in group housing and Maria island insurance populations?	• Resolved relatedness of founders (Hogg et al., 2015) • Resolved parentage in group housing within the metapopulation (Farquharson, Hogg & Grueber, 2019; Gooley et al., 2017; Gooley et al., 2018) • Reconstructed pedigree of island population (McLennan et al., 2018) • Informed translocation recommendations (Hogg et al., 2020)	• Tool for selecting individuals for translocations based on genetic complementation • Improved maintenance of genetic diversity across captive populations • Increased genetic diversity of hybrid individuals at wild release sites
• Characterisation of DFTD strains	• How many DFTD strains exist?	• Appropriate management of wild populations (Hogg et al., 2017a; Murchison et al., 2012; Pye et al., 2016b)	• Assisted in managing the spread of new DFTD strains
• Characterisation of immune genes • Primer design and SNP panel development • Targeted SNP analysis	• Can we develop a vaccine for DFTD? • Can we improve Tasmanian devil immune diversity?	• Immunisation development and deployment (Pye et al., 2018) • Immune gene diversity analysis for informed translocation recommendations (Cheng & Belov, 2014; Cheng et al., 2017; Cheng et al., 2019; Cui, Cheng & Belov, 2015; Grueber et al., 2019a; McLennan et al.,	• Improved immune responses of devils released to the wild • Improved immunogenetic diversity of released Tasmanian devils and their resultant offspring

		2020; Morris et al., 2015a; Morris et al., 2015b; Wright et al., 2015)	
• Development of blocking primer for metagenomics diet analysis	• What constitutes the complete diet of Tasmanian devils on Maria Island?	• Investigated the impact of an introduced carnivore to island wildlife (McLennan, *unpublished data*)	• Mitigation implemented to reduce impact on highly consumed species
• Alignment of resequenced genomes • SNP Analysis and Annotation • GWAS	• Are devils evolving host-parasite resistance to DFTD?	• Ongoing monitoring to ensure releases do not impact the evolution of potential resistance alleles (Epstein et al., 2016; Hohenlohe et al., 2019; Margres et al., 2018a; Wright et al., 2017)	• Assisted in understanding regions of the genome that are potentially involved in DFTD resistance

When compared to a microsatellite approach, only 3-8 biallelic SNPs are required to be as informative as one microsatellite marker (Rosenberg et al., 2003; Schopen et al., 2008). Reduced representation sequencing (RRS) is a simple, cost-effective approach for generating genome-wide SNP data and is gaining popularity in the conservation sector (Andrews et al., 2016; Fuentes-Pardo & Ruzzante, 2017; Wright et al., 2019). RRS relies on high-throughput sequencing of fragments generated by restriction enzyme digestion of the genome and can therefore easily be applied in any species. There are a variety of RRS methods currently available, including traditional RADseq (Davey & Blaxter, 2010), ddRAD (Peterson et al., 2012), DArTseq (Von Mark, Kilian & Dierig, 2013) and others (Andrews et al., 2016).

Both DArTseq and RADseq have been employed to collect RRS data from over 1,000 Tasmanian devils from the insurance population, Maria Island and a number of wild sites (Epstein et al., 2016; Hendricks et al., 2017; Margres et al., 2018a; McLennan et al., 2019; Wright et al., 2019). RRS methods have shown to be superior in accurately estimating diversity and inferring genome-wide heterozygosity compared with microsatellite analysis and other targeted techniques (McLennan et al., 2019). Although RRS does not require a reference genome for development and use, coupling RRS data with a reference genome is advantageous in that it: i) improves the reliability of genotype calls (Torkamaneh, Laroche & Belzile, 2016); ii) reduces the required coverage for accurate genotyping (Davey et al., 2011); iii) provides for a greater number of SNPs (Shafer et al., 2017); iv) improves downstream population genetic inferences (Shafer et al., 2017); v) allows for SNP annotation with gene information (Gurgul et al., 2019); and vi) provides the ability to compare results from differing RRS methods which is particularly important when different methods are used across time for endangered species.

Using a reference genome guided approach for the Tasmanian devil enabled 2060 SNPs to be identified (Wright et al., 2019) much more quickly than a *de novo* approach. Aligning the RRS data to the reference genome provides the ability to identify genes that may be targets of future analysis, and to separate functional vs non-functional genomic diversity, which can have conservation implications (Hoelzel, Bruford & Fleischer, 2019). For example, the reference genome was used to identify candidate genes within a genomic region that displayed signatures of selection in RRS data (Epstein et al., 2016), and to identify cancer-resistance candidate genes from phenotype association tests of RRS data (Margres et al., 2018a) (Table 1.1). A number of non-synonymous SNPs have also been identified within particular genes, which have the potential to impact phenotype. Furthermore, reference alignment allows SNPs from alternative RRS datasets to be compared and combined, such as the DArTseq and RADseq data, which is important for reusing previous investment of limited conservation dollars. Recent work investigating New Zealand threatened bird species also showed the benefits of calling SNPs against conordinal, confamilial, congeneric and conspecific reference genomes (Galla et al., 2019). This highlights that not every threatened species requires a reference genome, although the quality of the SNP data reduces as species relatedness moves away from the genus and family level.

Further Species-specific Applications

Reference Gene Characterisation

A valuable advantage of having access to a reference genome is the ability to characterise particular genes, or gene families, that are relevant to species-specific conservation (Johnson et al., 2018). Gene characterisation is often undertaken in two main ways: in-depth, manual characterisation of a specific set of genes of interest, and/or automatic, whole-genome annotation. The latter is achieved in two main stages: the computational phase and the annotation phase (Ekblom & Wolf, 2014; Yandell & Ence, 2012). During the computational phase, initial gene predictions are based on several lines of evidence including transcriptome and protein data from the species of interest and/or several closely-related, or well-annotated species (Ekblom & Wolf, 2014; Yandell & Ence, 2012). During the annotation phase, the most representative gene predictions (defined by the annotation pipeline) are synthesised into the final gene annotations (Ekblom & Wolf, 2014; Yandell & Ence, 2012). Whole-genome annotation of the Tasmanian devil reference genome was achieved using the Ensembl genome annotation pipeline (Curwen et al., 2004; Murchison et al., 2012; Potter et al., 2004). This automatic annotation of 18,775 protein-coding genes was critical to the development of targeted SNP panels to explore diversity at important immune genes in the Tasmanian devil (Morris et al., 2015a; Morris et al., 2015b; Wright et al., 2015) (see SNP Panel section below), and in the identification of genes that may be linked to DFTD (Epstein et al., 2016; Margres et al., 2018a; Margres et al., 2018b; Murchison et al., 2012; Wright et al., 2017) (Table 1.1).

While modern-day tools, such as trainable automated gene prediction algorithms, have increased the feasibility of genome annotation of newly sequenced species within individual research groups, complete genome annotation still requires considerable bioinformatic expertise (Ekblom & Wolf, 2014; Yandell & Ence, 2012). Manual annotation of a subset of target genes is often required. This is particularly relevant for genes that have undergone duplications and are therefore often unable to be automatically annotated (Ekblom & Wolf, 2014; Johnson et al., 2018). In the Tasmanian devil, duplication affected a number of gene families including the Major Histocompatibility Complex (MHC), toll-like receptors (TLR), natural killer (NK) receptors, cathelicidins, behaviour and reproductive genes which were all manually annotated (Cheng & Belov, 2014; Cui, Cheng & Belov, 2015; Morris et al., 2015a; Peel et al., 2016; van der Kraan et al., 2013). Annotation of these genes was essential in facilitating species-specific downstream research and informing conservation management decisions in the Tasmanian devil, such as: genetic variation analyses (Cheng & Belov, 2014; Cui, Cheng & Belov, 2015; Morris et al., 2015a; Morris et al., 2015b); selection of individuals for release to the wild (Hogg et al., 2020), individuals response to the immunotherapy (Pye et al., 2018); changes of immune function with the onset of puberty (Cheng et al., 2017); and the influence of age and DFTD on immune function (Cheng et al., 2019) (Table 1.1). This breadth of research highlights the potential of a reference genome for exploratory analysis of gene families involved in key biological processes of threatened species such as immunity, reproduction and behaviour.

Targeted SNP Panels

Targeted SNP panels enable diversity at particular genes to be investigated based on current conservation concerns/questions (van Tienderen et al., 2002). In the Tasmanian devil, a SNP panel targeting immune, behavioural and putatively neutral loci was developed and used to genotype over 300 individuals in the insurance population (Wright et al., 2015). This involved low-coverage resequencing of a number of individuals (see Whole-Genome Resequencing section below), alignment of data to the reference genome, identification of target SNPs, primer design, pilot sequencing and final genotyping. The SNP panel resolved parentage with higher confidence than microsatellite markers and provided representative measures of genetic diversity at both functional and non-functional loci (Wright et al., 2015). Development of another

SNP panel, which targeted a range of immune genes, showed considerably low immune diversity in the species (Morris et al., 2015b) which has led to further research into ways of breeding Tasmanian devils to improve genome-wide heterozygosity and functional diversity (Grueber et al., 2019a; McLennan et al., 2020). The Tasmanian devil reference genome was essential for aligning sequencing data and target SNP discovery, allowing management decisions to be based on both genome-wide and functional diversity (Table 1.1). Although custom SNP panel development can be expensive and is not simple, once developed it provides fast, accurate measures of diversity at particular genes, or genome regions, across a large number of individuals (Russell et al., 2019; Wright et al., 2015; Zhao et al., 2019).

Whole-genome Resequencing

Whole-genome resequencing (WGR) involves sequencing the genome of several individuals to a predetermined level of coverage (usually between 2× and 60×) and aligning this data to an available reference genome (for examples in non-model species, see Fuentes-Pardo & Ruzzante, 2017). A major application of whole-genome resequencing (WGR) is the identification of variation throughout the genome, enabling the development of more targeted approaches which can be used to explore diversity at key regions in a larger cohort of individuals (Morris et al., 2015b; Wright et al., 2015). The Tasmanian devil targeted SNP panels were created using low-coverage WGR (10-15×) data from 7-12 individuals aligned against the annotated reference genome (Morris et al., 2015b; Wright et al., 2017). A major limitation of using this low-coverage resequencing strategy is that genome regions with lower coverage can often contain sequencing errors that may not be distinguished from true SNPs (Li et al., 2009). This led to a number of the SNPs identified in the Tasmanian devil resequencing data not being present in the downstream SNP panel data (Morris et al., 2015b; Wright et al., 2017). While the best way to overcome this limitation is to increase the sequencing coverage of individuals, other methods, such as calling SNPs across individuals, can assist in more accurate variant calling in low-coverage WGR datasets (Cheng, Teo & Ong, 2014).

Higher-coverage sequence data enables variants and heterozygosity to be called much more accurately than low-coverage sequence data and hence allows for SNPs to be called more confidently without additional targeted sequencing (e.g., SNP panels) (Kishikawa et al., 2019). High-coverage (~45×) WGR of 25 Tasmanian devils has allowed for reliable estimates of genome-wide heterozygosity, which are being used to assess the accuracy of estimates from other techniques including microsatellites, SNP panels and RRS data. The higher cost of high-coverage data causes a trade-off between investigating the whole genome of a relatively small number of individuals versus using a target subset of loci across many individuals (as of 2019, WGR routinely costs over $1000 per individual whereas RRS costs less than $100 per individual). This trade-off needs to be acknowledged, is dependent on the conservation research questions, and requires careful consideration prior to the commencement of sequencing (Supple & Shapiro, 2018). Fortunately, a number of alternative cost-effective WGR approaches are available and may be suitable when high-coverage WGR is not possible. For a review of the different types of WGR and their differing applications in conservation, see Fuentes-Pardo & Ruzzante, 2017.

Whilst targeted sequencing approaches are useful for the exploration of genes known to be important to a species biology, sometimes genetic mechanisms driving particular phenomena that are vital to species adaptation and survival may not be known or detected in other reduced sequencing techniques like RRS (Hoban et al., 2016). Whole-genome resequencing (WGR) enables conservation researchers to ask and answer a wide range of questions that are not possible using other approaches. For example, WGR also enables the use of genome-wide association studies to determine the genetic basis of particular phenotypic traits that are important to species' conservation (Fuentes-Pardo & Ruzzante, 2017; Supple & Shapiro, 2018). In the case of the Tasmanian devil, some individuals have been found to display a resistant phenotype to DFTD, enabling spontaneous tumour regression (Pye et al., 2016a). Identifying the potential genetic basis of this phenotype is important to understanding which individuals may be more resilient to the disease and provide targets for the development of potential treatments (Epstein et al., 2016; Margres et al., 2018a; Wright et al., 2017) (Table 1.1). Low-coverage WGR of individuals showing tumour regression and those that succumbed to the disease enabled a genome-wide association study to be undertaken, which identified two genomic regions that may be associated with resistance to DFTD including PAX3 and TLL1 loci (Wright et al., 2017). A follow up study to Wright et al. (2017) re-sequenced 10 individuals to a higher coverage (20-30×) and was able to identify a larger number of genomic regions that may underlie tumour regression in the Tasmanian devil (Margres et al., 2018b). This work demonstrates the ability of WGR data, along with an annotated reference genome, in exploring the genetic basis of phenotypic traits that can have important conservation implications (Fuentes-Pardo & Ruzzante, 2017; Margres et al., 2018b; Supple & Shapiro, 2018; Wright et al., 2017) (Table 1.1). It is important to note that often larger numbers of individuals are required to identify genes underlying certain phenotypes, particularly in species with higher genetic diversity and/or reduced selective pressure on the phenotype of interest (Hong & Park, 2012). This requires careful consideration of trade-offs between the sequencing approach (targeted vs RRS vs WGR), number of samples and sequencing coverage, and will often depend upon some prior knowledge (or preliminary testing), budget, and access to samples. Overall, WGR data is better able to separate out and compare functional versus non-functional diversity than RRS methods (see Chapter 5), which is valuable in understanding the adaptive potential of species (Hoelzel, Bruford & Fleischer, 2019).

There are many other advantages of using this high-resolution genomic data, compared to RRS, including: i) more robust insights into the evolutionary and demographic histories of a species; ii) more accurate measures of diversity, inbreeding and population structure; and iii) the ability to identify and investigate signatures of selection and adaptive genetic variation (Fuentes-Pardo & Ruzzante, 2017; Khan et al., 2016; Larsen & Matocq, 2019). WGR data in the Tasmanian devil is currently being employed to assess selection and mutation rates within populations and in identifying runs of homozygosity (ROH) throughout the genome (for examples in other species see Ceballos, Hazelhurst & Ramsay, 2018; Hodgkinson et al., 2013). These analyses are useful in the investigation of well-known issues in conservation including inbreeding depression (Ceballos, Hazelhurst & Ramsay, 2018) and adaptation to captivity (Willoughby et al., 2017).

Some of the current limitations for using WGR in conservation contexts are the cost, the required compute power and respective expertise, and the availability of reference genomes (Fuentes-Pardo & Ruzzante, 2017; Supple & Shapiro, 2018). Costs vary greatly and depend on the number of individuals or loci you wish to use, and the required depth of sequencing (Fuentes-Pardo & Ruzzante, 2017). In addition, this approach requires significant expertise and compute power to execute, which limits its applicability to many conservation contexts (Fuentes-Pardo & Ruzzante, 2017). Creating partnerships between academic researchers with the required expertise and compute resources and conservation managers is key to overcoming many limitations of using genomics in conservation, and has been successfully implemented in conservation of the Tasmanian devil (Hogg et al., 2017b). A reference genome is essential for WGR, so the significant lack of published genomes (< 1%) for threatened species (or their closely-related counterparts) prevents many conservation managers from taking full advantage of high-resolution genomic data. However, in the dawn of large genomic consortia such as the Earth Biogenome Project, which aims to sequence the genomes of all of Earth's eukaryotic biodiversity over the next 10 years (Lewin et al., 2018), lack of a reference genome will soon become a thing of the past.

Overall, WGR paired with an annotated reference genome opens up a realm of possibilities for downstream conservation research by developing more cost-effective approaches when data from a large number of individuals is necessary for making informed conservation management decisions. As costs of sequencing continue to decrease, and the availability of reference genomes continue to rise, the use of this high-resolution genomic data in conservation research will likely become the norm (Johnson & Koepfli, 2014) and is already being applied to some bird species (Galla et al., 2019).

Reference Genome Quality

An important factor to consider in the creation of reference genomes is the quality of the assembly. Consortia such as the Vertebrate Genomes Project and the Earth Biogenome Project have proposed specific standards that reference genomes should meet (Genome 10K Community of Scientists, 2017; Lewin et al., 2018) (Table S1). However, it is important to understand whether such high standards are necessary or achievable for conservation management. A number of statistics are used to evaluate the different aspects of genome quality including accuracy (e.g., average read coverage and quality), continuity (e.g., N50, N90, number of contigs/scaffolds, average length of contigs/scaffolds, gap percentage etc), and completeness (e.g., BUSCO (Benchmarking sets of Universal Single-Copy Orthologs)/CEGMA (Core Eukaryotic Genes Mapping Approach) scores, number of genes etc) (see Wajid & Serpedin, 2014) for a more exhaustive list). While the ideal reference genome would consist of a completely annotated, gap free, chromosome-length assembly, even the some of the best model species genomes, such as the human genome, currently do not reach this standard. Furthermore, the ease and ability to reach chosen standards depends on many factors including genome size, genome structure (e.g. repetitive content), level of heterozygosity, sample availability/quantity, as well as the cost and expertise of the sequencing types and compute resources available (Koepfli et al., 2015) (for reviews on reference genome creation including available sequencing types and their associated advantages/disadvantages see Ekblom & Wolf, 2014; Sedlazeck et al., 2018; Wajid & Serpedin, 2014). It is important to note that the current Tasmanian devil reference genome, was sequenced in 2011 by Murchison et al. (2012), so does not meet the minimum standards set by the EBP (Earth Biogenome Project) or VGP (Vertebrate Genomes Project) (Table S1). Despite this, the Tasmanian devil genome has still been able to facilitate an enormous amount of conservation research. A higher-quality genome which is more complete, correct and contiguous, has a number of advantages such as: improved identification and characterisation of genes and other genomic regions; more accurate ROH (runs of homozygosity) analysis and structural variant analysis; and higher resolution of chromosomal organisation allowing for improved comparative genomic and evolutionary analyses (Lee et al., 2016).

Naturally, genome quality is also a factor of input DNA quality. High molecular weight DNA, generally greater than 40 kb in length, is required to generate the multiple sequencing types used to construct a high-quality genome (Rhoads & Au, 2015). Extracting high molecular weight DNA often requires additional consideration during the sample collection phase, such as flash-freezing tissues in liquid nitrogen, storage at -80˚C or below, and avoiding freeze-thaw. However, for species of high conservation concern, or those that inhabit difficult field locations, this can be challenging. In these scenarios, researchers may utilise museum specimens. However, this can introduce additional problems associated with sample preservation and degraded DNA, which may not be suited to long-read sequencing technologies (McDonough et al., 2018). As such, the ability to collect, store and extract high-quality DNA should not be underestimated, as this is an essential first step towards generating a high-quality genome. However, it is important to weigh-up whether the cost, compute resources, expertise and time of creating an improved or “Gold standard” assembly is necessary to answer the conservation research questions at hand. For example, Patton et al. (2019) showed that the improvement of contiguity of the newly released 2019 Tasmanian devil assembly had minimal impacts on inferred patterns of historical effective population size when compared to the current reference assembly. Hence, in many cases, a simple short-read genome assembly will be enough to answer many basic conservation management questions and also enable a number of more in-depth species-specific analyses mentioned in the sections above. Nevertheless, as sequencing technologies and computational infrastructure continue to advance and become more affordable, high-quality reference genomes will become easier to create and will overcome many of the limitations of current fragmented reference assemblies such as incomplete gene characterisation, comparative evolutionary limitations, and increased computational requirements (Lee et al., 2016). Despite this, without advances in sequencing chemistry and library preparation to reduce input DNA quality and quantity, the availability of high-quality samples and ensuing high molecular weight DNA may continue to limit the creation of high-quality reference genomes in some species.

Conclusions

The Tasmanian devil reference genome has enhanced our capacity to manage this species in the face of an infectious, clonal cancer. By having the reference genome we have been able to develop a range of genomic tools that have been used to investigate DFTD (e.g. Murchison et al., 2012), investigate the interplay between the Tasmanian devils and the disease (e.g. Epstein et al., 2016; Hohenlohe et al., 2019; Margres et al., 2018a; Wright et al., 2017), inform development of immunotherapy and vaccine protocols (Pye et al., 2018), inform the management of the insurance population (Hogg et al., 2017a; Hogg et al., 2019a) and provide advice on the translocation of Tasmanian devils to wild populations to improve both genome-wide and functional diversity (e.g. Hogg et al., 2020; McLennan et al., 2019). Tasmanian devils are not the only species threatened by disease, other examples include black-footed ferret and distemper (Thorne & Williams, 1988), bats and white-nose syndrome (Blehert et al., 2009), and frogs and chytrid (Berger, Speare & Hyatt, 1999). Here we have presented a strong case study of the benefits of using reference genomes for conservation of threatened species. As the threat to global biodiversity increases, the management of threatened species will become more pronounced. Reference genomes can be used by conservation managers to develop a range of genetic tools such as designing species-specific microsatellite markers for population data and differentiation; developing targeted SNP panels, or aligning and calling RRS data, for higher resolution population information or data on particular genes of interest; and/or conducting exploratory analyses (e.g., genome-wide association studies) using variant calling of whole-genome resequencing data.

Despite the challenges in obtaining high quality samples for genome sequencing, and expertise for the creation of reference genomes for threatened species, there is value in them. Reduced costs and lower input DNA requirements, as well as improved bioinformatic assembly and annotation pipelines based on non-model non-eutherian species, mean that these technologies are becoming more attainable by conservation programs and should be used more routinely where budgets allow (Ekblom & Wolf, 2014). Reference genomes enable a wealth of genetic/genomic applications and are an important asset in our ongoing fight to preserve global biodiversity. We would recommend that conservation managers who are seeking to use the types of methods we have described herein collaborate with global genome consortia (like the Earth Biogenome Project) or national/local consortia (like the Oz Mammal Genome Initiative) to utilise the full potential of genomic resources and join the genomics revolution. This allows conservation managers to focus on conservation and work with geneticists who can help them make adaptive management decisions in real time (Hogg et al., 2017b).

Although here we have presented a unique case study of a species with significantly low levels of genetic diversity and a large threatening disease process, the techniques described for the Tasmanian devil can be applied more broadly to many species of conservation concern. The applications of reference genomes in species conservation that have been described herein for devils are not unique to this species, as many of the questions we have answered are posed by those managing other threatened species. These include understanding historical demography and current population structure, minimising inbreeding, maximising adaptive potential, and identifying the basis of important phenotypic traits (whether these be related to disease, behaviour or reproduction). Hence, despite variation in threatening processes and status of vulnerable species, the nature of their small population sizes will result in a number of common conservation concerns that can be informed using genomic data (Fuentes-Pardo & Ruzzante, 2017; Khan et al., 2016). In the midst of the sixth mass extinction event, we advocate the use of reference genomes and associated genetic tools to arm conservation managers with ways to assist the long-term survival of species.

Acknowledgments

Thanks to all the staff, students and affiliates of the Australasian Wildlife Genomics Group who have contributed to discussions over the years in regards to the benefit of genomics data to conservation in particular R. Johnson, C. Grueber, B. Wright and Y. Cheng. Thanks also to the staff of the Save the Tasmanian Devil Program and the members of the Zoo and Aquarium Association Australasia without whom our devil work would not be possible.

1.3 THESIS AIMS AND OVERVIEW

The aim of this thesis is to assist researchers in bridging the conservation genomics research-implementation gap by providing a cohesive summary of the bioinformatic tools and approaches that can be used to assist in the conservation of threatened species. With a focus on threatened Australian marsupials, specifically, this thesis aims to:

1. Introduce the importance of reference genomes and genomic data as a tool for conservation management of threatened species (Chapter 1)
2. Provide researchers with a simple guide for getting started with command-line bioinformatics to facilitate the analysis of genomic data for species conservation (Chapter 2)
3. Provide clear examples that demonstrate how reference genomes, alongside other common genomic datasets, can be used to provide answers to key questions and assist in the conservation of threatened species across a range of different scenarios including:
    a. Utilising pre-existing reference genomes and genomic datasets for a well-studied endangered species to answer novel species-specific questions with conservation implications (Chapter 3)
    b. Creating a reference genome for a non-threatened species to act as a resource for threatened counterparts, and as a model to explore unique biological traits that can provide valuable insights with potentially broad conservation implications (Chapter 4)
    c. Generating a reference genome for a threatened species with limited pre-existing genetic data to answer a suite of crucial conservation questions and provide valuable resources for ongoing national population genetic management of the species (Chapter 5)
4. Employ a range of sequencing technologies and bioinformatic approaches in novel contexts to showcase a variety of ways that common genomic data types can be utilised to inform species conservation (Chapters 3-5)

This thesis addresses the above aims as follows:

- In Chapter 1, I provide a literature review that introduces the context and background of this thesis and demonstrates the importance of reference genomes as a tool for conservation management using the endangered Tasmanian devil as a model.
- In Chapter 2, I provide a general guide for getting started with command-line bioinformatics using the “Ten Simple Rules” framework. This chapter is fundamental to the usefulness of this thesis, as it provides the necessary background for other researchers wanting to employ the bioinformatic methods described in the following chapters.
- In Chapter 3, I extend the work presented in Chapter 2 by utilising the Tasmanian devil reference genome and pre-existing whole genome resequencing datasets to explore reproductive gene diversity using a targeted gene approach. I characterise diversity at 219 genes and identify 19 genes with variation that may have functional consequences on Tasmanian devil reproduction.
- In Chapter 4, I create a reference genome for the brown antechinus, a common Australian marsupial, to act as a resource for population genetic monitoring of other threatened antechinus species, and to facilitate its use as a model to examine the genetic interplay between stress, immunity and reproduction in marsupials.
- In Chapter 5, I generate a high-quality reference genome for the Greater Bilby and utilise the reference genome, along with two other common genomic data types, whole genome resequencing and reduced representation sequencing, to answer a number of key conservation questions and generate a suite of tools that will assist in long term monitoring and management of the national bilby metapopulation.
- In Chapter 6, I present the general discussion and conclusions for the thesis where I highlight the implications of this body of work and provide suggestions for future directions.

Together, this thesis arms researchers with the necessary background knowledge to harness the power of genomics for the conservation of threatened species and help close the research-implementation gap.

1.4 REFERENCES

Abdelkrim, J, Robertson, BC, Stanton, JL & Gemmell, NJ 2009, 'Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing', *BioTechniques,* vol. 46, no. 3, pp. 185-192.

Abdul-Muneer, PM 2014, 'Application of microsatellite markers in conservation genetics and fisheries management: Recent advances in population structure analysis and conservation strategies', *Genetics Research International,* vol. 2014, pp. 1-11.

Allendorf, FW 2017, 'Genetics and the conservation of natural populations: allozymes to genomes', *Molecular Ecology,* vol. 26, no. 2, pp. 420-430.

Andrews, KR, Good, JM, Miller, MR, Luikart, G & Hohenlohe, PA 2016, 'Harnessing the power of RADseq for ecological and evolutionary genomics', *Nature Reviews Genetics,* vol. 17, no. 2, pp. 81-92.

Armstrong, AJ, Dudgeon, CL, Bustamante, C, Bennett, MB & Ovenden, JR 2019, 'Development and characterization of 17 polymorphic microsatellite markers for the reef manta ray (*Mobula alfredi*)', *BMC Research Notes,* vol. 12, no. 233, pp. 1-5.

Ballou, JD, Lees, C, Faust, LJ, Long, S, Lynch, C, Bingaman Lackey, L & Foose, TJ 2010, 'Demographic and genetic management of captive populations', *Wild Mammals in Captivity: Principles and Techniques for Zoo Management,* 2nd edn, The University of Chicago Press, Chicago, IL, USA.

Berger, L, Speare, R & Hyatt, A 1999, 'Chytrid fungi and amphibian declines: overview, implications and future directions', in Environment Australia (ed.), *Declines and disappearances of australian frogs*, Campbell, A, Canberra, Australia.

Blehert, DS, Hicks, AC, Behr, M, Meteyer, CU, Berlowski-Zier, BM, Buckles, EL, Coleman, JT, Darling, SR, Gargas, A & Niver, R 2009, 'Bat white-nose syndrome: an emerging fungal pathogen?', *Science,* vol. 323, no. 5911, pp. 227-227.

Britt, M, Haworth, SE, Johnson, JB, Martchenko, D & Shafer, AB 2018, 'The importance of non-academic coauthors in bridging the conservation genetics gap', *Biological Conservation,* vol. 218, pp. 118-123.

Ceballos, FC, Hazelhurst, S & Ramsay, M 2018, 'Assessing runs of homozygosity: A comparison of SNP array and whole genome sequence low coverage data', *BMC Genomics,* vol. 19, no. 106, pp. 1-12.

Ceballos, G, Ehrlich, PR, Barnosky, AD, García, A, Pringle, RM & Palmer, TM 2015, 'Accelerated modern human–induced species losses: Entering the sixth mass extinction', *Science Advances,* vol. 1, no. 5, pp. e1400253.

Chapman, AD 2009, *Numbers of living species in Australia and the world*, 2 edn, Australian Government, Department of the Environment and Energy, Canberra, Australia.

Cheng, AY, Teo, Y & Ong, RT 2014, 'Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals', *Bioinformatics,* vol. 30, no. 12, pp. 1707-1713.

Cheng, Y & Belov, K 2012, 'Isolation and characterisation of 11 MHC-linked microsatellite loci in the Tasmanian devil (*Sarcophilus harrisii*)', *Conservation Genetics Resources,* vol. 4, no. 2, pp. 463-465.

Cheng, Y & Belov, K 2014, 'Characterisation of non-classical MHC class I genes in the Tasmanian devil (*Sarcophilus harrisii*)', *Immunogenetics,* vol. 66, no. 12, pp. 727-735.

Cheng, Y, Heasman, K, Peck, S, Peel, E, Gooley, RM, Papenfuss, AT, Hogg, CJ & Belov, K 2017, 'Significant decline in anticancer immune capacity during puberty in the Tasmanian devil', *Scientific Reports,* vol. 7, no. 44716, pp. 1-7.

Cheng, Y, Makara, M, Peel, E, Fox, S, Papenfuss, AT & Belov, K 2019, 'Tasmanian devils with contagious cancer exhibit a constricted T-cell repertoire diversity', *Communications Biology,* vol. 2, no. 99, pp. 1-9.

China National Genebank 2016, *B10K*, viewed 16 August 2019, https://b10k.genomics.cn/

Commonwealth of Australia 1999, *Environmental Protection and Biodiversity Conservation Act*, Canberra, Australia.

Cui, J, Cheng, Y & Belov, K 2015, 'Diversity in the Toll-like receptor genes of the Tasmanian devil (*Sarcophilus harrisii*)', *Immunogenetics,* vol. 67, no. 3, pp. 195-201.

Curwen, V, Eyras, E, Andrews, TD, Clarke, L, Mongin, E, Searle, SM & Clamp, M 2004, 'The Ensembl automatic gene annotation system', *Genome Research,* vol. 14, no. 5, pp. 942-950.

Davey, JW & Blaxter, ML 2010, 'RADSeq: next-generation population genetics', *Briefings in Functional Genomics,* vol. 9, no. 5-6, pp. 416-423.

Davey, JW, Hohenlohe, PA, Etter, PD, Boone, JQ, Catchen, JM & Blaxter, ML 2011, 'Genome-wide genetic marker discovery and genotyping using next-generation sequencing', *Nature Reviews Genetics,* vol. 12, no. 7, pp. 499-510.

Day, J, Gooley, RM, Hogg, CJ, Belov, K, Whittington, CM & Grueber, CE 2019, 'MHC-associated mate choice under competitive conditions in captive versus wild Tasmanian devils', *Behavioral Ecology,* vol. 30, pp. 1196-1204.

Department of the Environment and Energy 2019, *Recovery plans*, viewed 8 August 2019, https://www.environment.gov.au/biodiversity/threatened/recovery-plans

Ekblom, R & Wolf, JB 2014, 'A field guide to whole-genome sequencing, assembly and annotation', *Evolutionary Applications,* vol. 7, no. 9, pp. 1026-1042.

Epstein, B, Jones, M, Hamede, R, Hendricks, S, McCallum, H, Murchison, EP, Schönfeld, B, Wiench, C, Hohenlohe, P & Storfer, A 2016, 'Rapid evolutionary response to a transmissible cancer in Tasmanian devils', *Nature Communications,* vol. 7, no. 12684, pp. 1-7.

Faria, J, Pita, A, Rivas, M, Martins, GM, Hawkins, SJ, Ribeiro, P, Neto, AI & Presa, P 2016, 'A multiplex microsatellite tool for conservation genetics of the endemic limpet *Patella candei* in the Macaronesian archipelagos', *Aquatic Conservation: Marine and Freshwater Ecosystems,* vol. 26, no. 4, pp. 775-781.

Farquharson, KA, Hogg, CJ & Grueber, CE 2019, 'A case for genetic parentage assignment in captive group housing', *Conservation Genetics,* vol. 20, no. 5, pp. 1-7.

Fernández, ME, Goszczynski, DE, Lirón, JP, Villegas-Castagnasso, EE, Carino, MH, Ripoli, MV, Rogberg-Muñoz, A, Posik, DM, Peral-García, P & Giovambattista, G 2013, 'Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd', *Genetics and Molecular Biology,* vol. 36, no. 2, pp. 185-191.

Frankham, R, Ballou, JD & Briscoe, DA 2010, *Introduction to Conservation Genetics*, 2 edn, Cambridge University Press, Cambridge, UK.

Frankham, R, Ballou, JD, Ralls, K, Eldridge, M, Dudash, MR, Fenster, CB, Lacy, RC & Sunnucks, P 2017, *Genetic Management of Fragmented Animal and Plant Populations*, Oxford University Press, New York, USA.

Fuentes-Pardo, AP & Ruzzante, DE 2017, 'Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations', *Molecular Ecology,* vol. 26, no. 20, pp. 5369-5406.

Galla, SJ, Buckley, TR, Elshire, R, Hale, ML, Knapp, M, McCallum, J, Moraga, R, Santure, AW, Wilcox, P & Steeves, TE 2016, 'Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances', *Molecular Ecology,* vol. 25, no. 21, pp. 5267-5281.

Galla, SJ, Forsdick, NJ, Brown, L, Hoeppner, M, Knapp, M, Maloney, RF, Moraga, R, Santure, AW & Steeves, TE 2019, 'Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management', *Genes,* vol. 10, no. 9, pp. 1-19.

Genome 10k Community of Scientists 2009, 'Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species', *Journal of Heredity,* vol. 100, no. 6, pp. 659-674.

Genome 10k Community of Scientists 2017, *Vertebrate Genomes Project*, viewed 16 August 2019, https://vertebrategenomesproject.org

Giga Community of Scientists 2013, 'The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes', *Journal of Heredity,* vol. 105, no. 1, pp. 1-18.

Gooley, R, Hogg, CJ, Belov, K & Grueber, CE 2017, 'No evidence of inbreeding depression in a Tasmanian devil insurance population despite significant variation in inbreeding', *Scientific Reports,* vol. 7, no. 1830, pp. 1-11.

Gooley, RM, Hogg, CJ, Belov, K & Grueber, CE 2018, 'The effects of group versus intensive housing on the retention of genetic diversity in insurance populations', *BMC Zoology,* vol. 3, no. 2, pp. 1-12.

Groenen, MA, Archibald, AL, Uenishi, H, Tuggle, CK, Takeuchi, Y, Rothschild, MF, Rogel-Gaillard, C, Park, C, Milan, D & Megens, H-J 2012, 'Analyses of pig genomes provide insight into porcine demography and evolution', *Nature,* vol. 491, no. 7424, pp. 393-398.

Grueber, C, Peel, E, Wright, B, Hogg, C & Belov, K 2019a, 'A Tasmanian devil breeding program to support wild recovery', *Reproduction Fertility and Development,* vol. 31, no. 7, pp. 1296-1304.

Grueber, CE 2015, 'Comparative genomics for biodiversity conservation', *Computational and Structural Biotechnology Journal,* vol. 13, pp. 370-375.

Grueber, CE, Chong, R, Gooley, RM, McLennan, EA, Barrs, VR, Belov, K & Hogg, CJ 2020, 'Genetic analysis of scat samples to inform conservation of the Tasmanian devil', *Australian Zoologist,* vol. 40, no. 3, pp. 492-504.

Grueber, CE, Fox, S, McLennan, EA, Gooley, RM, Weiser, EL, Pemberton, D, Hogg, CJ & Belov, K 2019b, 'Complex problems need detailed solutions: Harnessing multiple data types to inform genetic rescue in the wild', *Evolutionary Applications,* vol. 12, pp. 280-291.

Gurgul, A, Miksza-Cybulska, A, Szmatoła, T, Jasielczuk, I, Piestrzyńska-Kajtoch, A, Fornal, A, Semik-Gurgul, E & Bugno-Poniewierska, M 2019, 'Genotyping-by-sequencing performance in selected livestock species', *Genomics,* vol. 111, no. 2, pp. 186-195.

Hendricks, S, Epstein, B, Schönfeld, B, Wiench, C, Hamede, R, Jones, M, Storfer, A & Hohenlohe, P 2017, 'Conservation implications of limited genetic diversity and population structure in Tasmanian devils (*Sarcophilus harrisii*)', *Conservation Genetics,* vol. 18, no. 4, pp. 977-982.

Hoban, S, Kelley, JL, Lotterhos, KE, Antolin, MF, Bradburd, G, Lowry, DB, Poss, ML, Reed, LK, Storfer, A & Whitlock, MC 2016, 'Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions', *The American Naturalist,* vol. 188, no. 4, pp. 379-397.

Hodgkinson, A, Casals, F, Idaghdour, Y, Grenier, J-C, Hernandez, RD & Awadalla, P 2013, 'Selective constraint, background selection, and mutation accumulation variability within and between human populations', *BMC Genomics,* vol. 14, no. 495, pp. 1-10.

Hoelzel, AR, Bruford, MW & Fleischer, RC 2019, 'Conservation of adaptive potential and functional diversity', *Conservation Genetics,* vol. 20, pp. 1-5.

Hogg, C, Lee, A, Srb, C & Hibbard, C 2017a, 'Metapopulation management of an endangered species with limited genetic diversity in the presence of disease: the Tasmanian devil Sarcophilus harrisii', *International Zoo Yearbook,* vol. 51, no. 1, pp. 137-153.

Hogg, CJ, Fox, S, Pemberton, D & Belov, K 2019a, *Saving the Tasmanian Devil*, CSIRO Publishing, Clayton South, VIC, Australia.

Hogg, CJ, Grueber, CE, Pemberton, D, Fox, S, Lee, AV, Ivy, JA & Belov, K 2017b, '“Devil Tools & Tech”: a synergy of conservation research and management practice', *Conservation Letters,* vol. 10, no. 1, pp. 133-138.

Hogg, CJ, Ivy, JA, Srb, C, Hockley, J, Lees, C, Hibbard, C & Jones, M 2015, 'Influence of genetic provenance and birth origin on productivity of the Tasmanian devil insurance population', *Conservation Genetics,* vol. 16, no. 6, pp. 1465-1473.

Hogg, CJ, McLennan, EA, Wise, P, Lee, A, Pemberton, D, Fox, S, Belov, K & Grueber, CE 2020, 'Preserving the integrity of a single source population during multiple translocations', *Biological Conservation,* vol. 241, no. 108318, pp. 1-7.

Hogg, CJ, Wright, B, Morris, KM, Lee, AV, Ivy, JA, Grueber, CE & Belov, K 2019b, 'Founder relationships and conservation management: empirical kinships reveal the effect on breeding programmes when founders are assumed to be unrelated', *Animal Conservation,* vol. 22, no. 4, pp. 348-361.

Hohenlohe, PA, Funk, WC & Rajora, OP 2021, 'Population genomics for wildlife conservation and management', *Molecular Ecology,* vol. 30, no. 1, pp. 62-82.

Hohenlohe, PA, McCallum, HI, Jones, ME, Lawrance, MF, Hamede, RK & Storfer, A 2019, 'Conserving adaptive potential: lessons from Tasmanian devils and their transmissible cancer', *Conservation Genetics,* vol. 20, no. 1, pp. 81-87.

Holderegger, R, Balkenhol, N, Bolliger, J, Engler, JO, Gugerli, F, Hochkirch, A, Nowak, C, Segelbacher, G, Widmer, A & Zachos, FE 2019, 'Conservation genetics: Linking science with practice', *Molecular Ecology,* vol. 28, no. 17, pp. 3848-3856.

Hong, EP & Park, JW 2012, 'Sample size and statistical power calculation in genetic association studies', *Genomics & informatics,* vol. 10, no. 2, pp. 117-122.

IUCN 2019, *The IUCN Red List of Threatened Species. Version 2019-3*, viewed 2 July 2019, http://www.iucnredlist.org

IUCN 2020, *The IUCN Red List of Threatened Species. Version 2020-3*, viewed 16 March 2021, http://www.iucnredlist.org

Johnson, CN & Isaac, JL 2009, 'Body mass and extinction risk in Australian marsupials: the ‘Critical Weight Range’revisited', *Austral Ecology,* vol. 34, no. 1, pp. 35-40.

Johnson, CN, Isaac, JL & Fisher, DO 2006, 'Rarity of a top predator triggers continent-wide collapse of mammal prey: dingoes and marsupials in Australia', *Proceedings of the Royal Society Biological Sciences Series B,* vol. 274, no. 1608, pp. 341-346.

Johnson, RN, O'Meally, D, Chen, Z, Etherington, GJ, Ho, SYW, Nash, WJ, Grueber, CE, Cheng, Y, Whittington, CM, Dennison, S, Peel, E, Haerty, W, O'Neill, RJ, Colgan, D, Russell, TL, Alquezar-Planas, DE, Attenbrow, V, Bragg, JG, Brandies, PA, Chong, AY-Y, Deakin, JE, Di Palma, F, Duda, Z, Eldridge, MDB, Ewart, KM, Hogg, CJ, Frankham, GJ, Georges, A, Gillett, AK, Govendir, M, Greenwood, AD, Hayakawa, T, Helgen, KM, Hobbs, M, Holleley, CE, Heider, TN, Jones, EA, King, A, Madden, D, Graves, JaM, Morris, KM, Neaves, LE, Patel, HR, Polkinghorne, A, Renfree, MB, Robin, C, Salinas, R, Tsangaras, K, Waters, PD, Waters, SA, Wright, B, Wilkins, MR, Timms, P & Belov, K 2018, 'Adaptation and conservation insights from the koala genome', *Nature Genetics,* vol. 50, no. 8, pp. 1102-1111.

Johnson, WE & Koepfli, K 2014, 'The role of genomics in conservation and reproductive sciences', *Reproductive Sciences in Animal Conservation*, Springer, Berlin, Germany.

Jones, ME, Paetkau, D, Geffen, E & Moritz, C 2003, 'Microsatellites for the Tasmanian devil (*Sarcophilus laniarius*)', *Molecular Ecology Notes,* vol. 3, no. 2, pp. 277-279.

Khan, S, Nabi, G, Ullah, MW, Yousaf, M, Manan, S, Siddique, R & Hou, H 2016, 'Overview on the role of advance genomics in conservation biology of endangered species', *International Journal of Genomics,* vol. 2016, pp. 1-8.

Kishikawa, T, Momozawa, Y, Ozeki, T, Mushiroda, T, Inohara, H, Kamatani, Y, Kubo, M & Okada, Y 2019, 'Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data', *Scientific Reports,* vol. 9, no. 1784, pp. 1-10.

Kitts, PA, Church, DM, Thibaud-Nissen, F, Choi, J, Hem, V, Sapojnikov, V, Smith, RG, Tatusova, T, Xiang, C & Zherikov, A 2015, 'Assembly: a resource for assembled genomes at NCBI', *Nucleic Acids Research,* vol. 44, no. D1, pp. D73-D80.

Knight, AT, Cowling, RM, Rouget, M, Balmford, A, Lombard, AT & Campbell, BM 2008, 'Knowing but not doing: selecting priority conservation areas and the research–implementation gap', *Conservation Biology,* vol. 22, no. 3, pp. 610-617.

Koepfli, K-P, Paten, B, Genome 10k Community of Scientists & O'Brien, SJ 2015, 'The Genome 10K Project: a way forward', *Annual Review of Animal Biosciences,* vol. 3, no. 1, pp. 57-111.

Lacy, RC 1997, 'Importance of genetic variation to the viability of mammalian populations', *Journal of Mammalogy,* vol. 78, no. 2, pp. 320-335.

Larsen, PA & Matocq, MD 2019, 'Emerging genomic applications in mammalian ecology, evolution, and conservation', *Journal of Mammalogy,* vol. 100, no. 3, pp. 786-801.

Lazenby, BT, Tobler, MW, Brown, WE, Hawkins, CE, Hocking, GJ, Hume, F, Huxtable, S, Iles, P, Jones, ME & Lawrence, C 2018, 'Density trends and demographic signals uncover the long-term impact of transmissible cancer in Tasmanian devils', *Journal of Applied Ecology,* vol. 55, no. 3, pp. 1368-1379.

Lee, H, Gurtowski, J, Yoo, S, Nattestad, M, Marcus, S, Goodwin, S, McCombie, WR & Schatz, M 2016, 'Third-generation sequencing and the future of genomics', *BioRxiv*, no. 048603.

Lewin, HA, Robinson, GE, Kress, WJ, Baker, WJ, Coddington, J, Crandall, KA, Durbin, R, Edwards, SV, Forest, F & Gilbert, MTP 2018, 'Earth BioGenome Project:

Sequencing life for the future of life', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 115, no. 17, pp. 4325-4333.

Li, Q, Xuan, Z, Li, Y, Zheng, H, Bai, Y, Li, B, Hu, Y, Liu, X, Zhang, Z, Li, D, Sun, X, Wang, J, Fan, W, Xu, L, Tian, F, Guo, X, Lin, S, Jiang, Z, Zhang, X, Cao, J, Bolund, L, Ye, C, Ren, Y, Ruan, J, Gong, T, Hu, W, Wang, X, Cook, K, Ni, P, Wang, J, Wang, H, Zhou, Y, Li, J, Lu, Z, Zheng, Y, Huang, Q, Li, Z, Zhang, G, Vinar, T, Wang, J, Wang, J, An, N, Xu, A, Olson, M, Wu, Q, Qian, W, Shen, F, Zhang, H, Cheng, S, Xie, X, Mu, B, Wang, W, Liu, S, Tian, J, Shi, Y, Fu, Y, Bruford, MW, Yiu, S-M, Zhang, D, Lin, R, Tian, G, Zhao, S, Liu, Q, Zhang, J, Lam, T-W, Li, G, Steiner, CC, Wang, B, Zheng, H, Guo, Y, Shan, G, Zhu, H, Hou, R, Li, J, Zhao, J, Liang, H, Shi, Z, Li, H, Zhang, Q, Yang, G, Liu, H, Nie, W, Ma, L, Wen, M, Wang, X, Qu, N, Yang, Z, Ren, X, Kosiol, C, Huang, Y, Yu, C, Lam, TT-Y, Wei, F, Liu, B, Wang, M, He, L, Nielsen, R, Fang, X, Ryder, OA & Li, D 2010, 'The sequence and *de novo* assembly of the giant panda genome', *Nature,* vol. 463, no. 7279, pp. 311-317.

Li, R, Li, Y, Fang, X, Yang, H, Wang, J, Kristiansen, K & Wang, J 2009, 'SNP detection for massively parallel whole-genome resequencing', *Genome Research,* vol. 19, no. 6, pp. 1124-1132.

Margres, MJ, Jones, ME, Epstein, B, Kerlin, DH, Comte, S, Fox, S, Fraik, AK, Hendricks, SA, Huxtable, S & Lachish, S 2018a, 'Large-effect loci affect survival in Tasmanian devils (*Sarcophilus harrisii*) infected with a transmissible cancer', *Molecular Ecology,* vol. 27, no. 21, pp. 4189-4199.

Margres, MJ, Ruiz-Aravena, M, Hamede, R, Jones, ME, Lawrance, MF, Hendricks, SA, Patton, A, Davis, BW, Ostrander, EA & McCallum, H 2018b, 'The genomic basis of tumor regression in Tasmanian devils (*Sarcophilus harrisii*)', *Genome Biology and Evolution,* vol. 10, no. 11, pp. 3012-3025.

Marx, V 2013, 'The big challenges of big data', *Nature,* vol. 498, no. 7453, pp. 255-260.

McDonough, MM, Parker, LD, Rotzel McInerney, N, Campana, MG & Maldonado, JE 2018, 'Performance of commonly requested destructive museum samples for mammalian genomic studies', *Journal of Mammalogy,* vol. 99, no. 4, pp. 789-802.

McLennan, EA, Gooley, RM, Wise, P, Belov, K, Hogg, CJ & Grueber, CE 2018, 'Pedigree reconstruction using molecular data reveals an early warning sign of gene diversity loss in an island population of Tasmanian devils (*Sarcophilus harrisii*)', *Conservation Genetics,* vol. 19, no. 2, pp. 439-450.

McLennan, EA, Grueber, CE, Wise, P, Belov, K & Hogg, CJ 2020, 'Mixing genetic lineages sucessfully boosts diversity of an endangered carnivore.', *Animal Conservation,* vol. 23, pp. 700-712.

McLennan, EA, Wright, BR, Belov, K, Hogg, CJ & Grueber, CE 2019, 'Too much of a good thing? Finding the most informative genetic data set to answer conservation questions', *Molecular Ecology Resources,* vol. 19, no. 3, pp. 659-671.

McMahon, BJ, Teeling, EC & Höglund, J 2014, 'How and why should we implement genomics into conservation?', *Evolutionary Applications,* vol. 7, no. 9, pp. 999-1007.

Miller, W, Hayes, VM, Ratan, A, Petersen, DC, Wittekindt, NE, Miller, J, Walenz, B, Knight, J, Qi, J, Zhao, F, Wang, Q, Bedoya-Reina, OC, Katiyar, N, Tomsho, LP, Kasson, LM, Hardie, R-A, Woodbridge, P, Tindall, EA, Bertelsen, MF, Dixon, D, Pyecroft, S, Helgen, KM, Lesk, AM, Pringle, TH, Patterson, N, Zhang, Y, Kreiss, A, Woods, GM, Jones, ME & Schuster, SC 2011, 'Genetic diversity and population structure of the endangered marsupial *Sarcophilus harrisii* (Tasmanian devil)', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 108, no. 30, pp. 12348-12353.

Mittermeier, RA 1997, *Megadiversity: Earth's biologically wealthiest nations*, Agrupacion Sierra Madre, Mexico City, Mexico.

Morris, KM, Cheng, Y, Warren, W, Papenfuss, AT & Belov, K 2015a, 'Identification and analysis of divergent immune gene families within the Tasmanian devil genome', *BMC Genomics,* vol. 16, no. 1017, pp. 1-12.

Morris, KM, Wright, B, Grueber, CE, Hogg, C & Belov, K 2015b, 'Lack of genetic diversity across diverse immune genes in an endangered mammal, the Tasmanian devil (*Sarcophilus harrisii*)', *Molecular Ecology,* vol. 24, no. 15, pp. 3860-3872.

Murchison, EP, Schulz-Trieglaff, OB, Ning, Z, Alexandrov, LB, Bauer, MJ, Fu, B, Hims, M, Ding, Z, Ivakhno, S & Stewart, C 2012, 'Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer', *Cell,* vol. 148, no. 4, pp. 780-791.

Patton, AH, Margres, MJ, Stahlke, AR, Hendricks, S, Lewallen, K, Hamede, RK, Ruiz-Aravena, M, Ryder, O, McCallum, HI, Jones, ME, Hohenlohe, PA & Storfer, A 2019, 'Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in Tasmanian Devils', *Molecular Biology and Evolution,* vol. 36, no. 12, pp. 2906–2921.

Peel, E, Cheng, Y, Djordjevic, J, Fox, S, Sorrell, T & Belov, K 2016, 'Cathelicidins in the Tasmanian devil (*Sarcophilus harrisii*)', *Scientific Reports,* vol. 6, no. 35019, pp. 1-9.

Peterson, BK, Weber, JN, Kay, EH, Fisher, HS & Hoekstra, HE 2012, 'Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species', *PloS One,* vol. 7, no. 5, pp. e37135.

Potter, S & Eldridge, M 2017, 'Oz Mammal Genomics', *Australasian Science,* vol. 38, pp. 19-21.

Potter, SC, Clarke, L, Curwen, V, Keenan, S, Mongin, E, Searle, SM, Stabenau, A, Storey, R & Clamp, M 2004, 'The Ensembl analysis pipeline', *Genome Research,* vol. 14, no. 5, pp. 934-941.

Pye, R, Hamede, R, Siddle, HV, Caldwell, A, Knowles, GW, Swift, K, Kreiss, A, Jones, ME, Lyons, AB & Woods, GM 2016a, 'Demonstration of immune responses against devil facial tumour disease in wild Tasmanian devils', *Biology Letters,* vol. 12, no. 10, pp. 20160553.

Pye, R, Patchett, A, McLennan, E, Thomson, R, Carver, S, Fox, S, Pemberton, D, Kreiss, A, Baz Morelli, A & Silva, A 2018, 'Immunization strategies producing a humoral IgG immune response against devil facial tumor disease in the majority of Tasmanian devils destined for wild release', *Frontiers in Immunology,* vol. 9, no. 259, pp. 1-12.

Pye, RJ, Pemberton, D, Tovar, C, Tubio, JM, Dun, KA, Fox, S, Darby, J, Hayes, D, Knowles, GW & Kreiss, A 2016b, 'A second transmissible cancer in Tasmanian devils', *Proceedings of the National Academy of Sciences,* vol. 113, no. 2, pp. 374-379.

Ralls, K, Ballou, JD, Dudash, MR, Eldridge, MD, Fenster, CB, Lacy, RC, Sunnucks, P & Frankham, R 2018, 'Call for a paradigm shift in the genetic management of fragmented populations', *Conservation Letters,* vol. 11, no. 2, pp. e12412.

Rhoads, A & Au, KF 2015, 'PacBio Sequencing and Its Applications', *Genomics Proteomics & Bioinformatics,* vol. 13, no. 5, pp. 278-289.

Rosenberg, NA, Li, LM, Ward, R & Pritchard, JK 2003, 'Informativeness of genetic markers for inference of ancestry', *American Journal of Human Genetics,* vol. 73, no. 6, pp. 1402-1422.

Russell, T, Cullingham, C, Kommadath, A, Stothard, P, Herbst, A & Coltman, D 2019, 'Development of a novel mule deer genomic assembly and species-diagnostic SNP panel for assessing introgression in mule deer, white-tailed deer, and their interspecific hybrids', *G3: Genes, Genomes, Genetics,* vol. 9, no. 3, pp. 911-919.

Schopen, G, Bovenhuis, H, Visker, M & Van Arendonk, J 2008, 'Comparison of information content for microsatellites and SNPs in poultry and cattle', *Animal Genetics,* vol. 39, no. 4, pp. 451-453.

Sedlazeck, FJ, Lee, H, Darby, CA & Schatz, MC 2018, 'Piercing the dark matter: bioinformatics of long-range sequencing and mapping', *Nature Reviews Genetics,* vol. 19, no. 6, pp. 329-346.

Selkoe, KA & Toonen, RJ 2006, 'Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers', *Ecology Letters,* vol. 9, no. 5, pp. 615-629.

Shafer, AB, Peart, CR, Tusso, S, Maayan, I, Brelsford, A, Wheat, CW & Wolf, JB 2017, 'Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference', *Methods in Ecology and Evolution,* vol. 8, no. 8, pp. 907-917.

Shafer, AB, Wolf, JB, Alves, PC, Bergström, L, Bruford, MW, Brännström, I, Colling, G, Dalén, L, De Meester, L & Ekblom, R 2015, 'Genomics and the challenging translation into conservation practice', *Trends in Ecology & Evolution,* vol. 30, no. 2, pp. 78-87.

Shaney, KJ, Adams, R, Kurniawan, N, Hamidy, A, Smith, EN & Castoe, TA 2016, 'A suite of potentially amplifiable microsatellite loci for ten reptiles of conservation concern from Africa and Asia', *Conservation Genetics Resources,* vol. 8, no. 3, pp. 307-311.

Short, J & Smith, A 1994, 'Mammal decline and recovery in Australia', *Journal of Mammalogy,* vol. 75, no. 2, pp. 288-297.

Siddle, HV, Marzec, J, Cheng, Y, Jones, M & Belov, K 2010, 'MHC gene copy number variation in Tasmanian devils: implications for the spread of a contagious cancer', *Proceedings of the Royal Society Biological Sciences Series B,* vol. 277, no. 1690, pp. 2001-2006.

Storfer, A, Epstein, B, Jones, M, Micheletti, S, Spear, SF, Lachish, S & Fox, S 2017, 'Landscape genetics of the Tasmanian devil: implications for spread of an infectious cancer', *Conservation Genetics,* vol. 18, no. 6, pp. 1287-1297.

Supple, MA & Shapiro, B 2018, 'Conservation of biodiversity in the genomics era', *Genome Biology,* vol. 19, no. 131, pp. 1-12.

Taberlet, P, Waits, LP & Luikart, G 1999, 'Noninvasive genetic sampling: look before you leap', *Trends in Ecology & Evolution,* vol. 14, no. 8, pp. 323-327.

Taylor, HR, Dussex, N & Van Heezik, Y 2017, 'Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners', *Global Ecology and Conservation,* vol. 10, pp. 231-242.

Teeling, EC, Vernes, SC, Dávalos, LM, Ray, DA, Gilbert, MTP, Myers, E & Bat1k Consortium 2018, 'Bat biology, genomes, and the Bat1K project: To generate chromosome-level genomes for all living bat species', *Annual Review of Animal Biosciences,* vol. 6, pp. 23-46.

Thorne, ET & Williams, ES 1988, 'Disease and endangered species: the black-footed ferret as a recent example', *Conservation Biology,* vol. 2, no. 1, pp. 66-74.

Tokarska, M, Marshall, T, Kowalczyk, R, Wójcik, J, Pertoldi, C, Kristensen, T, Loeschcke, V, Gregersen, V & Bendixen, C 2009, 'Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison', *Heredity,* vol. 103, no. 4, pp. 326–332.

Torkamaneh, D, Laroche, J & Belzile, F 2016, 'Genome-wide SNP calling from genotyping by sequencing (GBS) data: a comparison of seven pipelines and two sequencing technologies', *PLoS One,* vol. 11, no. 8, pp. e0161333.

United Nations Environment Programme 2021, *United Nations Decade on Ecosystem Restoration 2021-2030*, viewed 16 March 2021, https://www.decadeonrestoration.org

Van Der Kraan, LE, Wong, ES, Lo, N, Ujvari, B & Belov, K 2013, 'Identification of natural killer cell receptor genes in the genome of the marsupial Tasmanian devil (*Sarcophilus harrisii*)', *Immunogenetics,* vol. 65, no. 1, pp. 25-35.

Van Tienderen, PH, De Haan, AA, Van Der Linden, CG & Vosman, B 2002, 'Biodiversity assessment using markers for ecologically important traits', *Trends in Ecology & Evolution,* vol. 17, no. 12, pp. 577-582.

Von Mark, VC, Kilian, A & Dierig, DA 2013, 'Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species', *PLoS One,* vol. 8, no. 5, pp. e64062.

Voolstra, CR, Wörheide, G & Lopez, JV 2017, 'Corrigendum to: Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA)', *Invertebrate Systematics,* vol. 31, no. 2, pp. 231-231.

Wajid, B & Serpedin, E 2014, 'Do it yourself guide to genome assembly', *Briefings in Functional Genomics,* vol. 15, no. 1, pp. 1-9.

Willoughby, JR, Ivy, JA, Lacy, RC, Doyle, JM & Dewoody, JA 2017, 'Inbreeding and selection shape genomic diversity in captive populations: Implications for the conservation of endangered species', *PloS One,* vol. 12, no. 4, pp. e0175996.

Wright, B, Farquharson, KA, McLennan, EA, Belov, K, Hogg, CJ & Grueber, CE 2019, 'From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species', *BMC Genomics,* vol. 20, no. 453, pp. 1-10.

Wright, B, Morris, K, Grueber, CE, Willet, CE, Gooley, R, Hogg, CJ, O'Meally, D, Hamede, R, Jones, M & Wade, C 2015, 'Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population', *BMC Genomics,* vol. 16, no. 791, pp. 1-11.

Wright, B, Willet, CE, Hamede, R, Jones, M, Belov, K & Wade, CM 2017, 'Variants in the host genome may inhibit tumour growth in devil facial tumours: evidence from genome-wide association', *Scientific Reports,* vol. 7, no. 423, pp. 1-6.

Yandell, M & Ence, D 2012, 'A beginner's guide to eukaryotic genome annotation', *Nature Reviews Genetics,* vol. 13, no. 5, pp. 329-342.

Zhao, H, Fuller, A, Thongda, W, Mohammed, H, Abernathy, J, Beck, B & Peatman, E 2019, 'SNP panel development for genetic management of wild and domesticated white bass (*Morone chrysops*)', *Animal Genetics,* vol. 50, no. 1, pp. 92-96.

1.5 SUPPLEMENTARY

Table S1 Comparison of model and non-model mammalian/marsupial reference genomes to the G10K and EBP minimum reference genome quality standards. Green: metrics matching the G10K standards, Yellow: metrics matching the EBP Phase I standards, Red: metrics matching the EBP Phase II standards, Grey: metrics that fall below all VGP and EBP standards.

Reference Genome Minimum Quality Standards

Project	Phase	Contig N50	Scaffold N50	% Genome assembled into chromosomes	Inter-chromosomal rearrangements validated by >2 data sources	QV Cut-off Score*	Genome Quality Metric^
G10K (Genome 10K Community of Scientists, 2009; Koepfli et al., 2015)	-	1 Mb	10 Mb	>90%	Yes	40	3.4.2.QV40
EBP (Lewin et al., 2018)	I	0.1 Mb	1 Mb	>90%	Yes	40	2.3.2QV40
EBP (Lewin et al., 2018)	II	0.01 Mb	0.1 Mb	>90%	No	40	1.2.1QV40

Current Mammalian/Marsupial Reference Genome Metrics

Species	Genome	Contig N50 (Mb)	Scaffold N50 (Mb)	% Genome assembled into chromosomes	Inter-chromosomal rearrangements validated by >2 data sources	QV Cut-off Score	Genome Quality Metric	Date Published to NCBI
Human	GRCh38.p13	57.9	67.8	99.86%	Yes	ND	4.4.2QV?	28/2/19
Mouse	GRCm38.p6	32.8	54.5	99.97%	Yes	ND	4.4.2QV?	15/9/17
Dog	CanFam3.1	0.267	45.9	96.54%	Yes	ND	2.4.2QV?	2/11/11
Koala	phaCin_unsw_v4.1	11.6	-	0.00%	No	ND	4.4.0QV?	18/4/17
Tasmanian Devil	Devil_ref v7.0	0.0201	1.85	99.96%	Yes	30	1.3.2QV30	17/2/11

* ND = Not Determined.

^ The genome quality metric summarises all of the minimum standards from the previous columns whereby the first three numbers are the exponents of the N50 contig, N50 scaffold and level of chromosomal assembly and QV represents the minimum base-call quality error. Question marks represent unknown values.

CHAPTER 2

Ten simple rules for getting started with command-line bioinformatics

TEN SIMPLE RULES FOR GETTING STARTED WITH COMMAND-LINE BIOINFORMATICS

2.1 BACKGROUND

Chapter 2 comprises the published manuscript:

Brandies, PA & Hogg, CJ 2021, 'Ten simple rules for getting started with command-line bioinformatics', *PLOS Computational Biology,* vol. 17, no. 2, pp. e1008645.

As discussed in Chapter 1, one of the major limitations for the use of reference genomes and next-generation sequencing data in conservation contexts is the bioinformatic expertise and resources that are required to work with such datasets. With the advancement of sequencing technologies, reductions in sequencing costs, and abundance of sequencing consortia, there is a current influx of genomic data for threatened species worldwide. However, without the knowledge of how to manage and analyse big data, and an understanding of the computational resources required to do so, the downstream applications of these valuable genomic resources are limited. Many researchers are eager to harness the power of genomic datasets for answering key conservation questions, though the leap into the world of command-line bioinformatics can be challenging without a starting point. This chapter aims to equip researchers with the necessary background knowledge for undertaking and applying the bioinformatic methods presented throughout this thesis to their own species and genomic datasets. Presented as a "Ten simple rules" editorial, I provide a 10-step process encompassing a simple guide of key components to assist researchers in unlocking the true potential of genomic data.

I wrote this manuscript towards the end of my PhD, with assistance from Carolyn J. Hogg, with the aim to concisely summarise and translate some of the most useful bioinformatic tips I had learnt during the course of my doctorate degree to share with other researchers starting on the same journey. The published PDF version of this manuscript is provided in Appendix 1.

2.2 MAIN ARTICLE

Ten simple rules for getting started with command-line bioinformatics

Parice A. Brandies[1] and Carolyn J. Hogg[1] *

1. School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia.

* Corresponding Author

Introduction

Sequencing technologies are becoming more advanced and affordable than ever before. In response, growing international consortia such as the Earth BioGenomes Project (EBP) (Lewin et al., 2018), the Genome 10K project (G10K) (Genome 10K Community of Scientists, 2009; Koepfli et al., 2015), the Global Invertebrate Genomics Alliance (GIGA) (GIGA Community of Scientists, 2013; Voolstra, Wörheide & Lopez, 2017), the Insect 5K project (i5K) (Consortium, 2013; Levine, 2011), the 10,000 plants project (10KP) (Cheng et al., 2018), and many others, have big plans to sequence all life on earth. These consortia aim to utilise genomic data to uncover the biological secrets of our planet's biodiversity and apply this knowledge to real-world matters, such as improving our understanding of species' evolution, assisting with conservation of threatened species, and identifying new targets for medical, agricultural or industrial purposes (Lewin et al., 2018). All of these goals rely on someone to analyse and make sense of the tremendous amounts of biological data, making bioinformaticians more sought-after than ever. Many researchers with a background in biology and genetics are stepping up to the challenge of big data analysis, but it can be a little daunting to start down the path of bioinformatics, particularly using the command line, without a strong background in computing and/or computer science. A recent "Ten simple rules" article highlighted the importance of bioinformatics research support (Kumuthini et al., 2020). Here we provide ten simple rules for anyone interested in taking the leap into the realm of bioinformatics using the command line. We have put together these ten simple rules for those starting on their bioinformatics journey, whether you be a student, an experienced biologist or geneticist, or anyone else who may be interested in this emerging field. The rules are presented in chronological order, together encompassing a simple 10-step process for getting started with command-line bioinformatics (Figure

2.1). This is by no means an exhaustive introduction to bioinformatics, but rather a simple guide to the key components to get you started on your way to unlocking the true potential of biological big data.

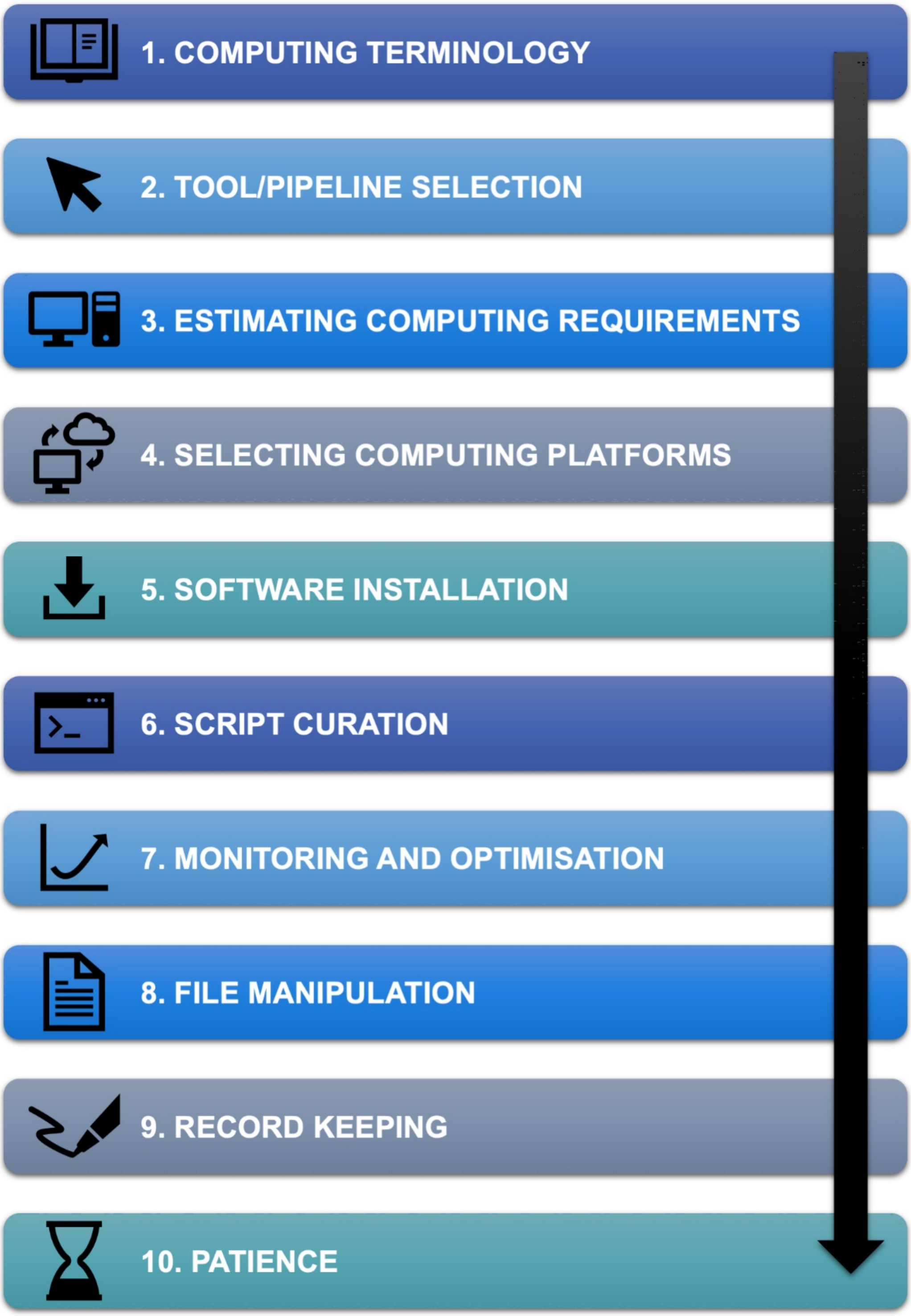

1. COMPUTING TERMINOLOGY
2. TOOL/PIPELINE SELECTION
3. ESTIMATING COMPUTING REQUIREMENTS
4. SELECTING COMPUTING PLATFORMS
5. SOFTWARE INSTALLATION
6. SCRIPT CURATION
7. MONITORING AND OPTIMISATION
8. FILE MANIPULATION
9. RECORD KEEPING
10. PATIENCE

Figure 2.1 Our 10-step process for getting started with command-line bioinformatics. Each step corresponds to each of our ten simple rules presented below.

Rule 1: Get familiar with computer terminology

The first step in your command-line bioinformatics journey can be overwhelming due to the wealth of new terminology. This is where you need to channel your inner computer geek and learn the new language of computer terminology. In fact, this very paper is riddled with it, so our first rule addresses this tricky obstacle. Having a basic understanding of computing and associated terminology can be really useful in determining how to run your bioinformatics pipelines effectively. It can also help you troubleshoot many errors along the way. Understanding the terminology allows you to talk with your institutional information technology (IT) departments and communicate your computational needs to answer your biological questions. This will allow you to be able to source the resources you will need. A number of basic definitions of the main terms that you will likely come across as you enter the world of bioinformatics is presented in Box 1.

Box 1. Some simple definitions of common computer terms

Algorithm: The set of rules or calculations that are performed by a computer program. Certain algorithms may be more suitable for particular datasets and may have differences in performance (e.g., in speed or accuracy).

Central processing unit (CPU): The chip that performs the actual computation on a compute node or VM.

Compute Node: An individual computer that contains a number of CPUs and associated RAM.

Core: Part of a CPU. Single-core processors contain one core per CPU, meaning CPUs and cores are often interchangeable terms.

CPU Time: The time CPUs have spent actually processing data (often CPU time ~= Walltime * Number of CPUs).

Dependency: Software that is required by another tool or pipeline for successful execution.

Executable: The file that contains a tool/program. Some software has a single executable while others have multiple executables for different commands/steps.

High Performance Computer (HPC): A collection of connected compute nodes.

Operating System (OS): The base software that supports a computer's basic functions. Some of the most common Linux-based operating systems include those of the Debian distribution (Ubuntu), and those of the RedHat distribution (Fedora and CentOS).

Pipeline: A pipeline is a workflow consisting of a variety of steps (commands) and/or tools that process a given set of inputs to create the desired output files.

Programming languages: Specific syntax and rules for instructing a computer to perform specific tasks. Common programming language used in bioinformatics include Bash, Python, Perl, R, C and C++.

Random access memory (RAM): Temporarily stores all the information the CPUs require (can be accessed by all of the CPUs on the associated node or VM).

Scheduler: Manages jobs (scripts) running on shared HPC environments. Some common schedulers include SLURM, PBS, Torque and SGE.

Script: A file which contains code to be executed in a single programming language.

Thread: Number of computations that a program can perform concurrently – depends on the number of cores (usually 1 core = 1 thread).

Tool: A software program that performs an analysis on an input dataset to extract meaningful outputs/information - Tool, software and program are often used interchangeably but refer to the core components of bioinformatics pipelines.

VM: Virtual machine - Similar to a compute node as it behaves as a single computer and contains a desired number of CPUs and associated RAM (usually associated with Cloud Computing).

Walltime: The time a program takes to run in our clock-on-the-wall time.

Rule 2: Know your data and needs to determine which tool or pipeline to use

This can often be one of the most difficult steps as there are usually many different tools and pipelines to choose from for each particular bioinformatic analysis. While you may think about creating your own tool to perform a particular task, more often than not there is already a pre-existing tool that will suit your needs, or perhaps only need minor tweaking to achieve the required result. Having a clear understanding of your data and the types of questions you are wanting to ask, will go a long way to assisting in your tool or pipeline selection. Selecting the most suitable pipeline or tool will be dependent on a number of factors including:

Your target species and quality of data

Some bioinformatic pipelines/software may work better for a particular species based on their unique features (e.g., genome size, repeat complexity, ploidy, etc.) or based on the quality of data (e.g., scaffold length, short reads vs long reads, etc.). Reading other published papers on similar species will assist with being able to define this.

Your available computing resources and time restrictions

Certain software may be based of different algorithms, which can result in significant reductions or increases of computational resources and walltime. Some shared HPC infrastructure may have walltime limitations in place, or the amount of RAM or cores may be a limiting factor when using personal computing resources. Make enquiries with your institutional IT department regarding limits on personal computing or HPC infrastructure before you start.

Which tools are readily available

Many bioinformatic pipelines and tools are freely available for researchers, though some require purchasing of a license. Additionally, some tools/pipelines may already be available on your desired computing infrastructure or through your local institution. There are a number of “standard” bioinformatic command line tools that have broad applicability across a variety of genomic contexts and are therefore likely already installed on shared infrastructure. Such examples include tabix, FastQC, samtools, vcftools/bcftools, bedtools, GATK, BWA, PLINK and BUSCO. Furthermore, collaborators or other researchers may have already tested and optimised a particular pipeline on a certain infrastructure and have therefore already overcome the first hurdle for you.

Talking with colleagues who are working on similar projects and reading through the literature is often the best way to decide on which software to use for a particular analysis. There are many publications that benchmark different tools and compare the advantages and disadvantages of similar pipelines. There are also many online web forums (e.g., BioStars [Parnell et al., 2011]) that may also assist with your decision-making process. Be sure to search through the different web forums to see whether another researcher has also asked the same or similar question as you (this is often the case). If you cannot find a solution ensure any questions you post are clear and detailed, with examples of code or errors provided to have the best chance of helpful replies and answers. Beginning with a pipeline that has previously been tested and optimised on a particular platform is helpful in getting a head start, though do not be scared to try out a new or different pipeline if it seems better suited to your data or desired outcome.

Rule 3: Estimate your computing requirements

Once you have selected your desired tool or pipeline, the next crucial step involves estimating the desired computing requirements for your chosen analysis. Estimating your requirements will not only allow you to determine which platforms may be most suitable to run your pipeline (e.g., cloud vs HPC; see Rule 4) but will also reduce time spent on troubleshooting basic resource errors (e.g., running out of RAM or storage space). Furthermore, this step is almost always necessary prior to running any tool or pipeline on any given compute infrastructure. For instance, on shared HPC environments, your job script will need to include your requested computational resources (cores, RAM, walltime), and you will need to make sure you have enough disk space available for your account. Similarly, for cloud computing, you will need to decide what size machine/s (cores and RAM) and how much attached storage you need for your analysis. Estimating incorrectly can be frustrating as you will waste time in queues on shared HPC infrastructure, only to have your analysis terminated prematurely, or waste money in the cloud specifying more resources than you actually need. Many bioinformatics tools can be run on a single core by default, but this can result in much greater walltimes (Kawalia et al., 2015) (which are often restricted on shared HPC infrastructure). Increasing the number of cores can greatly reduce your walltime though there is often a balance between this and other important factors such as RAM usage, cost, queueing time etc. (Kawalia et al., 2015).

It can be a little tricky estimating computing requirements for a pipeline you have never run before, or on a species that the pipeline has never been tested with before. Never fear though, as there are a number of places you can seek out information on computing requirements. First and foremost, read the documentation for the pipeline/tool you are running. Some tool documentation will provide an example of the compute resources required or provide suggestions. Additionally, many programs will provide a test dataset to ensure the pipeline is working correctly before employing your own datasets. These test datasets are a great start for estimating minimal computational requirements and to obtain some general benchmarks when using different parameters or computing resources. If the tool documentation does not provide a guide of computing requirements or an example dataset, you may wish to use a smaller subset of your own data for initial testing. The literature may also provide a guide for general computing requirements that have been used for a particular tool or pipeline for a similar species or sample size. There are many publications where common bioinformatics pipelines are compared with one another to assess performance and results across a variety of organisms (e.g., Cornish & Guda, 2015; Khan et al., 2018; Schilbert, Rempel & Pucker, 2020; Zhang et al., 2017). These can be found with a simple citation search. Finally, another great resource for estimating your computing requirements is from other researchers. Talking to others in your field, who may work with similar data or utilising online forums such as BioStars (Parnell et al., 2011), will assist in understanding the resources required.

In general, 32 cores and 128 Gb of RAM is usually sufficient for most common bioinformatics pipelines to run within a reasonable timeframe. With that being said, some programs might require much less than this while others may have much higher memory requirements or enable greater parallelisation.

Rule 4: Explore different computing options

After estimating your computing requirements for your chosen pipeline, you will then need to determine where such resources are available and which infrastructure will best suit your needs. Some tools may easily run on a personal computer, though many of the large bioinformatics pipelines (particularly when working on organisms with large genomes like mammals and plants) require computational resources that will well exceed a standard PC. Many institutions have a local HPC or access to national/international HPC infrastructure. However, the unprecedented generation of sequencing data has started to push these shared infrastructures to their limits. These resources are not always well suited to the requirements of bioinformatic pipelines such as their high I/O demands and “bursty” nature (see Rule 7) (O’Driscoll, Daugelaite & Sleator, 2013). This is why cloud computing is becoming increasingly popular for bioinformaticians (Kwon et al., 2015; O’Driscoll, Daugelaite & Sleator, 2013; Shanker, 2012; Stein, 2010; Zhao et al., 2017).

Cloud computing provides a number of key advantages over traditional shared HPC resources including:

- The ability to tailor your computing resources for each bioinformatic tool or pipeline you wish to use
- Complete control over your computing environment (i.e., operating system, software installation, file system structure etc.)
- Absence of a queuing system resulting in faster time to research
- Unlimited scalability and ease of reproducibility

Utilising cloud resources also prevents the need for researchers to purchase and maintain their own physical computer hardware (which can be time consuming, costly and nowhere near as scalable [Fox, 2011]). However, commercial cloud computing does come at a cost, and can be a bit of a steep learning curve. Fortunately, services like RONIN (https://ronin.cloud) have simplified the use of cloud computing for researchers and allow for simple budgeting and cost monitoring to ensure research can be conducted in a simple, cost-effective manner. Researchers at academic institutions may also have access to other free cloud compute services such as Galaxy (https://usegalaxy.org/), ecocloud (https://ecocloud.org.au/), nectar (https://nectar.org.au/cloudpage/) and CyVerse (https://www.cyverse.org).

Overall, deciding where to run your analysis will be dependent on your data/species, what platforms are most easily accessible to you, your prior experience, your timeline and your budget. Exploring different compute options will allow you to choose which infrastructure best suits your needs and enable you to adapt to the fast-evolving world of bioinformatics.

Rule 5: Understand the basics of software installation

When wanting to utilise a personal resource for your bioinformatic pipelines, such as a cloud VM or a personal computer, you will need to familiarise yourself with the various installation methods for your required tools. While software installation is sometimes provided as a service for some shared HPC platforms, understanding the basics of software installation is useful in helping you troubleshoot any installation-based errors, and identify which software you can likely install locally yourself (i.e., without requiring root user privileges). There are numerous ways software can be installed but we have provided four main methods that should cover most bioinformatics software (Box 2).

Box 2. Common software installation methods for bioinformatics tools

Package Managers

APT (Advanced Package Tool) (https://www.debian.org/doc/manuals/apt-guide/index.en.html) is a package manager that is often already installed by default on many Debian distributions and enables very simple installation of available tools. APT works with a variety of core libraries to automate the download, configuration and installation of software packages and their dependencies. A number of common bioinformatics tools are available through APT including NCBI blast+, samtools, hmmer, vcftools, bcftools, bedtools among others. If working on a RedHat operating system, the package manager YUM (Yellowdog Updater, Modified) (https://access.redhat.com/solutions/9934) is the equivalent of APT.

Conda

Conda (https://docs.conda.io/en/latest/) is also a package management tool, though it sits somewhere between package managers like APT and containers (see below) due to its ability to also manage environments (i.e., collections of software). This feature makes conda extremely useful, particularly for bioinformatics software where different pipelines may utilise the same tools but require different versions of a particular tool. Conda allows you to easily install and run pipelines in their own separate environments so they do not interfere with one another and also enables you to easily update software when new versions are made available. Bioconda (Grüning et al., 2018) is a channel for conda which specialises in bioinformatics software and includes a myriad of the most commonly used bioinformatic tools. Furthermore, conda also enables the installation and management of popular programming languages such as python or R, along with their respective libraries and packages. It is a great resource for bioinformaticians of all levels and is particularly helpful as a stepping-stone before stepping down a container lane.

Containers

Containers package up software and all dependencies, as well as all of the base system tools and system libraries into a separate environment so that they can be reliably run on different computing platforms. Containers are similar to conda environments, but they differ in the sense that containers include absolutely everything they need within the container itself (even including the base operating system). It is sometimes easier to think about containers as installing a whole separate machine that just utilises the same computing resources and hardware as the local machine it is installed on. The main advantage of a container over a conda environment is the ease of reproducibility due to the ability to pull a specific container each time you want to run, or re-run, a certain pipeline or use a particular tool, no matter what computing platform you are using. Reproducibility can be achieved with conda environments too, but this often requires exporting and keeping track of saved environments.

There are two main options when wanting to use a container: Docker (Merkel, 2014) or Singularity (Kurtzer, Sochat & Bauer, 2017). Docker is the most standard container service available with thousands of containers available from DockerHub (https://hub.docker.com) or from other container registries such as quay.io (https://quay.io). Bioinformatics software that is available via bioconda also has a respective docker container on quay.io through the BioContainers architecture (da Veiga Leprevost et al., 2017). This means many common bioinformatics software and pipelines are already available in a containerised environment. Otherwise some software developers make their own containers available, e.g. Trinity (for RNA-seq assembly) (see https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-in-Docker) or BUSCO v4 (for assessing assembly completeness) (see https://busco.ezlab.org/busco_userguide.html#docker-image). There are also thousands of other public docker containers across a range of online container registries that may have the software you are looking for, or there is always the option to create your own Docker container for reproducible pipelines. Obviously, Docker can be used to download and employ Docker containers, but singularity is another program that can also be used to download and employ Docker containers (particularly on HPC environments). Both have advantages and disadvantages, so it is usually down to user preference as to which to choose. If you are new to containers, we suggest starting with Singularity. Not only will this allow you to easily be able to scale up your containerised pipelines to HPC environments but also makes reading and writing files to and from the container from the local machine a bit more straightforward.

Manual Installation

If none of the above methods are available for your chosen software, you may need to install it manually. This process is usually explained step-by-step in the software documentation but typically involves a number of steps including: 1) Downloading a tar package (or zip file) of the source code (or cloning a Git repository) from GitHub (https://github.com) (or another website); 2) Unpacking the source code to extract its contents; 3) Configuring the software to check your environment and ensure all of the required dependencies are available; 4) Building the finished software from the source code; and 5) Installing the software i.e. copying the software executables, libraries and documentation to the required locations. This process is what package managers and containers do automatically for you. There are a number of standard dependencies that are usually required for manual installation (e.g., the build-essential package, the dh-autoreconf package and the libarchive-dev package) so it is often handy to install these using APT before attempting to manually install any other software. You will be notified of any other required dependencies you may be missing during the installation process.

Once you have your software installed it is good practice to try and run the program with the help command-line option (i.e. -h/--help/-help), or with no parameters, to ensure it has been installed correctly. If the help option displays some information about running the program and the different command-line options, it is usually a good sign that your software was installed successfully and is ready to go. If your tool does not seem to be working, you may need to ensure the executable for your tool (and sometimes its required dependencies) is available in your path. But what exactly is your path and why is it important? Well, whenever we call upon a particular input file or output directory within a command, we often use an absolute or relative path to show the program where that file or directory is sitting within the file system hierarchy. We can also call upon tools or executables the same way, though it is not efficient to provide a path to a tool every time we need to use it. The path environmental variable overcomes this issue by providing a list of directories that contain tools/executables you may wish to execute.

By default, the path variable is always set to include some standard directories that include a variety of system command-line utilities. So, to ensure a new program can be called upon anywhere without specifying the path to the program, you can either move or copy the tool/executable to a directory that is already listed in your path variable, or add a new directory to the path variable that contains the program. New directories can be added to your path either temporarily (by simply exporting the path variable with the added directory included) or permanently (by editing your .bash_profile). Another thing to be aware of is that the order of directories in your path is important because if the same program (or executable with the same name) is found in two different directories, the one that is found first in your path will be used. Always keep this in mind when adding new directories to your path to determine where they should sit in the list of paths. [The sheer number of times we mentioned the word "path" in this rule alone should emphasize how important paths really are – though we promise there are no more mentions of it for the rest of this article].

Rule 6: Carefully curate and test your scripts

In other words, always double-check (or triple-check) your scripts and perform test-runs at each step along the way. Before you run your pipeline, it is important to first read through the software documentation to ensure you understand the different inputs, outputs, and analysis options. Ensure that the documentation is for the correct version of the software as particular command-line options may change version-to-version. Many bioinformatics programs have extensive documentation online, either through their GitHub or another website. The basic documentation for most tools can be accessed using the command-line help options (which is also a great way to determine whether your required tool is available and installed correctly - see Rule 5). Sometimes more detailed information can be found in a README file in the source code directory. Most documentation should provide some example commands on how to run the program with basic or default options, which should assist you in curating a successful script.

Once you have your final script, it is essential to give it a quick test to determine if there are any immediate errors that will prevent your script from running successfully. From simple spelling mistakes or syntax errors which result in files or directories not being found or commands being confused with invalid options, to not being able to locate the desired software or the software being configured incorrectly with problematic dependencies. These are the “face-palm” errors that any bioinformatician is aware of as we have all been there, time and time again. The good news is that these errors are often quite simple to fix. Yet it is better to catch them early rather than waiting in queues only for your script to error as soon as it starts, or leaving your script to run in the cloud only to come back and realise the machine has been sitting there idle the whole time due to a minor scripting error. Testing your scripts in the cloud is usually as simple as running the script or command and watching to see whether any errors are immediately thrown on-screen, but to test scripts in a shared HPC environment, you may need to utilise an interactive queue. Interactive queues allow you to run commands directly from the command line with a small subset of HPC resources. These resources are usually not enough to run an entire pipeline but are quite useful for testing and debugging purposes. Obviously, your script may still run into errors later on in your pipeline but testing your script before you submit it properly should alert you to any preliminary errors that would prevent the pipeline from starting successfully and prevent any precious time being wasted in queues or precious dollars being wasted on idle cloud compute.

Rule 7: Monitor and optimise your pipelines

Once you have your script running, it is important to monitor your pipelines to determine whether it is effectively utilising the computational resources you have allocated to it. Understanding what resources your pipeline utilises can help you scale up or down your compute so that you are not wasting resources or hitting resource limits that may slow down your pipeline. On shared HPC infrastructure, you will usually be able to see a summary of the computational resources used from either the job log files or scheduler specific commands. Metrics such as maximum RAM and CPU usage as well as CPU time and walltime are useful in adjusting future scripts so that they request the optimum amount of resources needed. This enables the pipeline to run efficiently without any unnecessary queue time. Storage space of output files should also be monitored periodically to ensure you are not exceeding your allocated quota.

More specific monitoring is possible when running pipelines in the cloud as you have full control over all computing resources. Simple programs like htop (https://hisham.hm/htop/) can be used for fast real-time monitoring of basic metrics like CPU and RAM usage, while more in-depth programs like Netdata (https://www.netdata.cloud) can assist with tracking a large variety of metrics both in real-time and across an entire pipeline using hundreds of pre-configured interactive graphs. Many bioinformatic pipelines are “bursty” in nature, meaning different steps in a single pipeline may have vastly different computing requirements. Some steps/tools may have high memory requirements but only utilise a small number of cores, while others may multi-thread quite well across a large number of cores but require minimal memory. Knowing the required computing resources for each step may help you break up your pipeline and run each stage on a different machine type for greater cost efficiency. Monitoring disk space requirements throughout a pipeline is also important as many bioinformatics tools require large amounts of temporary storage that are often cleaned upon completion of the pipeline. Attached storage can be quite costly in the cloud, so ensuring you only request what is necessary will also reduce pipeline costs.

Overall, monitoring of bioinformatics pipelines is key to improving pipeline efficiency, optimising computing resources, reducing wasted queue time, and reducing cloud costs.

Rule 8: Get familiar with basic bash commands

As a bioinformatician, your main role is to make sense of biological datasets and this often means manipulating, sorting and filtering input and output files to and from various bioinformatic tools and pipelines. For example, you may want to extract information for a certain sample, or a certain gene of interest. Or in a file containing a table of data, you may want to sort an output file by a particular column or select rows that contain a particular value. You may want to replace a certain ID with a respective name from a list or perform a calculation on values within a column. Fortunately, many of the input and output files used in bioinformatics are regular text files, so these tasks can easily be achieved. One might think about using common spreadsheet applications such as Microsoft Excel to perform these tasks, however while this may suffice for small files, Excel is not too fond of the sometimes millions of rows of data that are characteristic of a number of common bioinformatic files. This is where some standard Unix shell command-line utilities come into play, namely the grep, AWK and sed utilities.

Global regular expression print (grep) is a command-line utility that searches a text file for a regular expression (i.e., a pattern of text) and returns lines containing the matched expression (Table 2.1). This tool is useful when wanting to filter or subset a file based on the presence of a particular word or pattern of text (e.g., a sample name or genomic location etc). AWK is much more extensive command-line utility that enables more specific file manipulation of column-based files (Table 2.1). For example, AWK can return lines where a column contains a particular value or regular expression, in addition it can output only particular columns, perform calculations on values within the columns and work with multiple files at once. The extensive abilities of AWK are too many to cover here, but just know that this clever little tool will likely hold a special place in any bioinformaticians heart. Lastly, stream editor (sed) has a basic "find and replace" usage allowing you to transform defined patterns in your text. In its most basic form, sed can replace a word with another given word (Table 2.1) but can also perform more useful functions like removing everything before or after a certain pattern or adding text at certain places in a file.

Table 2.1 Basic usage examples of the grep, awk and sed commands.

Command	Example	Description
grep	grep "chr5" file	Print all lines that contain the string "chr5" in the named file
awk	awk '$1 == 5 {print $2, $3}' file	For rows in the named file where the value in column 1 is equal to 5, print columns 2 and 3
sed	sed 's/sample1/ID7037/g' file	Replace all occurrences of "sample1" with "ID7037" in the named file and print the result

Of course, grep, AWK and sed all have their limitations and more extensive file manipulation may be better suited to a python or perl script (and there is already a great "Ten simple rules" article for biologists wanting to learn how to program [Carey & Papin, 2018]); but for simple processing, filtering and manipulation of bioinformatics files, look no further than these three useful command-line utilities.

Rule 9: Write it down!

A previous "Ten simple rules" article has highlighted the importance of keeping a laboratory notebook for computational biologists (Schnell, 2015), and another covered some best practices around the documentation of scientific software (Lee, 2018). Many components from these articles apply to our rule of writing it down and keeping helpful notes when getting started with command-line bioinformatics. The number of pipelines or analyses that can be run on a single set of biological data can sometimes be quite extensive, and usually coincides with a lot of trial and error of different parameters, computing resources, and/or tools. Even those with a great memory will often look back at results at the time of publication and ponder “why did we use that tool?”, or “what parameters did we end up deciding on for that analysis?”. Keeping detailed notes is crucial to research integrity and can be a real lifesaver. Not only is it important to keep track of your different script files, and the required computing resources for each script, but also the accompanied notes about why you chose a particular tool and any troubleshooting you had to do to run the pipeline successfully. An easy-to-access document of all of your favourite commands and nifty pieces of code that may come in handy time and time again is also a must! Getting familiar with helpful code text editors like Visual Studio Code (https://code.visualstudio.com), or Atom (https://atom.io), as well as investing some time into learning helpful mark-up languages like Markdown will assist with keeping detailed, organised and well-formatted scripts and documentation for the pipelines you are using. Exactly how you decide to keep your notes is completely up to you, but just ensure to keep everything well-organised, up-to-date, and backed up. Also publishing your scripts as markdown files in supplementary material ensures the utility, transparency (and citability) of your work.

Rule 10: Patience is key

The number one key (that we’ve saved until last) to being a successful bioinformatician is patience. A large proportion of your time will be spent troubleshooting software installation, computing errors, pipeline errors, scripting errors or weird results. Some problems are simple to solve while others may take quite some time. You will likely feel that with every step forward there is just another hurdle to cross. Yet if you are patient and push through every error that is thrown your way, the euphoria of conquering a bioinformatics pipeline and turning a big lump of numeric data or As, Ts, Cs and Gs into something biologically meaningful is well worth it. Also, as many past “Ten simple rules” articles in this field have addressed, do not be afraid to raise your hand and ask for help when you get stuck. Most of the time, someone before you has been in the exact same situation and encountered the same error or tackled a similar problem. Google will become your best friend and first port of call when things are not going as planned. And on the rare occasion where endless googling leads you nowhere, talk with your peers and reach out to the bioinformatic community, people are often more than happy to share their knowledge and put their problem-solving skills to the test.

Conclusion

In the new era of whole genome sequencing, bioinformaticians are now more sought-after than ever before. Stepping into the world of command-line bioinformatics can be a steep learning curve but is a challenge well worth undertaking. We hope these ten simple rules will give any aspiring bioinformatician a head-start on their journey to unlocking the meaningful implications hidden within the depths of their biological datasets.

Acknowledgements

Our journey into bioinformatics and this article would not have been possible without the support of the teams at the Sydney Informatics Hub, Australian BioCommons, RONIN, Amazon Web Services (AWS), Pawsey Supercomputing Centre, National Computational Infrastructure and Microsoft Azure. Special thanks to Rosie Sadsad, Nathan Albrighton, Johan Gustafsson, Jeff Christiansen, Kathy Belov and the team members of the Australasian Wildlife Genomics Group.

References

Carey, MA & Papin, JA 2018, 'Ten simple rules for biologists learning to program', *PLoS Computational Biology,* vol. 14, no. 1, pp. e1005871.

Cheng, S, Melkonian, M, Smith, SA, Brockington, S, Archibald, JM, Delaux, P-M, Li, F-W, Melkonian, B, Mavrodiev, EV & Sun, W 2018, '10KP: A phylodiverse genome sequencing plan', *Gigascience,* vol. 7, no. 3, pp. giy013.

Cornish, A & Guda, C 2015, 'A comparison of variant calling pipelines using genome in a bottle as a reference', *BioMed Research International,* vol. 2015, no. 456479, pp. 1-11.

Da Veiga Leprevost, F, Grüning, BA, Alves Aflitos, S, Röst, HL, Uszkoreit, J, Barsnes, H, Vaudel, M, Moreno, P, Gatto, L & Weber, J 2017, 'BioContainers: an open-source and community-driven framework for software standardization', *Bioinformatics,* vol. 33, no. 16, pp. 2580-2582.

Fox, A 2011, 'Cloud computing—what's in it for me as a scientist?', *Science,* vol. 331, no. 6016, pp. 406-407.

Genome 10k Community of Scientists 2009, 'Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species', *Journal of Heredity,* vol. 100, no. 6, pp. 659-674.

Giga Community of Scientists 2013, 'The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes', *Journal of Heredity,* vol. 105, no. 1, pp. 1-18.

Grüning, B, Dale, R, Sjödin, A, Chapman, BA, Rowe, J, Tomkins-Tinch, CH, Valieris, R & Köster, J 2018, 'Bioconda: sustainable and comprehensive software distribution for the life sciences', *Nature Methods,* vol. 15, no. 7, pp. 475-476.

I5k Consortium 2013, 'The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment', *Journal of Heredity,* vol. 104, no. 5, pp. 595-600.

Kawalia, A, Motameny, S, Wonczak, S, Thiele, H, Nieroda, L, Jabbari, K, Borowski, S, Sinha, V, Gunia, W & Lang, U 2015, 'Leveraging the power of high performance computing for next generation sequencing data analysis: tricks and twists from a high throughput exome workflow', *PloS One,* vol. 10, no. 5, pp. e0126321.

Khan, AR, Pervez, MT, Babar, ME, Naveed, N & Shoaib, M 2018, 'A comprehensive study of de novo genome assemblers: current challenges and future prospective', *Evolutionary Bioinformatics,* vol. 14, no. 1176934318758650, pp. 1-8.

Koepfli, K-P, Paten, B, Genome 10k Community of Scientists & O'Brien, SJ 2015, 'The Genome 10K Project: a way forward', *Annual Review of Animal Biosciences,* vol. 3, no. 1, pp. 57-111.

Kumuthini, J, Chimenti, M, Nahnsen, S, Peltzer, A, Meraba, R, McFadyen, R, Wells, G, Taylor, D, Maienschein-Cline, M & Li, J-L 2020, 'Ten simple rules for providing effective bioinformatics research support', *PLoS Computational Biology,* vol. 13, no. 3, pp. e1007531.

Kurtzer, GM, Sochat, V & Bauer, MW 2017, 'Singularity: Scientific containers for mobility of compute', *PloS One,* vol. 12, no. 5, pp. e0177459.

Kwon, T, Yoo, WG, Lee, W-J, Kim, W & Kim, D-W 2015, 'Next-generation sequencing data analysis on cloud computing', *Genes & Genomics,* vol. 37, no. 6, pp. 489-501.

Lee, BD 2018, 'Ten simple rules for documenting scientific software', *PLoS Computational Biology,* vol. 14, no. 12, pp. e1006561.

Levine, R 2011, 'i5k: the 5,000 insect genome project', *American Entomologist,* vol. 57, no. 2, pp. 110-113.

Lewin, HA, Robinson, GE, Kress, WJ, Baker, WJ, Coddington, J, Crandall, KA, Durbin, R, Edwards, SV, Forest, F & Gilbert, MTP 2018, 'Earth BioGenome Project: Sequencing life for the future of life', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 115, no. 17, pp. 4325-4333.

Merkel, D 2014, 'Docker: lightweight linux containers for consistent development and deployment', *Linux Journal,* vol. 2014, no. 239, pp. 2.

O'Driscoll, A, Daugelaite, J & Sleator, RD 2013, ''Big data', Hadoop and cloud computing in genomics', *Journal of Biomedical Informatics,* vol. 46, no. 5, pp. 774-781.

Parnell, LD, Lindenbaum, P, Shameer, K, Dall'olio, GM, Swan, DC, Jensen, LJ, Cockell, SJ, Pedersen, BS, Mangan, ME & Miller, CA 2011, 'BioStar: an online question & answer resource for the bioinformatics community', *PLoS Computational Biology,* vol. 7, no. 10, pp. e1002216.

Schilbert, HM, Rempel, A & Pucker, B 2020, 'Comparison of read mapping and variant calling tools for the analysis of plant NGS data', *Plants,* vol. 9, no. 4, pp. 439.

Schnell, S 2015, 'Ten simple rules for a computational biologist's laboratory notebook', *PLoS Computational Biology,* vol. 11, no. 9, pp. e1004385.

Shanker, A 2012, 'Genome research in the cloud', *OMICS A Journal of Integrative Biology,* vol. 16, no. 7-8, pp. 422-428.

Stein, LD 2010, 'The case for cloud computing in genome informatics', *Genome Biology,* vol. 11, no. 207, pp. 1-7.

Voolstra, CR, Wörheide, G & Lopez, JV 2017, 'Corrigendum to: Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA)', *Invertebrate Systematics,* vol. 31, no. 2, pp. 231-231.

Zhang, C, Zhang, B, Lin, L-L & Zhao, S 2017, 'Evaluation and comparison of computational tools for RNA-seq isoform quantification', *BMC Genomics,* vol. 18, no. 583, pp. 1-11.

Zhao, S, Watrous, K, Zhang, C & Zhang, B 2017, 'Cloud computing for next-generation sequencing data analysis', in Jaydip Sen (ed.), *Cloud Computing-Architecture and Applications*, InTech, Rijeka.

CHAPTER 3

Characterisation of reproductive gene diversity in the endangered Tasmanian devil

CHARACTERISATION OF REPRODUCTIVE GENE DIVERSITY IN THE ENDANGERED TASMANIAN DEVIL

3.1 BACKGROUND

Chapter 3 comprises the published manuscript:

Brandies, PA, Wright, BR, Hogg, CJ, Grueber, CE & Belov, K 2020, 'Characterisation of reproductive gene diversity in the endangered Tasmanian devil', Molecular Ecology Resources, vol. 00, pp. 1-12.

This Chapter is an extension of the work presented in Chapter 1 on the Tasmanian devil. Here I show how previously existing genomic data for well-researched Australian marsupial species can be further explored to investigate species-specific questions that have implications in downstream conservation management. Specifically, I use the Tasmanian devil reference genome and whole genome resequencing data from previous studies to explore reproductive gene diversity in the Tasmanian devil. This research identifies a number of polymorphic genes across 37 individuals that may have functional consequences on reproduction and provides the crucial foundation for future work to examine the effects of genetic diversity on reproductive fitness in Tasmanian devil populations. This work is important in demonstrating the usefulness of pre-existing reference genomes and associated genomic data.

I compiled the list of target genes, performed bioinformatic gene characterisation and SNP prediction and wrote the manuscript. Katherine Belov, Catherine E. Grueber and Carolyn J. Hogg contributed to the design of the study and sourced funding. Belinda Wright performed alignments of resequencing data to the reference genome and assisted me with the analysis of SNP genotypes. All authors revised the manuscript. Supplementary material is presented at the end of this chapter. The published PDF version of this manuscript is provided in Appendix 1.

3.2 MAIN ARTICLE

Characterisation of reproductive gene diversity in the endangered Tasmanian devil

Parice A. Brandies[1], Belinda R. Wright1, Carolyn J. Hogg[1], Catherine E. Grueber[1,2] and Katherine Belov[1]*

1. School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia.

2. San Diego Zoo Global, San Diego, CA, USA.

* Corresponding Author

Abstract

Inter-individual variation at genes known to play a role in reproduction may impact reproductive fitness. The Tasmanian devil is an endangered Australian marsupial with low genetic diversity. Recent work has shown concerning declines in productivity in both wild and captive populations over time. Understanding whether functional diversity exists at reproductive genes in the Tasmanian devil is a key first step in identifying genes that may influence productivity. We characterised single nucleotide polymorphisms (SNPs) at 214 genes involved in reproduction in 37 Tasmanian devils. Twenty genes contained non-synonymous substitutions, with genes involved in embryogenesis, fertilisation and hormonal regulation of reproduction displaying greater numbers of nonsynonymous SNPs than synonymous SNPs. Two genes, *ADAMTS9* and *NANOG*, showed putative signatures of balancing selection indicating that natural selection is maintaining diversity at these genes despite the species exhibiting low overall levels of genetic diversity. We will use this information in future to examine the interplay between reproductive gene variation and reproductive fitness in Tasmanian devil populations.

Introduction

Globally the number of species threatened with extinction is increasing as a result of human-induced activities including habitat fragmentation, invasive predators, and pollution. Genetic diversity at functional gene families can have long-term consequences on species adaptation and survival in a changing world (Holderegger, Kamm & Gugerli, 2006; Mimura et al., 2017). Understanding the causes and consequences of inter-individual variation sits at the core of evolution and ecology, yet despite decades of molecular research, the genetic basis of phenotypic variation, i.e., genetic polymorphism, remains poorly quantified for the vast majority of species and traits (Forsman & Wennersten, 2016; Mimura et al., 2017). However, recent advances in sequencing technology have better enabled researchers to investigate inter-individual variation at gene families and determine how this variation is linked to important phenotypic traits. For example, genetic diversity at immune genes, particularly genes of the major histocompatibility complex (MHC), have been associated with a range of key biological phenomena such as disease susceptibility and mate choice (Brandies et al., 2018; Sommer, 2005). These phenomena have significant implications on fitness, and as a result, inter-individual variation at MHC loci has been extensively studied across a number of threatened species (Ujvari & Belov, 2011). Studies of MHC and other immune genes demonstrate how characterising genetic variation is crucial to predicting which genes may contribute to variable phenotypes, and the resultant implications for species conservation. However, little is currently known about diversity at other important gene families in threatened species.

Variation at reproductive genes may contribute to key productivity traits that impact the survival of threatened species. Relationships between gene variants and reproductive phenotypes have been extensively studied across a range of model organisms, from *Drosophila* to humans. For example, polymorphisms in male reproductive genes have been associated with variation in sperm competitive ability in *Drosophila* (Fiumera, Dumont & Clark, 2005) and a range of gene mutations have been linked to infertility in humans (Layman, 2002). Associations between variants of key reproductive genes (e.g., those involved in the production or binding of reproductive hormones) and reproductive traits have also been reported in livestock species where high productivity is important (Kirkpatrick, 2002). Examining diversity at genes known to be involved in reproduction is a fundamental first step in determining which loci have the potential to underlie important reproductive traits. However, little is currently known about the variation at reproductive genes in wildlife species, particularly in threatened species that exhibit low levels of genetic diversity overall.

The Tasmanian devil is one such threatened species that is suffering from a range of threatening processes, in addition to having low genome-wide diversity.

Devils are the largest extant carnivorous marsupial and are native to the island state of Tasmania, Australia (Owen & Pemberton, 2005). Populations have declined by up to 80% across this species' range due to a contagious cancer, known as devil facial tumour disease (DFTD). Historical population declines and contemporary habitat fragmentation have resulted in the erosion of genetic diversity (Jones et al., 2004; Miller et al., 2011), particularly at immune gene loci that are highly polymorphic in other species (Cheng et al., 2012; Morris et al., 2015). Tasmanian devils exhibit a number of interesting life-history strategies such the ability of females to undergo up to three oestrous cycles per breeding season (Keeley et al., 2012), the production of up to 30 embryos, of which only 4 can be supported by the 4 teats (Guiler, 1970; Hughes, 1982), precocial breeding (Lachish, McCallum & Jones, 2009; Russell et al., 2019), and multiple paternity litters (Russell et al., 2019). Despite these unique reproductive traits, Tasmanian devils have shown concerning declines in productivity in both captivity (Farquharson, Hogg & Grueber, 2017) and the wild (Farquharson et al., 2018). So, an understanding of whether diversity exists at reproductive genes is a fundamental step in identifying genes that may be associated with differential reproductive phenotypes, and hence may influence reproductive fitness. Armed with this basic knowledge, conservation managers can then use this information in their management decisions pertaining to captive breeding and translocations.

Here, we aimed to identify and characterise reproductive genes, and then examine single nucleotide polymorphism (SNP) diversity at these genes using 37 resequenced Tasmanian devil genomes. We explore signatures of selection to identify polymorphic genes with adaptive potential (i.e., genes where specific alleles may result in differential phenotypes that are beneficial under particular circumstances). The results from this study provide a resource for future research to examine the association between reproductive diversity and productivity in the Tasmanian devil.

Materials and Methods

Gene Identification & Characterisation

In total, 250 genes that have previously been associated with reproduction in mammalian species were selected based on literature searches using the search terms "reproduction" and "gene", as well as mining the human gene database GeneCards (www.genecards.org, Stelzer et al., 2016) using the keyword "reproduction". The identified genes are involved in a variety of reproductive stages including: the hormonal regulation of reproduction, sexual/reproductive development, gametogenesis, fertilisation, and embryogenesis. Predicted complete and partial gene sequences from NCBI's or Ensembl's automatic annotation process were identified in the Tasmanian devil genome reference assembly on NCBI (Devil_ref v7.0 [GCA_000189315.1], Murchison et al., 2012).

Gene predictions in the Tasmanian devil genome were checked using a number of methods including: 1. confirming gene synteny against model organisms (human and mouse) and the current highest-quality marsupial genome (koala) using NCBI's genome viewer (NCBI Resource Coordinators, 2017); 2. mapping the predicted coding sequences (CDS) back to the reference genome using Splign (Kapustin et al., 2008) to ensure all exons were correctly identified and confirm that coding sequences were complete and did not contain any premature stop codons or frameshift mutations; and 3. performing a BLASTP (Altschul et al., 1990) search on the predicted translated sequences against the UniProt (Consortium, 2018) database to confirm identity and protein lengths. For genes with multiple isoforms, the first-named isoform (Variant X1) was investigated (usually the longest). All genes were utilised in downstream analyses.

For partial gene predictions, any missing exons were identified by comparison to well-annotated model organism orthologs using the NCBI genome viewer (NCBI Resource Coordinators, 2017) and TBLASTN (Altschul et al., 1990) searches. Where exons were unable to be fully resolved (i.e., due to gaps in the reference sequence or genome fragmentation etc.) partial sequences were utilised in downstream analyses. For any genes not automatically annotated in the reference genome by NCBI or Ensembl, the predicted location of these genes was identified through gene synteny and TBLASTN searches with model organisms (human and mouse) and gene prediction was performed using FGENESH+ (Solovyev, 2004) with koala orthologs as an input. If an orthologous sequence was not available in koala, human or mouse orthologs were used as an input instead.

Sample Collection and Genome Resequencing

Two existing datasets of resequenced genomes were used to explore reproductive gene diversity in the Tasmanian devil. The first dataset was comprised of twenty-five individuals (including twelve wild-born founders [Figure S1] and 9 parent-offspring trios [Figure S2]) that were sequenced to a high coverage of ~45×

(SRA accessions: SRX6096677- SRX6096696, Wright et al., 2020). The second dataset included twelve wild individuals from a separate wild population (Figure S1) sequenced to a low coverage of 10-15× (SRA accessions: ERS682204-ERS682210; ERS1202857-ERS1202861 Wright et al., 2015; Wright et al., 2017). This low-coverage dataset was only included following the preliminary SNP identification to minimise the risk of this dataset introducing false SNPs. We refer to the twelve low-coverage genomes as "12L" to differentiate it from the dataset encompassing the 25 high-coverage resequenced genomes ("25H").

Preliminary SNP Identification

To identify an initial high-confidence target SNP set, whole-genome alignment and SNP calling was performed on the 25H following the methods in Wright et al. (2020). Briefly, reads were aligned to the Tasmanian devil reference genome assembly version 7.0 (GenBank: GCA_000189315.1, Murchison et al., 2012) using Burrows-Wheeler aligner v 0.7.15 (Li & Durbin, 2009). PCR duplicates were removed with picardtools v1.119 (http://broadinstitute.github.io/picard/) and indel realignment was performed with GATK v3.6 (McKenna et al., 2010). SNPs were called using SAMtools v 1.6 (Li et al., 2009) with minimum base and mapping quality of 30 and a coefficient for downgrading mapping quality for reads containing excessive mismatches of 50. Annovar v 20180416 (Yang & Wang, 2015) gene-based annotation was used to annotate all variants from each of the 25H resequenced genomes aligned to the reference genome using the corresponding genome annotation file from NCBI (O'Leary et al., 2015). Any genes not included in the NCBI annotation were checked for SNPs manually in Geneious (Kearse et al., 2012). SNPs associated with the reproductive genes in the 25H Tasmanian devils were identified by filtering the Annovar output and the total number of each type of SNP (synonymous, nonsynonymous, splicing, UTR5, UTR3, intronic, upstream, downstream) was calculated for each gene. Reproductive genes containing nonsynonymous SNPs were targeted for further analysis. The 12L dataset was not included in the initial SNP identification procedure in order to minimise the risk of false positive SNPs, which may have resulted in inaccurate target gene identification, because SNPs from low-coverage datasets cannot be called as confidently as from higher-coverage data.

Nonsynonymous SNP Confirmation and Analysis

Reproductive genes containing nonsynonymous SNPs were investigated further in both the original 25H resequenced genomes as well as the 12L resequenced genomes. Variants within the target reproductive genes of the 12L resequenced genomes were called together with the 25H resequenced genomes using the same parameters, as above. This method was chosen as multi-sample callers result in the best accuracy when lower coverage samples are called simultaneously with a larger number of higher coverage individuals (Cheng, Teo & Ong, 2014). Individual sample VCF files were then subset from the multisample VCF file and filtered to exclude variants below a filtered depth threshold using BCFtools v1.3.1 (Li et al., 2009). We chose a minimum filtered read depth of 10 for the 25H resequenced genomes and a minimum filtered read depth of five for the 12L resequenced genomes to increase confidence in the variant calls while preventing excessive data loss. The remaining variants were then merged into a multisample VCF file and converted to transposed PLINK format (Purcell et al., 2007) using VCFtools v0.1.14 (Danecek et al., 2011). PLINK v 1.90 was used to calculate minor allele frequencies (MAFs) and determine genotypes for all variants present within the coding regions of the target reproductive genes. Any variants with a MAF below 0.05 that were called in only one individual and had a low allelic depth (below 10), were removed in Geneious. Any positions that were called as variants relative to the reference, but which were monomorphic across the 37 resequenced genomes (i.e., MAF = 0), were also filtered out using GATK and BCFtools. The final variant call files were used to create consensus sequences for each individual using GATK. IUPAC ambiguity codes were used to represent heterozygous positions in the individual consensus sequences and any positions below the specified filtered read depth (as above), or with a missing genotype, were masked. Extraction of CDS for the target genes was performed using bedtools v2.25 (Quinlan & Hall, 2010) with a custom bed file containing the target gene regions and exon positions. Alignments of the CDS were mapped to the reference in Geneious to confirm all synonymous and nonsynonymous SNPs. Missing data/genotyping rate (by locus and individual), MAFs, heterozygosity, and deviations from Hardy-Weinberg equilibrium were calculated for the identified nonsynonymous SNPs in PLINK v1.90 (Purcell et al., 2007). These analyses were performed on all samples and again with the nine known offspring removed to ensure the measures were not influenced by relatedness.

Population Diversity Analysis

CDS alignments of genes confirmed to contain SNPs were converted to PHASE format using SeqPHASE (Flot, 2010). PHASE v2.1 (Stephens & Donnelly, 2003; Stephens, Smith & Donnelly, 2001) was used to construct haplotypes using the original model with default iteration parameters and output probability thresholds (-*p* and -*q*) set to 0. This was performed to ensure any missing SNPs were imputed (based on the distributions of known haplotypes and allele frequencies across the entire dataset, see Stephens & Donnelly, 2003; Stephens et al., 2001) prior to performing the population diversity analysis. The -*x* flag was used to run the algorithm five times (with random seeds for each run) for each gene and the run with the highest goodness-of-fit statistic was selected for the output. SeqPHASE was used to convert the PHASE output files to FASTA format and CERVUS 3.0.7 (Kalinowski, Taper & Marshall, 2007) was used to test whether the phased haplotypes were consistent across the nine trios present in the dataset. DnaSP v6 (Rozas et al., 2017) was used to infer the number of haplotypes (*h*), haplotype diversity (*hd*) and nucleotide diversity per site (π) for each gene. Deviations from the neutral model of molecular evolution were tested using Tajima's *D* (Tajima, 1989) in DnaSP and codon-based *Z*-tests of selection were performed in MEGA7 (Kumar, Stecher & Tamura, 2016) using the Nei-Gojobori method (Nei & Gojobori, 1986) with variance estimated from 500 bootstraps. These statistics were repeated with the nine known offspring excluded to ensure any significant findings were not influenced by relatedness.

Results

Gene Characterisation

Of 250 genes examined, 214 had predicted (complete or partial) CDS (Table S1). These 214 predicted genes were confirmed through analysis of gene synteny, CDS, and BLASTP searches and were investigated in the subsequent SNP analysis. The remaining 36 genes were not automatically annotated by NCBI or Ensembl and could not be identified in the Tasmanian devil genome (Table S2).

SNP identification and Analysis

Using our 25H resequenced genomes, we identified over 5,000 putative SNPs associated with the 214 reproductive genes investigated (Figure 3.1) with an average of 28 putative SNPs per gene (range 0–549) (Table S3). Approximately 90% of these SNPs were intronic (Table S3). Forty-nine genes (23% of all genes investigated) were predicted to contain exonic SNPs, with 34 of these genes predicted to contain at least one nonsynonymous SNP (Table S3). Genes involved in embryogenesis, fertilisation and hormonal regulation of reproduction displayed greater numbers of nonsynonymous SNPs than synonymous SNPs (Figure 3.1).

Confirmation of putative nonsynonymous SNPs was performed by analysing data from the 12L and 25H resequenced genomes together, along with additional filtering (see Methods). After filtering, 33 nonsynonymous SNPs across 20 of the genes remained (Table S4). These 20 genes represented molecular processes across a range of reproductive roles in females, males or both sexes (Table 3.1). For these nonsynonymous SNPs, the genotyping rate (percentage of individuals successfully genotyped at each SNP) was 82% (77% when excluding the nine known offspring) (Table S5). All nonsynonymous SNPs conformed to Hardy-Weinberg expectations (Table S5).

Haplotypes at 18 of the 20 genes were consistent with the known trio information. *DIAPH2* showed inconsistencies in 5 sire-dam-offspring trios (offspring haplotypes were not observed in the parents), possibly due to sequence complexity or particular motifs in this gene region resulting in sequencing difficulty (Nakamura et al., 2011). This gene was excluded from further analysis due to the high error rate (25% of SNPs were inconsistent across the 9 trios). *PIP* showed two occurrences of trio phasing inconsistency but was included in subsequent analysis due to the low error rate (2.2% of SNPs were inconsistent across the 9 trios). This resulted in 19 final genes (following exclusion of *DIAPH2*) that were included in subsequent population diversity analysis.

The total number of SNPs (both synonymous and nonsynonymous) in the coding regions of each of the 19 final genes across the 37 resequenced genomes ranged from 1 to 10; number of haplotypes per gene ranged from 2 to 4 (Table 3.2). Mean haplotype diversity was 0.36 (SD 0.20) and mean nucleotide diversity was 4.3 x 10-4 (SD 5.4 x 10-4) (Table 3.2).

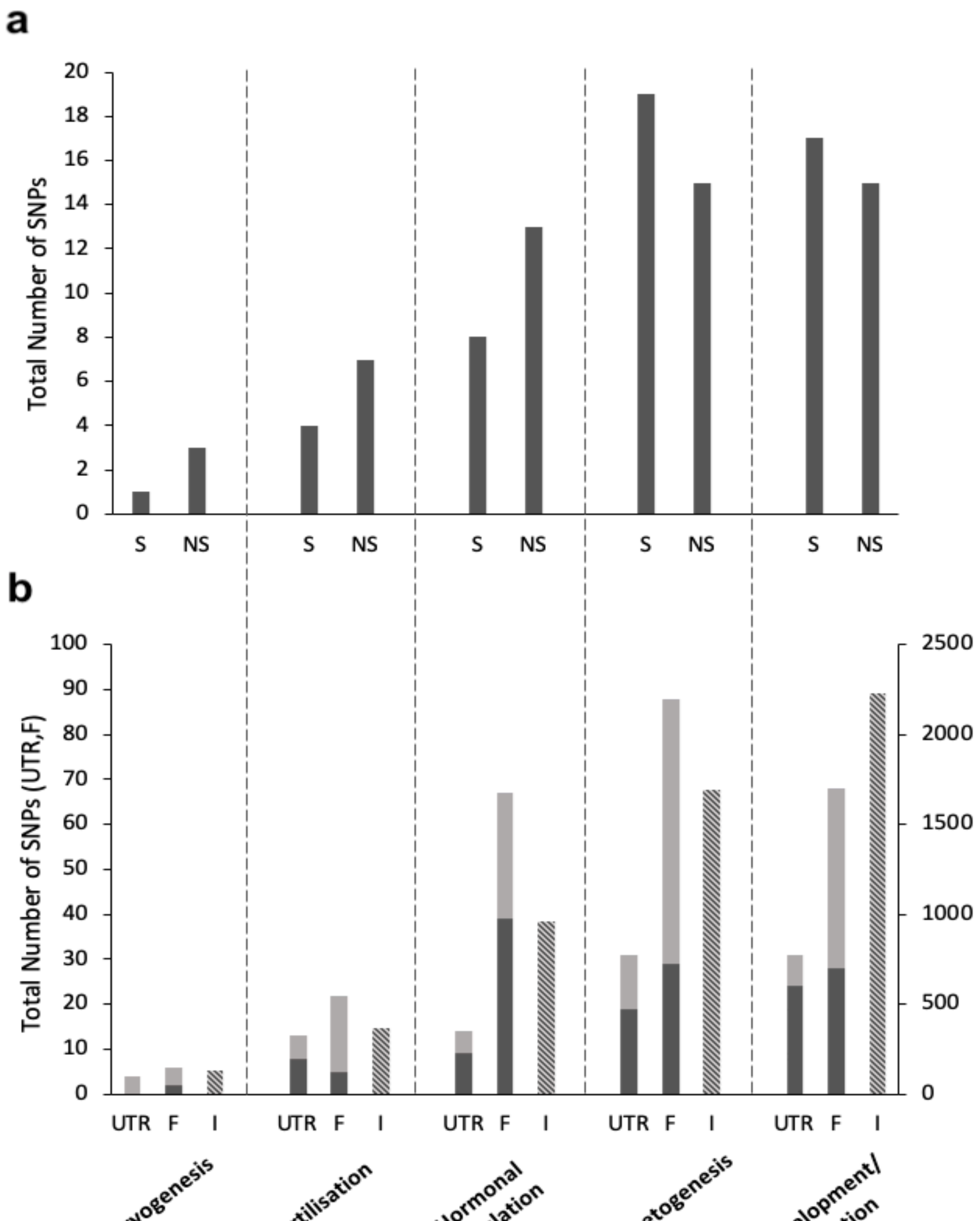


Figure 3.1 Total number of SNPs identified in genes known or predicted to be involved in a variety of reproductive functions including embryogenesis (N = 13 genes), fertilisation (N = 26), hormonal regulation of reproduction (N = 43), gametogenesis (N = 74), and general reproductive development and function (N = 58). **a)** Exonic SNPs including synonymous (S) and nonsynonymous (NS). **b)** Other major SNP types including untranslated regions (UTR), flanking regions (F) and intronic regions (I). Stripes indicate intronic SNPs are plotted on the secondary axis. Light shading indicates SNPs that are 5' (upstream), dark shading indicates SNPs that are 3' (downstream). See Table S3 for more information.

Table 3.1 Reproductive roles of genes found to contain nonsynonymous SNPs.

Gene	Role in Reproduction	Sex affected	Ref
ADAMTS9	Important in uterine remodelling of implantation, placentation and parturition	Female	Russell, Brown, & Dunning, 2015
ADAMTS10	Important for adhesion between the sperm and egg zona pellucida	Male	Dun et al., 2012
ADAMTSL1	Involved in embryonic gonadogenesis	Female	Carré, Couty, Hennequet-Antier, & Govoroun, 2011
AIRE	Mutations result in autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) which can lead to infertility	Both	Aaltonen et al., 1997
BMP5	Predicted to play a role in ovarian folliculogenesis	Female	Pierre, Pisselet, Dupont, Bontoux, & Monget, 2005
CHD7	Mutations result in CHARGE syndrome (pubertal failure and infertility)	Both	Kim et al., 2008
CLU	Increased expression results in reduced sperm quality and infertility	Male	Zalata et al., 2012
CYP19A1	A key enzyme in oestrogen biosynthesis and influences female fertility	Female	Simpson et al., 1994; Altmäe et al., 2009
DIAPH2	Important for normal ovarian development and function	Female	Bione et al., 1998
DZIP1	Regulator of hedgehog signalling and may participate in spermatogenesis via its interaction with DAZ	Male	Moore, Jaruzelska, Dorfman, & Reijo-Pera, 2004; Sekimizu et al., 2004

IRS4	Null mutations can lead to defects in reproduction	Both	Fantin, Wang, Lienhard, & Keller, 2000
KIT	Plays a key role in germ cell development, spermatogenesis and oogenesis	Both	Rossi, 2013; Russell, Brown, & Dunning, 2015
LEP	Deficiencies can lead to hypogonadotrophic hypogonadism and infertility	Both	Chehab, Lim, & Lu, 1996
NANOG	Transcription regulator important for embryonic stem cell pluripotency	Both	Pan & Thomson, 2007
PIP	Functions in seminal fluid, important for fertilisation	Male	Hassan, Waheed, Yadav, Singh, & Ahmad, 2009
PRDM14	Required for the proper initiation and coordination of the primordial germ cell specific gene expression program and promotes pluripotency	Both	Hohenauer & Moore, 2012
PTCH1	Mediates hedgehog signalling in developing and adult marsupial gonads	Both	O'Hara, Azar, Behringer, Renfree, & Pask, 2011
PTCH2	Mediates hedgehog signalling in developing and adult marsupial gonads	Both	O'Hara, Azar, Behringer, Renfree, & Pask, 2011
PTGFRN	Inhibitor of the Prostaglandin F2 Receptor which has multiple roles in reproduction e.g., progesterone synthesis and ovulation	Female	Craig, 1975
SPACA6	Involved in sperm-oocyte fusion - gene knockouts result in failed fusion	Male	Lorenzetti et al., 2014

Table 3.2 Diversity statistics and neutrality tests performed on the target reproductive genes.

Gene	*n*	CDS Length	SNPs (ns:s)	*h*	*hd*	π	Tajima's D	Z-test
ADAMTS9	74	5919	9 (1:8)	4	0.666	7.32	3.52***	-2.63**
ADAMTS10	74	3342	1 (1:0)	2	0.104	0.31	-0.60	0.98
ADAMTSL1	74	5298	1 (1:0)	2	0.294	0.55	0.53	1.00
AIRE	74	1590	10 (4:6)	4	0.451	21.71	1.83	-1.73
BMP5	74	1368	1 (1:0)	2	0.053	0.39	-0.90	1.04
CHD7	74	9093	3 (3:0)	3	0.586	1.15	1.34	1.27
CLU	74	1178	2 (2:0)	3	0.445	4.06	0.27	1.29
CYP19A1	74	1512	1 (1:0)	2	0.053	0.35	-0.90	1.07
DZIP1	74	2433	3 (1:2)	3	0.283	3.08	0.41	-1.26
IRS4	74	2751	1 (1:0)	2	0.053	0.19	-0.90	1.06
KIT	74	2901	3 (2:1)	4	0.545	3.76	1.47	-0.67
LEP	74	504	1 (1:0)	2	0.217	4.30	0.07	1.04
NANOG	74	936	2 (2:0)	2	0.494	10.55	2.30*	1.01
PIP	74	534	3 (3:0)	4	0.588	12.54	0.17	0.16
PRDM14	74	1662	2 (1:1)	2	0.217	2.61	0.09	-0.69
PTCH1	74	3891	1 (1:0)	2	0.344	0.88	0.82	1.01
PTCH2	74	4524	3 (3:0)	3	0.527	2.34	1.37	1.50
PTGFRN	74	2892	2 (2:0)	3	0.416	1.54	0.14	1.40
SPACA6	74	1122	1 (1:0)	2	0.462	4.12	1.53	1.02

n, number of sequences (2 allele sequences per individual); h, number of inferred haplotypes; hd, haplotype diversity; π, nucleotide diversity (x10^4)

*p < 0.05. Did not remain significant after Holm-Bonferroni multiple tests correction.

**p < 0.01. Did not remain significant after Holm-Bonferroni multiple tests correction.

***p < 0.001. Remained significant after Holm-Bonferroni multiple tests correction.

ADAMTS9 and *NANOG* showed statistically significant deviation from neutrality at the sequence level with positive Tajima's *D* values suggesting population decline or balancing selection (Table 3.2). *ADAMTS9* also showed evidence of purifying selection at the codon level with a statistically significant negative *Z*-test ($p < 0.01$; Table 3.2). There were no qualitative changes to the results when the nine known offspring were excluded from the analyses (Table S6).

Discussion

As wildlife populations continue to decline globally, understanding the genetic basis of inter-individual variation is crucial for determining which genes may govern important phenotypes and contribute to species' long-term survival and fitness. Here we show how genomic data can be used to explore functional genetic diversity in an endangered species. This study identified a surprising amount of putatively functional variation at reproductive genes in an otherwise genetically depauperate species. Tasmanian devils have shown concerning declines in productivity over time in both captivity (Farquharson, Hogg & Grueber, 2017) and the wild (Farquharson et al., 2018). It is predicted that genetic variation may play a role in such changes (Farquharson, Hogg & Grueber, 2017; Gooley et al., 2020), although until now there was limited knowledge of whether diversity even exists at their reproductive genes.

We characterised genetic variation at 214 reproductive genes in 37 Tasmanian devils and identified 5,933 putative SNPs. Signatures of selection were examined at a subset of 19 target genes that contained nonsynonymous variation, and hence may have functional consequences for reproduction. To the best of our knowledge, this is the first study to examine within-species reproductive gene diversity to this extent in a threatened species.

Tasmanian devils exhibit very low levels of genetic diversity overall (Cheng et al., 2012; Jones et al., 2004; Miller et al., 2011; Morris et al., 2015). Most (77%) of the reproductive genes we examined had monomorphic coding regions in our sample set of 37 resequenced genomes; a low level of diversity that is comparable to that seen in a previous study which examined genetic diversity at 167 immune genes in ten Tasmanian devils (7 of which were included in the current study) (Morris et al 2015). However, within those reproductive genes that showed nonsynonymous variation, we found surprisingly high diversity relative to a similar subset of immune genes that also contained nonsynonymous SNPs (Morris et al., 2015). For example, despite a much larger sample size of up to 196 individuals across multiple captive and wild populations (with majority of individuals presumed to be unrelated), Morris et al. (2015) found a maximum of 3 SNPs per gene across nine polymorphic immune genes, compared with a maximum of 10 SNPs per reproductive gene here (across the final 19 polymorphic reproductive genes). Mean haplotype diversity was also higher in the current study. Differences in sample origin may contribute to the observed increased levels of diversity herein; however, the finding of higher genetic diversity at reproductive genes compared with immune genes is unexpected given the smaller sample size and presence of related individuals within the current study. We note that Morris et al. (2015) used amplicon sequencing to confirm SNP diversity in the subset of target genes, which resulted in fewer SNPs than predicted by genome resequencing data. Although we did not employ gene-targeted sequencing methods in this study, we believe that the SNPs identified are likely to reflect real diversity, not sequencing artefacts, due to the number of resequenced genomes (particularly those with high coverage, around 45×) and the strict variant calling and filtering parameters employed.

Thirty-six reproductive genes (14% of all genes investigated) present in model species could not be characterised in the Tasmanian devil genome by the methods applied here. For example, there were no TBLASTN hits for a number of genes including *DPPA3/STELLA*, *SEMG1*, *SEMG2*, *TNP2* and *PRM2*, which are either too divergent from known orthologs to be identified by this method, or do not exist in marsupials (Johnson et al., 2018). Additionally, members of the *NLRP* (Nucleotide-binding oligomerisation domain, leucine rich repeat and pyrin domain containing proteins) gene family have shown extensive duplication and diversification in mammalian lineages (Tian, Pascal & Monget, 2009) and were unable to be identified in the Tasmanian devil genome. Fragmentation and gaps in the current reference genome precluded characterising a number of genes such as *KLK3*, *ZPBP* and others (Table S2). Genes located on the Y chromosome (e.g., *ATRY*, *DAZ1*, *USP9Y* and *DDX3Y*) were unable to be identified due to the unavailability of Y-chromosome data in the female reference genome. Sequencing the Y chromosome will be important in the future to focus on male reproduction, as a number of important male reproductive genes are found on the Y (Murtagh, Waters & Graves, 2010; Toder, Wakefield & Graves, 2000).

Twenty genes were found to contain nonsynonymous SNPs in the current study (with *DIAPH2* later excluded due to phasing inconsistencies). Since nonsynonymous mutations result in amino acid changes, genes that contain nonsynonymous SNPs may influence phenotype (Shastry, 2009). Although other SNPs, such as synonymous polymorphisms or variants outside the coding sequence, may contribute to phenotype via processes such as mRNA stability (Chamary & Hurst, 2005), these are expected to have a weaker effect on gene function compared with mutations that alter the protein sequence (Tomoko, 1995). The genes found to contain nonsynonymous SNPs in the current study are involved in a variety of reproductive functions in both males and females, and influence fertility-associated phenotypes in humans and other species (see Table 3.1 for more information). For example, mutations in the *CHD7* gene cause idiopathic hypogonadotropic hypogonadism and Kallmann syndrome in humans, resulting in impaired sexual development in both males and females (Kim et al., 2008). Mutations in the *AIRE* gene cause autoimmune polyendocrinopathy, candidiasis and ectodermal dystrophy (APECED) (Aaltonen et al., 1997) which has also been linked to infertility in both men and women (Perheentupa, 2006). *ADAMTS* proteases influence a range of reproductive processes in humans and mice (Russell, Brown & Dunning, 2015), three of which (*ADAMTS9*, *ADAMTS10* and *ADAMTSL1*) displayed nonsynonymous variation in the current study.

The majority of the individuals in our sample set are known to have successfully reproduced based on breeding records in captive facilities (Figure S2), so most of the nonsynonymous SNPs identified in the current study are unlikely to cause the extreme infertile phenotypes that have been reported in humans and mice. However, these variants may result in more subtle phenotypic effects such as reduced fertilisation success or reduced offspring survival etc. We note that a number of non-synonymous homozygoyte genotypes were not observed in our dataset (Table S5). They may encode more severe phenotypes, which could be associated with pregnancy loss or infertility and may exist in a larger sample set or could potentially be lethal and hence never appear in homozygous form. Further research is required to explore the functional consequences of the identified nonsynonymous variants herein. Interestingly, we found that genes involved in embryogenesis, fertilisation and hormonal regulation of reproduction displayed greater numbers of nonsynonymous SNPs than synonymous SNPs. This suggests that functional diversity may be important at genes involved in such processes. Tasmanian devils exhibit a number of unique reproductive characteristics including undergoing up to three oestrous cycles within their annual breeding season (Keeley et al., 2012); producing a greater number of embryos (up to 30) than can be supported by their four teats (Guiler, 1970; Hughes, 1982); and multiple paternity litters (Russell et al., 2019) even though mate-guarding is a behavioural reproductive strategy (Hamilton et al., 2019). We hypothesise that these unique reproductive traits may drive functional diversity across genes involved in particular reproductive processes through adaptive evolution. For example, multiple mating by females is known to drive sperm competition which may result in selective pressures on genes involved in fertilisation (Dapper & Wade, 2016; Fiumera, Dumont

& Clark, 2005). Similarly, fitness advantages associated with the timing or number of oestrous cycles, or the number of viable embryos, could potentially drive natural selection at genes involved in the hormonal regulation of reproduction or embryogenesis respectively. To explore these ideas further we investigated signatures of selection at the reproductive genes containing nonsynonymous SNPs.

Of the 19 final reproductive genes (following exclusion of *DIAPH2* due to phasing inconsistencies), two genes (*ADAMTS9* and *NANOG*) showed statistically significant signatures of selection, suggesting their variants may be linked to important phenotypic traits. After correcting for multiple testing using the Holm-Bonferroni method (Holm, 1979), the Tajima's D for *ADAMTS9* remained statistically significant, indicating that this gene may be under balancing selection at the sequence level within the population. Demographic factors such as population bottlenecks can contribute to the value of Tajima's D (Tajima, 1989), however demographic factors are likely to affect loci across the whole genome. Since similar patterns of selection were not observed across all of the target loci, we hypothesise that *ADAMTS9* may be a candidate for long-term balancing selection. Balancing selection actively maintains multiple alleles in a population, suggesting that the associated phenotypes may be advantageous under certain circumstances (e.g., Gos, Slotte, & Wright, 2012). *ADAMTS9* is a pleiotropic gene that belongs to a large, diversified family of *ADAMTS* genes and has been implicated in several crucial female reproductive processes, namely: ovulation, implantation, placentation and parturition (Russell, Brown & Dunning, 2015). *ADAMTS9* is also a novel tumour suppressor (Du et al., 2013) and has undergone strong selection for increased longevity in a number of small-bodied mammal lineages (Lambert & Portfors, 2017). This is particularly interesting in our context, as Tasmanian devils have a short lifespan (maximum 5 years in the wild) in comparison to other mammals of their size and show unusually high vulnerability to tumours (Griner, 1979). Although our data cannot disentangle whether selection on the *ADAMTS9* gene in Tasmanian devils may be attributed to that gene's role in reproduction and/or its role in tumour suppression and longevity. The attributes of this gene, such as its role in a number of key processes, make it a plausible candidate for adaptation and warrants further investigation.

The *NANOG* gene also showed a putative pattern of balancing selection in the Tasmanian devil, though this result did not remain statistically significant after correcting for multiple testing. *NANOG* is a key transcription factor involved in embryonic stem cell pluripotency (Pan & Thomson, 2007). We identified a multi-nucleotide nonsynonymous polymorphism within the coding sequence of *NANOG*. It is currently unknown whether these variants are associated with differential phenotypes. Investigations into whether the identified nonsynonymous SNP is correlated with embryonic survival traits and may influence reproductive success within the Tasmanian devil are required.

Although reproductive genes and their variants have been well studied in model and livestock species (see Hunt et al., 2018), there is little data on reproductive variants in threatened species, many of which typically show low overall levels of genome-wide diversity. As a result, it is difficult to ascertain whether Tasmanian devil reproductive gene diversity is higher or lower than expected compared to other threatened species. Furthermore, our study focused on a relatively small sample set from a limited number of locations in Tasmania and may not have captured the true extent of genetic diversity across the species' range. As whole genome sequencing technology becomes cheaper with time, sampling Tasmanian devils across their range would improve our understanding of their reproductive gene diversity. The full benefit of understanding reproductive gene diversity in Tasmanian devils can be realised by studying the relationship between genetic variation and reproductive phenotypes. For this threatened species, this is possible as the Tasmanian devil insurance population is Australia's largest captive breeding program (Hogg et al., 2019) with a large number of individuals across multiple generations with DNA samples and extensive reproductive records. This resource will allow us to investigate diversity across a range of candidate genes to determine whether variation in reproductive genes influences reproductive fitness. For example, the *SPACA6* gene has been implicated in fertilisation ability of male mice (Lorenzetti et al., 2014) and was found to contain a nonsynonymous SNP among the sampled Tasmanian devils in the current study. By sequencing the *SPACA6* gene across hundreds of male Tasmanian devils using specific PCR primers, or a targeted capture approach, we could statistically determine whether this variant is correlated with an individual's siring ability. Candidate gene approaches have several advantages over whole-genome approaches, namely the higher inherent statistical power and reduced sequencing costs. However, it is possible that other genes or genomic regions that influence reproductive phenotypes may be missed and so a genome-wide association study (GWAS) may be more informative (for a review of candidate gene vs GWAS approaches see Suh & Vijg, 2005). A combination of these approaches will likely be the best way forward to understanding the interplay between reproductive genotype and phenotype. The rise of whole genome sequencing and global consortia developing reference genomes for wildlife means that our understanding of functional gene diversity in a range of threatened species can only improve with time, particularly in those species where range reduction and population contraction has led them to be genetically depauperate. The approach used in this study demonstrates how these growing genomic resources can be utilised to explore functional diversity in threatened species and how this information can assist with their conservation management.

Conclusion

Our study has bioinformatically characterised diversity at 219 reproductive genes in 37 Tasmanian devils. We have identified and examined diversity at 19 polymorphic genes containing nonsynonymous SNPs that may have functional consequences on reproduction. The results from this study provide the foundation for future research to explore whether any of these genes are associated with variable reproductive phenotypes and hence may be involved in the generational productivity declines that have been observed in the Tasmanian devil insurance population (Farquharson, Hogg & Grueber, 2017; Hogg et al., 2015). If specific genotypes are found to influence productivity, preserving the functional variation described herein may be key to minimising these declines and facilitating the success of conservation breeding programs. Beyond assisting with conservation decisions for the Tasmanian devil the candidate gene approach described here may also be applied to reproductive management in other threatened species conservation programs.

Acknowledgements

This work was funded by the Australian Research Council (LP140100508; DP170101253) and supported by the Save the Tasmanian Devil Program, the Zoo and Aquarium Association Australasia and San Diego Zoo Global. We thank Professor Graham Wallis for insightful discussions on methodology and general themes of the paper. Dr Camilla Whittington provided helpful suggestions of genes important to reproduction. We thank Drew Lee for creating the supplemental map overlay. Computing resources were provided by the Sydney Informatics Hub and the University of Sydney's high-performance computing cluster Artemis.

References

Aaltonen, J, Björses, P, Perheentupa, J, Horelli–Kuitunen, N, Palotie, A, Peltonen, L, Lee, YS, Francis, F, Henning, S & Thiel, C 1997, 'An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains', *Nature Genetics,* vol. 17, no. 4, pp. 399-403.

Altmäe, S, Haller, K, Peters, M, Saare, M, Hovatta, O, Stavreus-Evers, A, Velthut, A, Karro, H, Metspalu, A & Salumets, A 2009, 'Aromatase gene (CYP19A1) variants, female infertility and ovarian stimulation outcome: a preliminary report', *Reproductive Biomedicine Online,* vol. 18, no. 5, pp. 651-657.

Altschul, SF, Gish, W, Miller, W, Myers, EW & Lipman, DJ 1990, 'Basic local alignment search tool', *Journal of Molecular Biology,* vol. 215, no. 3, pp. 403-410.

Bione, S, Sala, C, Manzini, C, Arrigo, G, Zuffardi, O, Banfi, S, Borsani, G, Jonveaux, P, Philippe, C & Zuccotti, M 1998, 'A human homologue of the *Drosophila melanogaster* diaphanous gene is disrupted in a patient with premature ovarian failure: evidence for conserved function in oogenesis and implications for human sterility', *The American Journal of Human Genetics,* vol. 62, no. 3, pp. 533-541.

Brandies, PA, Grueber, CE, Hogg, CJ & Belov, K 2018, 'MHC Genes and Mate Choice', in Choe, J (ed.), *Encyclopedia of Animal Behaviour,* 2 edn, Elsevier, Massachusetts, USA.

Carré, G-A, Couty, I, Hennequet-Antier, C & Govoroun, MS 2011, 'Gene expression profiling reveals new potential players of gonad differentiation in the chicken embryo', *PLoS One,* vol. 6, no. 9, pp. e23959.

Chamary, J & Hurst, LD 2005, 'Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals', *Genome Biology,* vol. 6, no. 9, pp. R75.

Chehab, FF, Lim, ME & Lu, R 1996, 'Correction of the sterility defect in homozygous obese female mice by treatment with the human recombinant leptin', *Nature Genetics,* vol. 12, no. 3, pp. 318-320.

Cheng, AY, Teo, Y & Ong, RT 2014, 'Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals', *Bioinformatics,* vol. 30, no. 12, pp. 1707-1713.

Cheng, Y, Sanderson, C, Jones, M & Belov, K 2012, 'Low MHC class II diversity in the Tasmanian devil (*Sarcophilus harrisii*)', *Immunogenetics,* vol. 64, no. 7, pp. 525-533.

Consortium, U 2018, 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids Research,* vol. 47, no. D1, pp. D506-D515.

Craig, GM 1975, 'Prostaglandins in reproductive physiology', *Postgraduate Medical Journal,* vol. 51, no. 592, pp. 74-84.

Danecek, P, Auton, A, Abecasis, G, Albers, CA, Banks, E, Depristo, MA, Handsaker, RE, Lunter, G, Marth, GT & Sherry, ST 2011, 'The variant call format and VCFtools', *Bioinformatics,* vol. 27, no. 15, pp. 2156-2158.

Dapper, AL & Wade, MJ 2016, 'The evolution of sperm competition genes: the effect of mating system on levels of genetic variation within and between species', *Evolution,* vol. 70, no. 2, pp. 502-511.

Du, W, Wang, S, Zhou, Q, Li, X, Chu, J, Chang, Z, Tao, Q, Ng, E, Fang, J & Sung, J 2013, 'ADAMTS9 is a functional tumor suppressor through inhibiting AKT/mTOR pathway and associated with poor survival in gastric cancer', *Oncogene,* vol. 32, no. 28, pp. 3319-3328.

Dun, M, Anderson, A, Bromfield, E, Asquith, K, Emmett, B, McLaughlin, E, Aitken, R & Nixon, B 2012, 'Investigation of the expression and functional significance of the novel mouse sperm protein, a disintegrin and metalloprotease with thrombospondin type 1 motifs number 10 (ADAMTS10)', *International Journal of Andrology,* vol. 35, no. 4, pp. 572-589.

Fantin, VR, Wang, Q, Lienhard, GE & Keller, SR 2000, 'Mice lacking insulin receptor substrate 4 exhibit mild defects in growth, reproduction, and glucose homeostasis', *American Journal of Physiology-Endocrinology And Metabolism,* vol. 278, no. 1, pp. E127-E133.

Farquharson, KA, Gooley, RM, Fox, S, Huxtable, SJ, Belov, K, Pemberton, D, Hogg, CJ & Grueber, CE 2018, 'Are any populations 'safe'? Unexpected reproductive decline in a population of Tasmanian devils free of devil facial tumour disease', *Wildlife Research,* vol. 45, no. 1, pp. 31-37.

Farquharson, KA, Hogg, CJ & Grueber, CE 2017, 'Pedigree analysis reveals a generational decline in reproductive success of captive Tasmanian devil (*Sarcophilus harrisii*): implications for captive management of threatened species', *Journal of Heredity,* vol. 108, no. 5, pp. 488-495.

Fiumera, AC, Dumont, BL & Clark, AG 2005, 'Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins', *Genetics,* vol. 169, no. 1, pp. 243-257.

Flot, JF 2010, 'SeqPHASE: a web tool for interconverting PHASE input/output files and FASTA sequence alignments', *Molecular Ecology Resources,* vol. 10, no. 1, pp. 162-166.

Forsman, A & Wennersten, L 2016, 'Inter-individual variation promotes ecological success of populations and species: Evidence from experimental and comparative studies', *Ecography,* vol. 39, no. 7, pp. 630-648.

Gooley, RM, Hogg, CJ, Fox, S, Pemberton, D, Belov, K & Grueber, CE 2020, 'Inbreeding depression in one of the last DFTD-free wild populations of Tasmanian devils', *PeerJ,* vol. 8, pp. e9220.

Gos, G, Slotte, T & Wright, SI 2012, 'Signatures of balancing selection are maintained at disease resistance loci following mating system evolution and a population bottleneck in the genus *Capsella*', *BMC Evolutionary Biology,* vol. 12, no. 152, pp. 1-10.

Griner, L 1979, 'Neoplasms in Tasmanian devils (*Sarcophilus harrisii*)', *Journal of the National Cancer Institute,* vol. 62, no. 3, pp. 589-595.

Guiler, E 1970, 'Observations on the Tasmanian devil, *Sarcophilus harrisii* (Marsupialia: Dasyuridae) II. Reproduction, breeding and growth of pouch young', *Australian Journal of Zoology,* vol. 18, no. 1, pp. 63-70.

Hamilton, DG, Jones, ME, Cameron, EZ, McCallum, H, Storfer, A, Hohenlohe, PA & Hamede, RK 2019, 'Rate of intersexual interactions affects injury likelihood in Tasmanian devil contact networks', *Behavioral Ecology,* vol. 30, no. 4, pp. 1087-1095.

Hassan, MI, Waheed, A, Yadav, S, Singh, T & Ahmad, F 2009, 'Prolactin inducible protein in cancer, fertility and immunoregulation: structure, function and its clinical implications', *Cellular and Molecular Life Sciences,* vol. 66, no. 3, pp. 447-459.

Hogg, CJ, Fox, S, Pemberton, D & Belov, K 2019, *Saving the Tasmanian Devil*, CSIRO Publishing, Clayton South, VIC, Australia.

Hogg, CJ, Ivy, JA, Srb, C, Hockley, J, Lees, C, Hibbard, C & Jones, M 2015, 'Influence of genetic provenance and birth origin on productivity of the Tasmanian devil insurance population', *Conservation Genetics,* vol. 16, no. 6, pp. 1465-1473.

Hohenauer, T & Moore, AW 2012, 'The Prdm family: expanding roles in stem cells and development', *Development,* vol. 139, no. 13, pp. 2267-2282.

Holderegger, R, Kamm, U & Gugerli, F 2006, 'Adaptive vs. neutral genetic diversity: implications for landscape genetics', *Landscape Ecology,* vol. 21, no. 6, pp. 797-807.

Holm, S 1979, 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics,* vol. 6, no. 2, pp. 65-70.

Hughes, R 1982, 'Reproduction in the Tasmanian devil *Sarcophilus harrisii* (Dasyuridae, Marsupialia)', *Carnivorous Marsupials,* vol. 1, pp. 49-63.

Hunt, SE, McLaren, W, Gil, L, Thormann, A, Schuilenburg, H, Sheppard, D, Parton, A, Armean, IM, Trevanion, SJ & Flicek, P 2018, 'Ensembl variation resources', *Database,* vol. 2018, no. 2018, pp. bay119.

Johnson, RN, O'Meally, D, Chen, Z, Etherington, GJ, Ho, SYW, Nash, WJ, Grueber, CE, Cheng, Y, Whittington, CM, Dennison, S, Peel, E, Haerty, W, O'Neill, RJ, Colgan, D, Russell, TL, Alquezar-Planas, DE, Attenbrow, V, Bragg, JG, Brandies, PA, Chong, AY-Y, Deakin, JE, Di Palma, F, Duda, Z, Eldridge, MDB, Ewart, KM, Hogg, CJ, Frankham, GJ, Georges, A, Gillett, AK, Govendir, M, Greenwood, AD, Hayakawa, T, Helgen, KM, Hobbs, M, Holleley, CE, Heider, TN, Jones, EA, King, A, Madden, D, Graves, JaM, Morris, KM, Neaves, LE, Patel, HR, Polkinghorne, A, Renfree, MB, Robin, C, Salinas, R, Tsangaras, K, Waters, PD, Waters, SA, Wright, B, Wilkins, MR, Timms, P & Belov, K 2018, 'Adaptation and conservation insights from the koala genome', *Nature Genetics,* vol. 50, no. 8, pp. 1102-1111.

Jones, ME, Paetkau, D, Geffen, E & Moritz, C 2004, 'Genetic diversity and population structure of Tasmanian devils, the largest marsupial carnivore', *Molecular Ecology,* vol. 13, no. 8, pp. 2197-2209.

Kalinowski, ST, Taper, ML & Marshall, TC 2007, 'Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment', *Molecular Ecology,* vol. 16, no. 5, pp. 1099-1106.

Kapustin, Y, Souvorov, A, Tatusova, T & Lipman, D 2008, 'Splign: algorithms for computing spliced alignments with identification of paralogs', *Biology Direct,* vol. 3, no. 20, pp. 1-13.

Kearse, M, Moir, R, Wilson, A, Stones-Havas, S, Cheung, M, Sturrock, S, Buxton, S, Cooper, A, Markowitz, S & Duran, C 2012, 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics,* vol. 28, no. 12, pp. 1647-1649.

Keeley, T, O'Brien, J, Fanson, B, Masters, K & McGreevy, P 2012, 'The reproductive cycle of the Tasmanian devil (*Sarcophilus harrisii*) and factors associated with reproductive success in captivity', *General and Comparative Endocrinology,* vol. 176, no. 2, pp. 182-191.

Kim, H-G, Kurth, I, Lan, F, Meliciani, I, Wenzel, W, Eom, SH, Kang, GB, Rosenberger, G, Tekin, M & Ozata, M 2008, 'Mutations in CHD7, encoding a chromatin-remodeling protein, cause idiopathic hypogonadotropic hypogonadism and Kallmann syndrome', *The American Journal of Human Genetics,* vol. 83, no. 4, pp. 511-519.

Kirkpatrick, B 'QTL and candidate gene effects on reproduction in livestock: progress and prospects', Proceedings of the 7th Word Congress on Genetics Applied to Livestock Production, August 19-23, Montpellier, France, Paper 18-01.

Kumar, S, Stecher, G & Tamura, K 2016, 'MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets', *Molecular Biology and Evolution,* vol. 33, no. 7, pp. 1870-1874.

Lachish, S, McCallum, H & Jones, M 2009, 'Demography, disease and the devil: life-history changes in a disease-affected population of Tasmanian devils (*Sarcophilus harrisii*)', *Journal of Animal Ecology,* vol. 78, no. 2, pp. 427-436.

Lambert, MJ & Portfors, CV 2017, 'Adaptive sequence convergence of the tumor suppressor ADAMTS9 between small-bodied mammals displaying exceptional longevity', *Aging,* vol. 9, no. 2, pp. 573-582.

Layman, LC 2002, 'Human gene mutations causing infertility', *Journal of Medical Genetics,* vol. 39, no. 3, pp. 153-161.

Li, H & Durbin, R 2009, 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics,* vol. 25, no. 14, pp. 1754-1760.

Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G & Durbin, R 2009, 'The sequence alignment/map format and SAMtools', *Bioinformatics,* vol. 25, no. 16, pp. 2078-2079.

Lorenzetti, D, Poirier, C, Zhao, M, Overbeek, PA, Harrison, W & Bishop, CE 2014, 'A transgenic insertion on mouse chromosome 17 inactivates a novel immunoglobulin superfamily gene potentially involved in sperm–egg fusion', *Mammalian Genome,* vol. 25, no. 3-4, pp. 141-148.

McKenna, A, Hanna, M, Banks, E, Sivachenko, A, Cibulskis, K, Kernytsky, A, Garimella, K, Altshuler, D, Gabriel, S & Daly, M 2010, 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research,* vol. 20, no. 9, pp. 1297-303.

Miller, W, Hayes, VM, Ratan, A, Petersen, DC, Wittekindt, NE, Miller, J, Walenz, B, Knight, J, Qi, J, Zhao, F, Wang, Q, Bedoya-Reina, OC, Katiyar, N, Tomsho, LP, Kasson, LM, Hardie, R-A, Woodbridge, P, Tindall, EA, Bertelsen, MF, Dixon, D, Pyecroft, S, Helgen, KM, Lesk, AM, Pringle, TH, Patterson, N, Zhang, Y, Kreiss, A, Woods, GM, Jones, ME & Schuster, SC 2011, 'Genetic diversity and population structure of the endangered marsupial *Sarcophilus harrisii* (Tasmanian devil)', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 108, no. 30, pp. 12348-12353.

Mimura, M, Yahara, T, Faith, DP, Vázquez-Domínguez, E, Colautti, RI, Araki, H, Javadi, F, Núñez-Farfán, J, Mori, AS & Zhou, S 2017, 'Understanding and monitoring the consequences of human impacts on intraspecific variation', *Evolutionary Applications,* vol. 10, no. 2, pp. 121-139.

Moore, FL, Jaruzelska, J, Dorfman, DM & Reijo-Pera, RA 2004, 'Identification of a novel gene, DZIP (DAZ-interacting protein), that encodes a protein that interacts with DAZ (deleted in azoospermia) and is expressed in embryonic stem cells and germ cells', *Genomics,* vol. 83, no. 5, pp. 834-843.

Morris, KM, Wright, B, Grueber, CE, Hogg, C & Belov, K 2015, 'Lack of genetic diversity across diverse immune genes in an endangered mammal, the Tasmanian devil (*Sarcophilus harrisii*)', *Molecular Ecology,* vol. 24, no. 15, pp. 3860-3872.

Murchison, EP, Schulz-Trieglaff, OB, Ning, Z, Alexandrov, LB, Bauer, MJ, Fu, B, Hims, M, Ding, Z, Ivakhno, S & Stewart, C 2012, 'Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer', *Cell,* vol. 148, no. 4, pp. 780-791.

Murtagh, VJ, Waters, PD & Marshall Graves, JA 2010, 'Compact but Complex – The Marsupial Y Chromosome', in Deakin, J, Waters, P & Marshall Graves, J (eds.), *Marsupial Genetics and Genomics*, Springer, Dordrecht.

Nakamura, K, Oshima, T, Morimoto, T, Ikeda, S, Yoshikawa, H, Shiwa, Y, Ishikawa, S, Linak, MC, Hirai, A & Takahashi, H 2011, 'Sequence-specific error profile of Illumina sequencers', *Nucleic Acids Research,* vol. 39, no. 13, pp. e90.

NCBI Resource Coordinators 2017, 'Database resources of the national center for biotechnology information', *Nucleic acids research,* vol. 45, no. D1, pp. D12-D17.

Nei, M & Gojobori, T 1986, 'Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions', *Molecular Biology and Evolution,* vol. 3, no. 5, pp. 418-426.

O'Hara, WA, Azar, WJ, Behringer, RR, Renfree, MB & Pask, AJ 2011, 'Desert hedgehog is a mammal-specific gene expressed during testicular and ovarian development in a marsupial', *BMC Developmental Biology,* vol. 11, no. 72, pp. 1-12.

O'Leary, NA, Wright, MW, Brister, JR, Ciufo, S, Haddad, D, McVeigh, R, Rajput, B, Robbertse, B, Smith-White, B & Ako-Adjei, D 2015, 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research,* vol. 44, no. D1, pp. D733-D745.

Owen, D & Pemberton, D 2005, *Tasmanian devil: a unique and threatened animal*, Allen & Unwin, Sydney.

Pan, G & Thomson, JA 2007, 'Nanog and transcriptional networks in embryonic stem cell pluripotency', *Cell Research,* vol. 17, no. 1, pp. 42-49.

Perheentupa, J 2006, 'Autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy', *The Journal of Clinical Endocrinology & Metabolism,* vol. 91, no. 8, pp. 2843-2850.

Pierre, A, Pisselet, C, Dupont, J, Bontoux, M & Monget, P 2005, 'Bone morphogenetic protein 5 expression in the rat ovary: biological effects on granulosa cell proliferation and steroidogenesis', *Biology of Reproduction,* vol. 73, no. 6, pp. 1102-1108.

Purcell, S, Neale, B, Todd-Brown, K, Thomas, L, Ferreira, MA, Bender, D, Maller, J, Sklar, P, De Bakker, PI & Daly, MJ 2007, 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *The American Journal of Human Genetics,* vol. 81, no. 3, pp. 559-575.

Quinlan, AR & Hall, IM 2010, 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics,* vol. 26, no. 6, pp. 841-842.

Rossi, P 2013, 'Transcriptional control of KIT gene expression during germ cell development', *International Journal of Developmental Biology,* vol. 57, no. 2-4, pp. 179-184.

Rozas, J, Ferrer-Mata, A, Sánchez-Delbarrio, JC, Guirao-Rico, S, Librado, P, Ramos-Onsins, SE & Sánchez-Gracia, A 2017, 'DnaSP 6: DNA sequence polymorphism analysis of large data sets', *Molecular Biology and Evolution,* vol. 34, no. 12, pp. 3299-3302.

Russell, DL, Brown, HM & Dunning, KR 2015, 'ADAMTS proteases in fertility', *Matrix Biology,* vol. 44-46, pp. 54-63.

Russell, T, Lane, A, Clarke, J, Hogg, C, Morris, K, Keeley, T, Madsen, T & Ujvari, B 2019, 'Multiple paternity and precocial breeding in wild Tasmanian devils, *Sarcophilus harrisii* (Marsupialia: Dasyuridae)', *Biological Journal of the Linnean Society,* vol. 128, no. 1, pp. 201-210.

Sekimizu, K, Nishioka, N, Sasaki, H, Takeda, H, Karlstrom, RO & Kawakami, A 2004, 'The zebrafish iguana locus encodes Dzip1, a novel zinc-finger protein required for proper regulation of Hedgehog signaling', *Development,* vol. 131, no. 11, pp. 2521-2532.

Shastry, BS 2009, 'SNPs: impact on gene function and phenotype', in Komar, A (ed.), *Single Nucleotide Polymorphisms*, Springer, Cleveland, OH.

Simpson, ER, Mahendroo, MS, Means, GD, Kilgore, MW, Hinshelwood, MM, Graham-Lorence, S, Amarneh, B, Ito, Y, Fisher, CR & Michael, MD 1994, 'Aromatase cytochrome P450, the enzyme responsible for estrogen biosynthesis', *Endocrine Reviews,* vol. 15, no. 3, pp. 342-355.

Solovyev, V 2004, 'Statistical approaches in Eukaryotic gene prediction', in Balding, D, Cannings, C & Bishop, M (eds.), *Handbook of Statistical genetics,* 3rd edn, John Wiley & Sons, England.

Sommer, S 2005, 'The importance of immune gene variability (MHC) in evolutionary ecology and conservation', *Frontiers in Zoology,* vol. 2, no. 16, pp. 1-18.

Stelzer, G, Rosen, N, Plaschkes, I, Zimmerman, S, Twik, M, Fishilevich, S, Stein, TI, Nudel, R, Lieder, I & Mazor, Y 2016, 'The GeneCards suite: from gene data mining to disease genome sequence analyses', *Current Protocols in Bioinformatics,* vol. 54, no. 1, pp. 1.30.1 - 1.30.33.

Stephens, M & Donnelly, P 2003, 'A comparison of bayesian methods for haplotype reconstruction from population genotype data', *The American Journal of Human Genetics,* vol. 73, no. 5, pp. 1162-1169.

Stephens, M, Smith, NJ & Donnelly, P 2001, 'A new statistical method for haplotype reconstruction from population data', *The American Journal of Human Genetics,* vol. 68, no. 4, pp. 978-989.

Suh, Y & Vijg, J 2005, 'SNP discovery in associating genetic variation with human disease phenotypes', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* vol. 573, no. 1-2, pp. 41-53.

Tajima, F 1989, 'Statistical method for testing the neutral mutation hypothesis by DNA polymorphism', *Genetics,* vol. 123, no. 3, pp. 585-595.

Tian, X, Pascal, G & Monget, P 2009, 'Evolution and functional divergence of NLRP genes in mammalian reproductive systems', *BMC Evolutionary Biology,* vol. 9, no. 202.

Toder, R, Wakefield, M & Graves, J 2000, 'The minimal mammalian Y chromosome–the marsupial Y as a model system', *Cytogenetic and Genome Research,* vol. 91, no. 1-4, pp. 285-292.

Tomoko, O 1995, 'Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory', *Journal of Molecular Evolution,* vol. 40, no. 1, pp. 56-63.

Ujvari, B & Belov, K 2011, 'Major histocompatibility complex (MHC) markers in conservation biology', *International Journal of Molecular Sciences,* vol. 12, no. 8, pp. 5168-5186.

Wright, B, Morris, K, Grueber, CE, Willet, CE, Gooley, R, Hogg, CJ, O'Meally, D, Hamede, R, Jones, M & Wade, C 2015, 'Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population', *BMC Genomics,* vol. 16, no. 791, pp. 1-11.

Wright, B, Willet, CE, Hamede, R, Jones, M, Belov, K & Wade, CM 2017, 'Variants in the host genome may inhibit tumour growth in devil facial tumours: evidence from genome-wide association', *Scientific Reports,* vol. 7, no. 423, pp. 1-6.

Wright, BR, Farquharson, KA, McLennan, EA, Belov, K, Hogg, CJ & Grueber, CE 2020, 'A demonstration of conservation genomics for threatened species management', *Molecular Ecology Resources,* vol. 00, pp. 1-16.

Yang, H & Wang, K 2015, 'Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR', *Nature Protocols,* vol. 10, no. 10, pp. 1556-1566.

Zalata, A, El-Samanoudy, AZ, Shaalan, D, El-Baiomy, Y, Taymour, M & Mostafa, T 2012, 'Seminal clusterin gene expression associated with seminal variables in fertile and infertile men', *The Journal of Urology,* vol. 188, no. 4, pp. 1260-1264.

3.3 SUPPLEMENTARY

SUPPLEMENTAL TABLES

Supplemental tables are too large for print but are available as a supplementary excel file which can be downloaded using the following link: https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.13295&file=men13295-sup-0002-TableS1-S6.xlsx

Table headers for each supplemental table are provided below.

Table S1 Reproductive genes characterised in the Tasmanian devil with the corresponding annotation information and analysis notes.

Table S2 Reproductive genes unable to be characterised in the Tasmanian devil.

Table S3 The total number of putative SNPs identified in each of the characterised reproductive genes across the 25 high-coverage Tasmanian devil genomes.

Table S4 Nonsynonymous SNPs identified in the 37 resequenced Tasmanian devil genomes.

Table S5 Tests of Hardy-Weinberg equilibrium performed on the target SNPs and related statistics.

Table S6 Diversity statistics and neutrality tests performed on the target reproductive genes with the 9 known offspring excluded.

SUPPLEMENTAL FIGURES

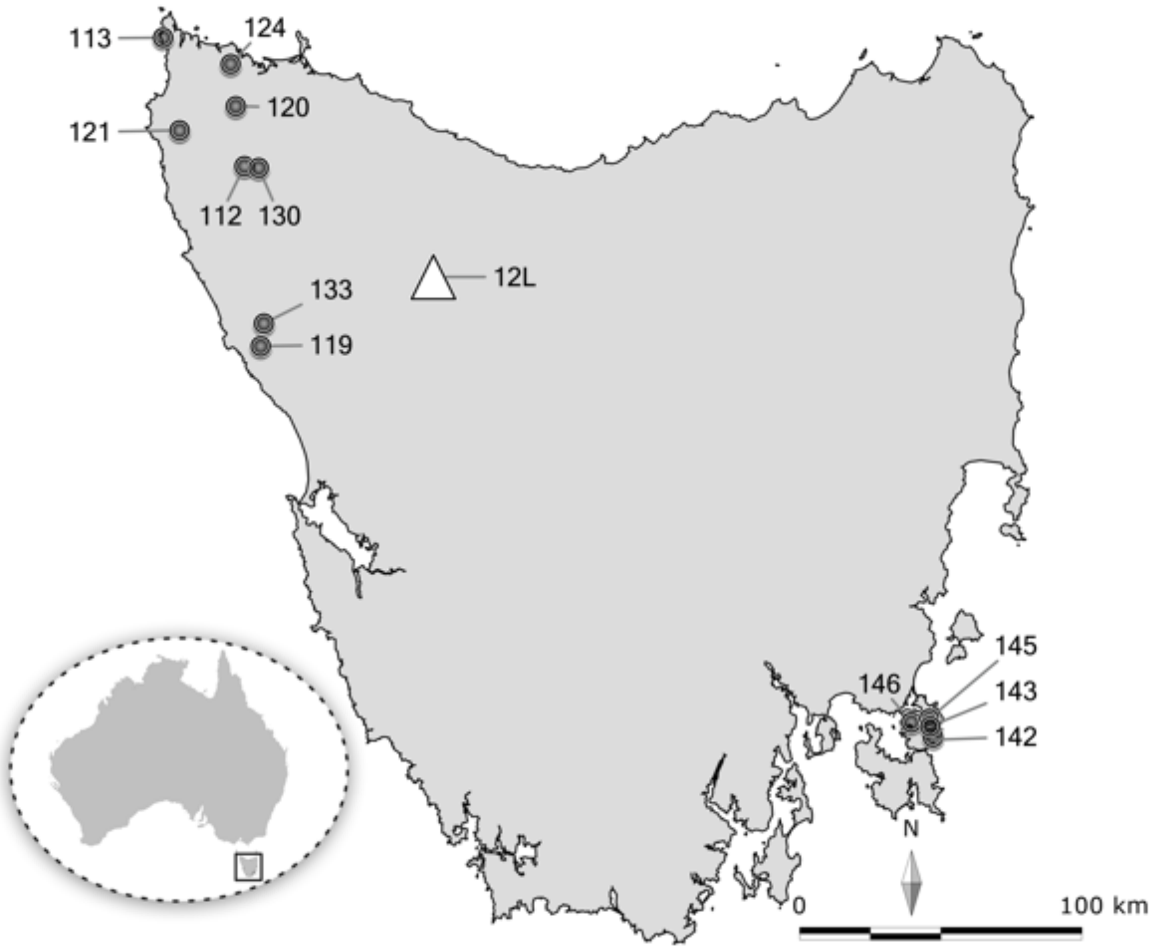


Figure S1 Map showing source locations of the twelve high-coverage wild-born founder individuals (circles) and the population where the 12 low-coverage (12L) resequenced genomes were sampled (triangle). Labels for the wild-born founders correspond to the last three digits of the respective sample name.

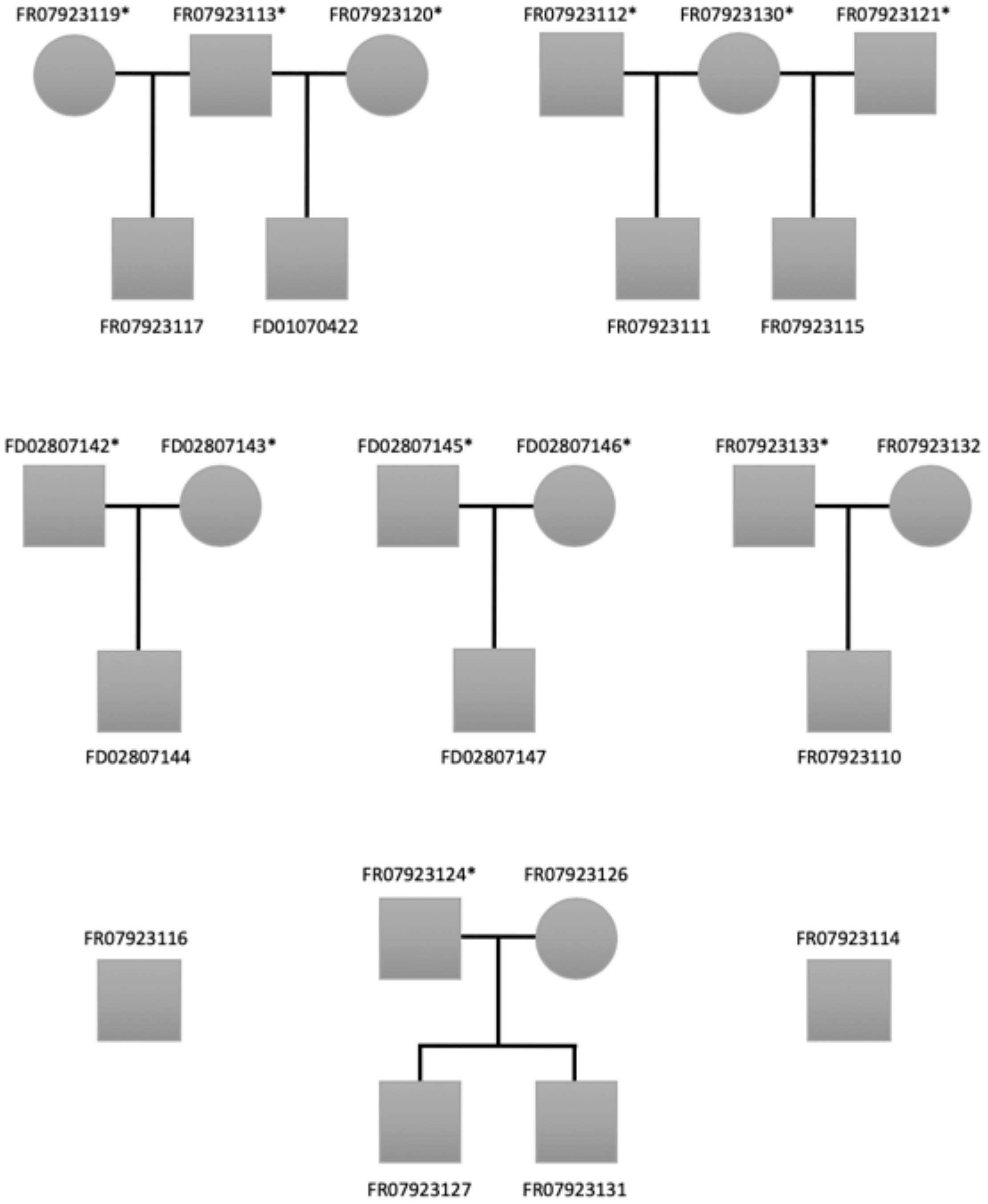


Figure S2 Pedigree showing the relationships between the twenty-five high-coverage resequenced genomes. Asterisks (*) indicate wild founder individuals.

CHAPTER 4

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

THE FIRST *ANTECHINUS* REFERENCE GENOME PROVIDES A RESOURCE FOR INVESTIGATING THE GENETIC BASIS OF SEMELPARITY AND AGE-RELATED NEUROPATHOLOGIES

4.1 BACKGROUND

Chapter 4 comprises the published manuscript:

Brandies, PA, Tang, S, Johnson, RSP, Hogg, CJ & Belov, K 2020, 'The first *Antechinus* reference genome provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies', *Gigabyte,* vol. 1, no. 7, pp. 1-22.

Chapter 1 and Chapter 3 focus on the value of genomic resources available for a well-researched threatened marsupial species. This chapter aims to demonstrate the value of generating reference genomes for non-threatened species as a genomic resource for closely related threatened counterparts and as a model to explore the genetic basis of biological traits with potentially broad implications. I created a reference genome for the brown antechinus, a common native Australian marsupial that exhibits a rare reproductive strategy that makes it an ideal model species for investigating the genetic interplay between stress, reproduction and immunity. The first antechinus reference genome and associated findings provides a key resource for future research to better understand how genetics modulates the relationship between extreme life history trade-offs, which could have crucial implications on threatened marsupial species. Additionally, the antechinus reference genome acts as a valuable tool to assist with population monitoring and conservation of all species in the *Antechinus* genus, particularly those vulnerable to extinction.

Katherine Belov, Carolyn J. Hogg and I conceived and designed the project. Carolyn J. Hogg and I collected the samples with assistance from Robert S.P. Johnson. I prepared the samples, created the reference genome, performed downstream analysis and drafted the manuscript. Simon Tang assisted with downstream analysis.

All authors revised the manuscript. The published PDF version of this manuscript is provided in Appendix 1.

4.2 MAIN ARTICLE

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

* Corresponding Author

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 73 Mb 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.

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 (Braithwaite & Lee, 1979). Although this reproductive strategy is common among bacteria, plant and invertebrate species (Cole, 1954), it is rarely seen in mammalian species and is restricted to didelphid and dasyurid marsupials (Lee & Cockburn, 1985; Naylor, Richardson & McAllan, 2008). 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 (Bradley, McDonald & Lee, 1980; Braithwaite & Lee, 1979; Promislow & Harvey, 1990; Woolley, 1966). Increased levels of plasma corticosteroid assist antechinus males in utilising their energy reserves to maximise reproductive potential during the breeding season (Lee & Cockburn, 1985). However, elevation of these corticosteroids results in total immune system collapse leading to gastrointestinal haemorrhage, parasite/pathogen invasion and death (Bradley, McDonald & Lee, 1980; Lee, Bradley & Braithwaite, 1977). 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) (McAllan, 2006; McAllan, Hobbs & Norris, 2006; Naylor, Richardson & McAllan, 2008). 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 (Ulep, Saraon & McLea, 2018). Traditionally, transgenic mouse models have been utilised to study AD (Elder, Gama Sosa & De Gasperi, 2010; Götz et al., 2004; Schwab, Hosokawa & McGeer, 2004); however, mice do not naturally develop β-amyloid plaques and neurofibrillary tangles (King, 2018; Reardon, 2018). Both of these have been found to develop naturally in mature male and female antechinus, particularly after the breeding season (McAllan, 2006; McAllan, Hobbs & Norris, 2006). 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 (Bradley, McDonald & Lee, 1980; Holleley et al., 2006; Wood, 1970). 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*). We use a bioinformatic approach (Bidon et al., 2015) 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 (Toder, Wakefield & Graves, 2000). 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 brown 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, NSW Scientific License number SL101204) (Tasker & Dickman, 2001), one male and one female adult brown antechinus (*Antechinus stuartii*) were trapped in June 2019 at Lane Cove National Park, NSW. Individuals were euthanased using pentobarbitone (60mg/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 150 bp PE reads. *De novo* genome assembly was performed for both sexes independently with Supernova v2.1.1 (Weisenfeld et al., 2017) using all reads, obtaining approximately 75× raw coverage and 55× effective (deduplicated) coverage. BBTools v38.73 (Bushnell, 2014) was used to generate assembly statistics and BUSCO (Simão et al., 2015) 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) (O'Leary et al., 2015) using nucmer v4.0.0beta2 (Kurtz et al., 2004) 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 (Bidon et al., 2015) and confirmed through BLAST searches of known marsupial Y genes.

Firstly, both the male and female 10x reads were trimmed to remove the 10x Chromium barcode and low-quality sequence using FastQC v0.11.5 (Andrews, 2010) and BBTools. Male and female trimmed reads were aligned to the male genome assembly separately using BWA (Burrows-Wheeler Aligner) v0.7.17-r1188 (Li & Durbin, 2009) with shorter split hits marked as secondary using the -*M* flag, duplicates were removed using samblaster v0.1.24 (Faust & Hall, 2014) with duplicates excluded using the -*e* flag, and alignments with quality scores <20 were removed with samtools v1.10 (Li et al., 2009) 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 (Pedersen & Quinlan, 2017) in fast mode. Following a previous study (Bidon et al., 2015), the AD-ratio of each scaffold was calculated for each scaffold whereby a normalised ratio of female reads to male reads should result in a value of ~1 (0.7 < 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 < 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 (AD-ratio ≤ 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 (Altschul et al., 1990; Camacho et al., 2009) to map 20 known marsupial Y genes and their autosomal or X homologs (if available) from a previous study (Cortez et al., 2014) against the male brown antechinus assembly. Scaffolds with an AD-ratio <0.3 and strong BLAST matches (1e^{-}

^{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 150 bp 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 (Grabherr et al., 2011; Haas et al., 2013) with default Trimmomatic (Bolger, Lohse & Usadel, 2014) 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 (Consortium, 2018) BLAST searches, and completeness was assessed using BUSCO v3 and v4. Trimmed reads were mapped back to the assembly using bowtie2 v2.3.5.1 (Langmead & Salzberg, 2012) 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 and Salmon v1.0.0 (Patro et al., 2017) with default parameters to create a cross-sample TMM normalised matrix of expression values (Dillies et al., 2013; Robinson & Oshlack, 2010). 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 (Haas et al., 2013).

Functional annotation of the global transcriptome was performed using Trinotate v3.2.0 (Bryant et al., 2017). Briefly, TransDECODER v5.5.0 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 database and the Tasmanian devil reference genome annotations from NCBI (RefSeq assembly mSarHar1.11) (O'Leary et al.,

2015) with an e-value cut-off of 1e^{-5}. HMMER v3.2.0 (Eddy, 2018) was used to identify conserved protein domains with the Pfam (El-Gebali et al., 2019) database, SignalP v4.1 (Nielsen, 2017) was used to predict signal peptides and RNAmmer v1.2 (Lagesen et al., 2007) 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 database.

Repeat Identification and Genome Annotation

A custom repeat database was generated with RepeatModeler v2.0.1 (Smit, Hubley & Green, 2008-2015) and repeats (excluding low complexity regions and simple repeats with the *-nolow* flag) were masked with RepeatMasker v4.0.6 (Smit, Hubley & Green, 2013-2015). Genome annotation was performed using Fgenesh++ v7.2.2 (Salamov & Solovyev, 2000; Solovyev et al., 2006; Solovyev, 2002) 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 and seqkit v0.10.1 (Shen et al., 2016). 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 searches against the Swiss-Prot database with an e-value cut-off of 1e^{-5} to identify genes with matches to known high quality proteins from other species.

Variant Annotation

The male reference genome was altered following the 10x Genomics Long Ranger (Zheng et al., 2016) 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 (Kent et al., 2002) 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 (Zheng et al., 2016) with the FreeBayes option. Male and female VCF files were merged with bcftools v1.10.1 (Li et al., 2009) and variants were annotated using ANNOVAR v20180416 (Wang, Li & Hakonarson, 2010; Yang & Wang, 2015) gene-based annotation.

Gene Family Analysis

Gene ontology (GO) annotation (using the generic GO slim subset) was performed on brown antechinus proteins based on Swiss-Prot matches using GOnet (Pomaznoy, Ha & Peters, 2018) to identify genes associated with key biological functions.

To identify any rapidly evolving gene families in the brown antechinus, proteomes from six other target species (Tasmanian devil, koala, opossum, human, mouse and platypus) were downloaded from NCBI (O'Leary et al., 2015) and the longest isoform for each gene was extracted using custom bash scripts. Protein sequences from the brown antechinus Fgenesh++ annotation were also extracted and OrthoFinder v2.4.0 (Emms & Kelly, 2015; Emms & Kelly, 2019) was run with default parameters to identify orthogroups between the 7 target species. CAFE v5 (De Bie et al., 2006; Hahn et al., 2005) was run on the output data from OrthoFinder 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 brown 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 (Stelzer et al., 2016) using the keyword “Alzheimer's” were used to identify forty common genes that have previously been associated with Alzheimer's disease in humans or mice disease models. Human coding sequences (CDS) for the genes of interest were downloaded from Swiss-Prot and were used in BLAST searches against the Fgenesh++ genome annotations to identify the predicted gene sequences within the male brown antechinus reference genome. The predicted protein sequences were matched against the predicted coding sequences of the global transcriptome using BLAST 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 searches back to the Human Swiss-Prot proteome to confirm orthology through reciprocal best hits (RBH) and were aligned to human protein sequences with MUSCLE v3.8.425 (Edgar, 2004) in order to determine sequence similarity and identity. SNVs associated with the target genes were explored using the ANNOVAR output.

Findings

Genome Assembly

The male and female brown 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 4.1). 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 4.1). 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 [Weisenfeld et al., 2017]) respectively. The male assembly was chosen to be the reference genome as it showed the highest contiguity and 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 2n = 14 karyotypes (Deakin, 2018). Antechinus chromosomes were therefore bioinformatically assigned by alignment of the male brown 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 4.1a). The length of assigned antechinus chromosomes was similar to that of the Tasmanian devil as expected (Figure 4.1b).

Table 4.1 Comparison of brown 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 BUSCOs v3 (%) ↑**	**Complete Mammalian BUSCOs 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

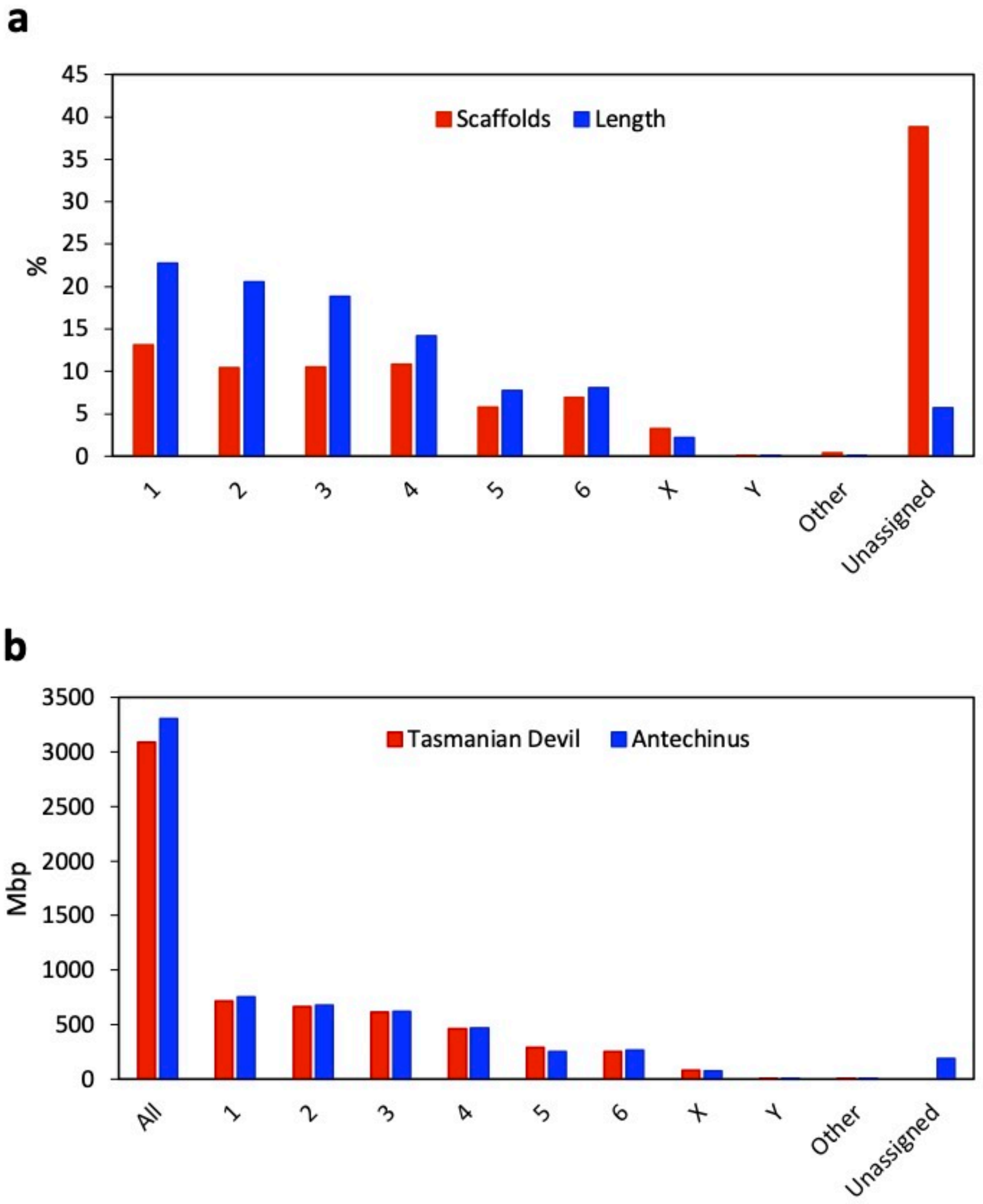


Figure 4.1 Assignment of antechinus scaffolds to chromosomes by alignment to the Tasmanian devil reference genome. **a)** Proportion (%) of scaffolds (blue) and genome length (red) assigned to chromosomes. **b)** Comparison of length of sequence assigned to each chromosome from the Tasmanian devil reference 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.

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 Bidon et al., 2015). 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 4.2). The results from this approach showed that ~92% of the genome was in agreeance with the chromosome assignment results from mapping the brown 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 containing known Y genes and Y-specific transcripts. Out of 20 known marsupial Y chromosome genes from a previous study (Cortez et al., 2014), 13 showed hits to scaffolds with AD-ratios ≤0.01 indicating a high probability 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 4.3). These scaffolds were deemed Y-chromosome scaffolds and comprise 0.78 Mb of the genome. They represent the largest amount of Y-chromosome sequence characterised 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 brown antechinus will assist with future research wanting to explore male semelparity and key male-specific reproductive genes.

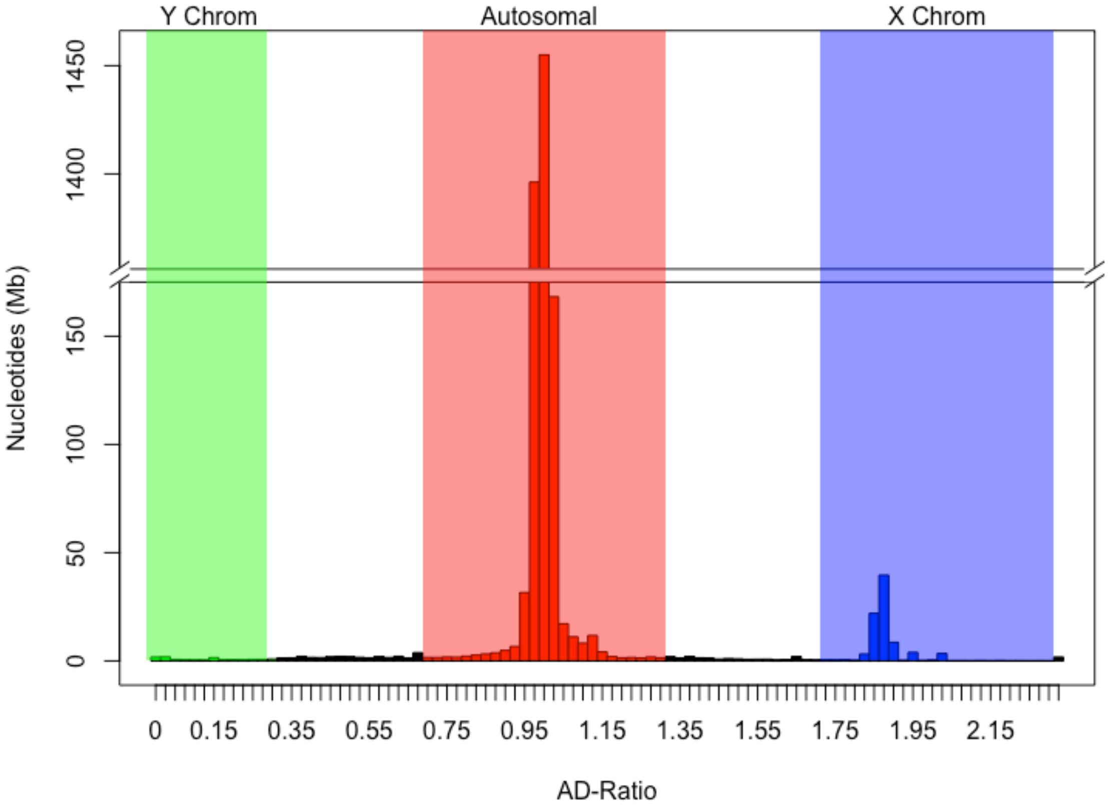


Figure 4.2 AD-Ratio histogram of antechinus scaffolds. Figure shows the total length of sequence within each 0.025 AD-ratio bin. Scaffolds clustering around an AD-ratio of 0 represent Y-linked sequence (Green), scaffolds clustering around an AD-ratio of 1 represent Autosomal sequence (Red), scaffolds clustering around an AD-ratio of 2 represent X-linked sequence (Blue) and scaffolds between these regions represent unassigned sequence (Black).

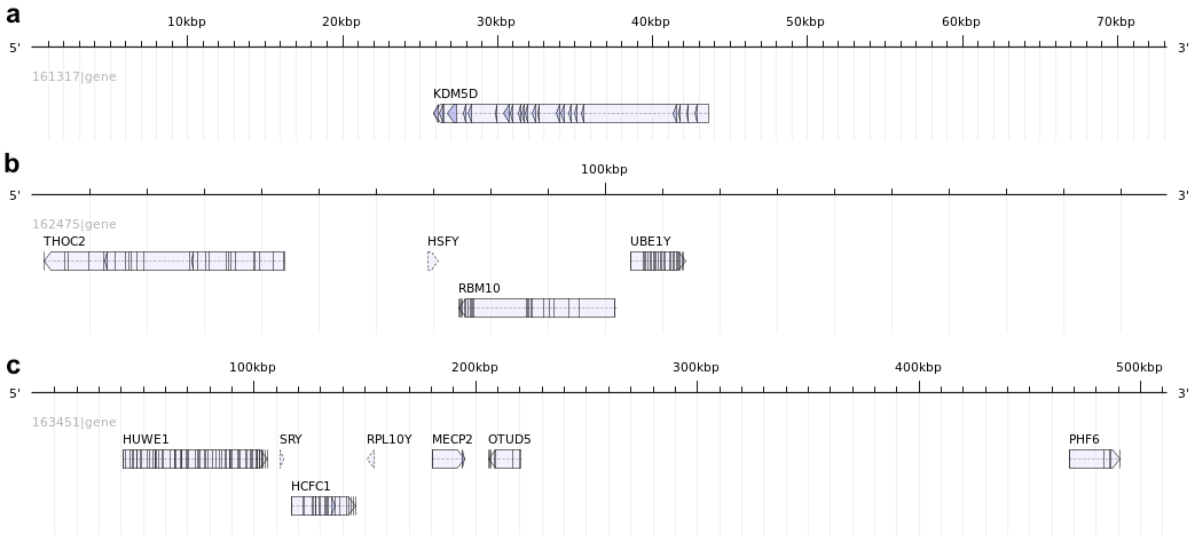


Figure 4.3 Mapping of known marsupial Y gene homologs on antechinus Y chromosome scaffolds. **a)** Scaffold 161317, **b)** Scaffold 162475, **c)** Scaffold 163451. Figure 4.3 was created using the AnnotationSketch module from GenomeTools (Gremme, Steinbiss & Kurtz, 2013).

Transcriptome Assembly and Annotation

The global brown 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 773 bp and the contig N50 was 1,367 bp. Considering only the top 95% most highly expressed transcripts gave an ExN50 (a more useful indicator of transcriptome quality) of 3,020 bp which is similar to the average mRNA length in humans (3,392 bp) (Piovesan et al., 2016). 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 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 4.2), with 44.82% of the genome being masked as repetitive; a similar repeat content to that of other marsupial and mammalian genomes (Margulies et al., 2005). 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.

Table 4.2 Summary of repeat classes identified and masked in the brown 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%

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 (Crowther & Braithwaite, 2008; Van Dyck & Crowther, 2000). The large population size and range of *A. stuartii* implies that this species would likely exhibit healthy levels of genomic diversity, though 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 1 kb upstream or downstream of a gene) of which 91% were intronic and 2% were exonic (Figure 4.4a). Within the exonic variants, 58% were nonsynonymous (result in alteration of the protein sequence) and 39% were synonymous (Figure 4.4b). 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 (Wright et al., 2020). 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 (Gray, Baker & Firn, 2017). Most antechinus species diverged in the Pilocene (~5mya) with the brown antechinus and its close relatives separating more recently in the Pleistocene (~2.5mya) (Mutton et al., 2019). Humans and chimpanzees are predicted to have diverged 7-8mya (Langergraber et al., 2012) but still share 99% of their DNA (Mikkelsen et al., 2005). The genetic similarity of human 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.

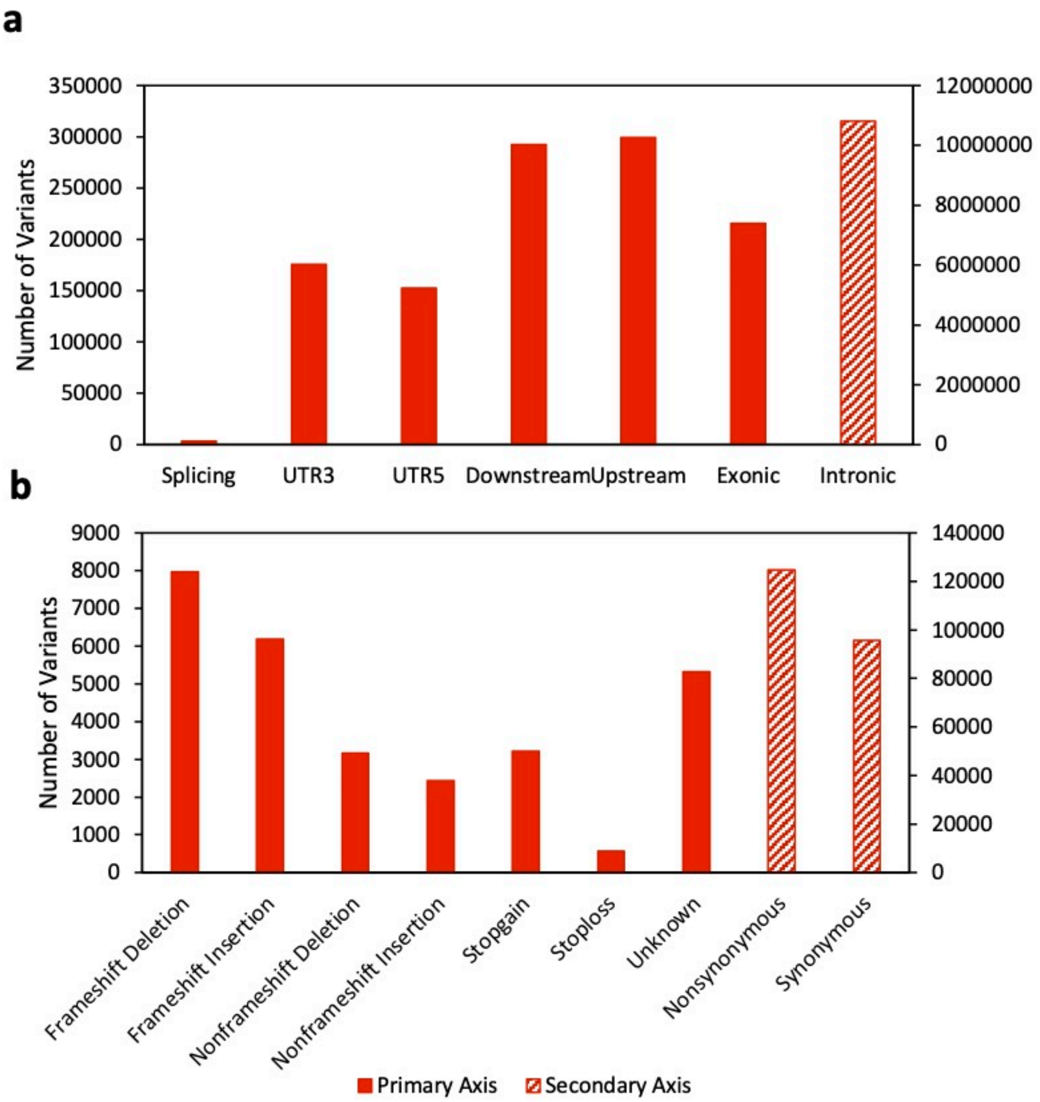


Figure 4.4 Functional annotation of brown antechinus variants. **a)** Total number of variants annotated to various gene regions including: Splicing (within a splice site of a gene), UTR3 (3' untranslated region), UTR5 (5' untranslated region), Downstream (within 1 kb downstream of a gene), Upstream (within 1 kb 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.

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 (Feuk, Carson & Scherer, 2006; Mahmoud et al., 2019). A few interchromosomal and intrachromosomal rearrangements have been identified in the Dasyuridae family using previous G-banding techniques (Deakin & Kruger-Andrzejewska, 2016); however, advancements in sequencing technologies, such as the linked-read approach utilised in the current study, allow for more fine-scale characterisation of structural variants in a cost-effective and reliable manner. (Balachandran & Beck, 2020). Using the linked-read datasets, 700 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 4.5). 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.

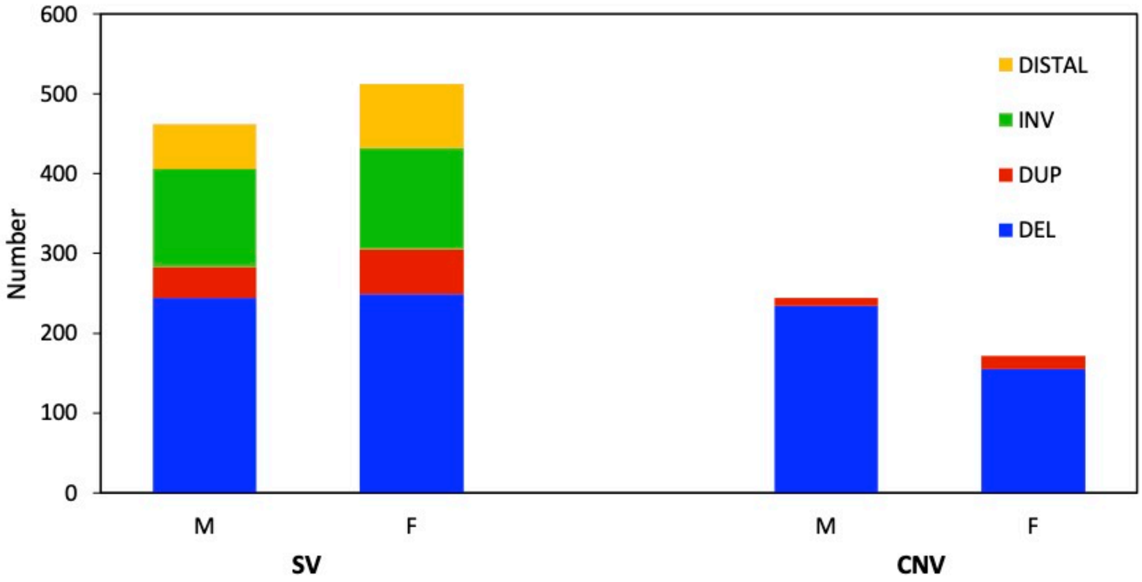


Figure 4.5 Breakdown of high-quality large structural variants (SVs) and copy number variants (CNVs) in the brown antechinus. Figure shows both male (M) and female (F) deletions (blue), tandem duplications (red), inversions (green) and distal structural variants (i.e., across two scaffolds, yellow).

Gene Family Analysis

GO analysis of the brown 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 brown 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 ($<1e^{-15}$) 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 brown antechinus (Orthogroup OG0000350).

Protocadherins (Pcdhs) belong to the cadherin superfamily and are organised into 3 main gene clusters: α, β and γ (Hayashi & Takeichi, 2015). Pcdhs, like all cadherins, are primarily responsible for mediating cell-cell adhesion (Chen & Maniatis, 2013). The brown antechinus displayed similar numbers of putative *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 4.6). *Pcdh*-γ genes specifically have been implicated in neuronal processes (Hayashi & Takeichi, 2015) and have previously been associated with Alzheimer's disease (Li et al., 2017). 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 brown antechinus. It is possible that the expansion of *Pcdh*-γ genes in the brown antechinus is 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.

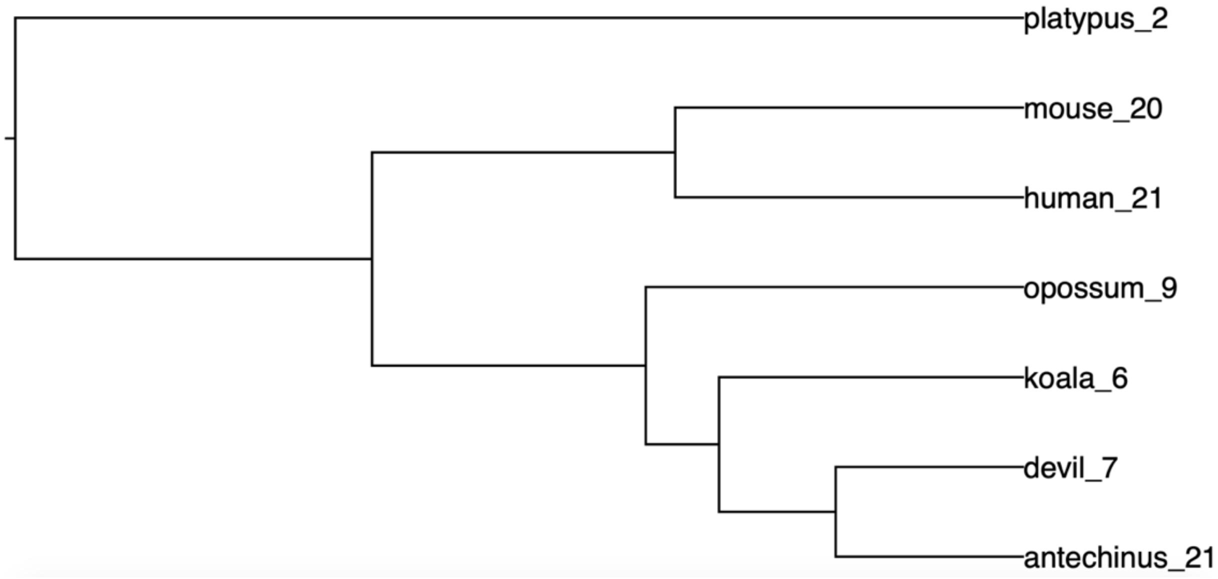
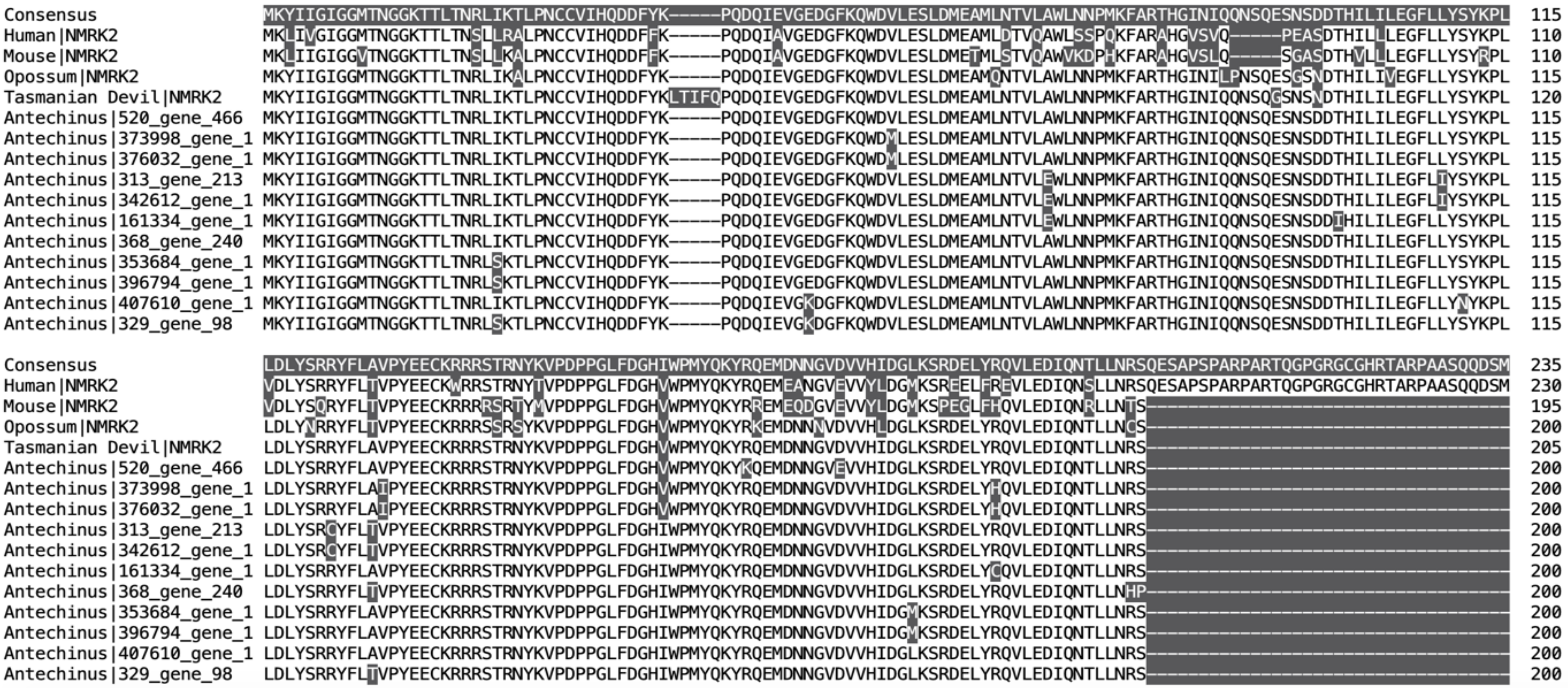


Figure 4.6 Gene tree showing numbers of *Pcdh*-γ genes across 7 species.

The brown 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 (Figure 4.7). 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 (Johnson & Imai, 2018; Yang & Sauve, 2016). The brown antechinus contains 11 full-length copies of this gene in its genome (Figure 4.7). 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 brown 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 (Johnson & Imai, 2018). NAD+ has also been associated with Alzheimer's disease whereby increased levels of the molecule may be a protective factor of the disease (Hou et al., 2018). The antechinus used 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 brown antechinus may provide crucial insights into aging and age-related dementias in humans.

Alzheimer's Genes Analysis

To investigate further the potential of antechinus as a disease model for AD (McAllan, 2006; Naylor, Richardson & McAllan, 2008), 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 brown antechinus reference genome and all 40 were expressed in the global transcriptome (Table 4.3). The *CD2AP* gene was not annotated by Fgenesh++ so was not included in downstream analysis. All of the annotated brown antechinus proteins except *PLD3* were found to be orthologous to the human proteins using a RBH strategy (Table 4.3). 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 (Kuzniar et al., 2008) (Table 4.3). Of the seven antechinus gene annotations with 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 4.3). The remaining four genes (*CD33*, *ZCWPW1*, *ABCA7* and *CR1*) did not have homologous genome annotations nor transcripts in the antechinus (large gaps were displayed in all sequences compared to the human genes) and were therefore excluded from downstream analysis.



```
Consensus                  MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Human|NMRK2                MKLIVGIGGMTNGGKTTLTNSLLRALPNCCVIHQDDFFK-----PQDQIAVGEDGFKQWDVLESLDMEAMLDTVQAWLSSPQKFARAHGVSVQ-----PEASDTHILLLEGFLLYSYKPL  110
Mouse|NMRK2                MKLIIGIGGVTNGGKTTLTNSLLKALPNCCVIHQDDFFK-----PQDQIAVGEDGFKQWDVLESLDMETMLSTVQAWVKDPHKFARAHGVSLQ-----SGASDTHVLLLEGFLLYSYRPL  110
Opossum|NMRK2              MKYIIGIGGMTNGGKTTLTNRLIKALPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAQNTVLAWLNNPMKFARTHGINILPNSQESGSNDTHILIVEGFLLYSYKPL  115
Tasmanian Devil|NMRK2      MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYKLTIFQPQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQGSNSNDTHILILEGFLLYSYKPL  120
Antechinus|520_gene_466    MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|373998_gene_1   MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDMLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|376032_gene_1   MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDMLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|313_gene_213    MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLEWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLIYSYKPL  115
Antechinus|342612_gene_1   MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLEWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLIYSYKPL  115
Antechinus|161334_gene_1   MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLEWLNNPMKFARTHGINIQQNSQESNSDIHILILEGFLLYSYKPL  115
Antechinus|368_gene_240    MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|353684_gene_1   MKYIIGIGGMTNGGKTTLTNRLSKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|396794_gene_1   MKYIIGIGGMTNGGKTTLTNRLSKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|407610_gene_1   MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGKDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYNYKPL  115
Antechinus|329_gene_98     MKYIIGIGGMTNGGKTTLTNRLSKTLPNCCVIHQDDFYK-----PQDQIEVGKDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115

Consensus                  LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRSQESAPSPARPARTQGPGRGCGHRTARPAASQQDSM  235
Human|NMRK2                VDLYSRRYFLTVPYEECKWRRSTRNYTVPDPPGLFDGHVWPMYQKYRQEMEANGVEVVYLDGMKSREELFREVLEDIQNSLLNRSQESAPSPARPARTQGPGRGCGHRTARPAASQQDSM  230
Mouse|NMRK2                VDLYSQRYFLTVPYEECKRRRRSRTYMVPDPPGLFDGHVWPMYQKYRREMEQDGVEVVYLDGMKSPEGLFHQVLEDIQNRLLNTS-----------------------------------  195
Opossum|NMRK2              LDLYNRRYFLTVPYEECKRRRSSRSYKVPDPPGLFDGHVWPMYQKYRKEMDNNNVDVVHLDGLKSRDELYRQVLEDIQNTLLNCS-----------------------------------  200
Tasmanian Devil|NMRK2      LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHVWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  205
Antechinus|520_gene_466    LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHVWPMYQKYKQEMDNNGVEVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|373998_gene_1   LDLYSRRYFLAIPYEECKRRRSTRNYKVPDPPGLFDGHVWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYHQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|376032_gene_1   LDLYSRRYFLAIPYEECKRRRSTRNYKVPDPPGLFDGHVWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYHQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|313_gene_213    LDLYSRCYFLTVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|342612_gene_1   LDLYSRCYFLTVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|161334_gene_1   LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYCQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|368_gene_240    LDLYSRRYFLTVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNHP-----------------------------------  200
Antechinus|353684_gene_1   LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGMKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|396794_gene_1   LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGMKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|407610_gene_1   LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|329_gene_98     LDLYSRRYFLTVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
```

Figure 4.7 Protein sequence alignment showing expansion of *NMRK2* genes in the brown antechinus. Single copy genes in the human, mouse, gray short-tailed opossum and Tasmanian devil are shown for comparison.

Table 4.3 Summary of Alzheimer's related genes explored in the brown 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

Gene	Gene ID*	Evidence^	Trans ID†	Protein Length (Tran) (BP)	Human Protein Length (bP)	RBH§	% Ident (Tran)	% Sim (Tran)
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)

Gene	Gene ID*	Evidence^	Trans ID†	Protein Length (Tran) (BP)	Human Protein Length (bP)	RBH§	% Ident (Tran)	% Sim (Tran)
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)

Gene	Gene ID*	Evidence^	Trans ID†	Protein Length (Tran) (BP)	Human Protein Length (bP)	RBH§	% Ident (Tran)	% Sim (Tran)
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

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 brown antechinus brain tissue. *APP* is the precursor for the amyloid beta (Aβ) proteins that form amyloid plaques in the brain and is predicted to contribute to early-onset AD in humans (O'Brien & Wong, 2011). The *MAPT* gene was also most highly expressed in brown antechinus brain tissue and is responsible for the creation of tau proteins which form the neurofibrillary tangles associated with AD (Iqbal et al., 2010). *APOE* (apolipoprotein E) is the most common risk-factor gene associated with late-onset AD (Liu et al., 2013) and was highly expressed across a range of brown antechinus tissues including the brain. *PICALM* is another common gene which has been associated with an increased risk of developing late-onset AD (Xu, Tan & Yu, 2015). *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 (Zhao et al., 2015). This gene was found to be quite lowly expressed in brown antechinus brain tissue when compared to 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 (Lambert et al., 2013; Tábuas-Pereira et al., 2020) 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 utilised 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 (Cuyvers & Sleegers, 2016; Mendez, 2019; Rosenthal & Kamboh, 2014; Sims, Hill & Williams, 2020; Sun et al., 2017; Tábuas-Pereira et al., 2020). 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 10x Genomics filters) were associated with the 40 target genes with majority of these being intronic (Figure 4.8). 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 4.8c). While the phenotypic effects of these putatively functional variants are currently unknown, mutations in these genes are commonly associated with AD neuropathologies in humans (Cuyvers & Sleegers, 2016; Mendez, 2019; Rosenthal & Kamboh, 2014; Sims, Hill & Williams, 2020; Sun et al., 2017; Tábuas-Pereira et al., 2020) and may also be associated with the age-related development of neuropathologies observed in mature antechinus brains (Naylor, Richardson & McAllan, 2008).

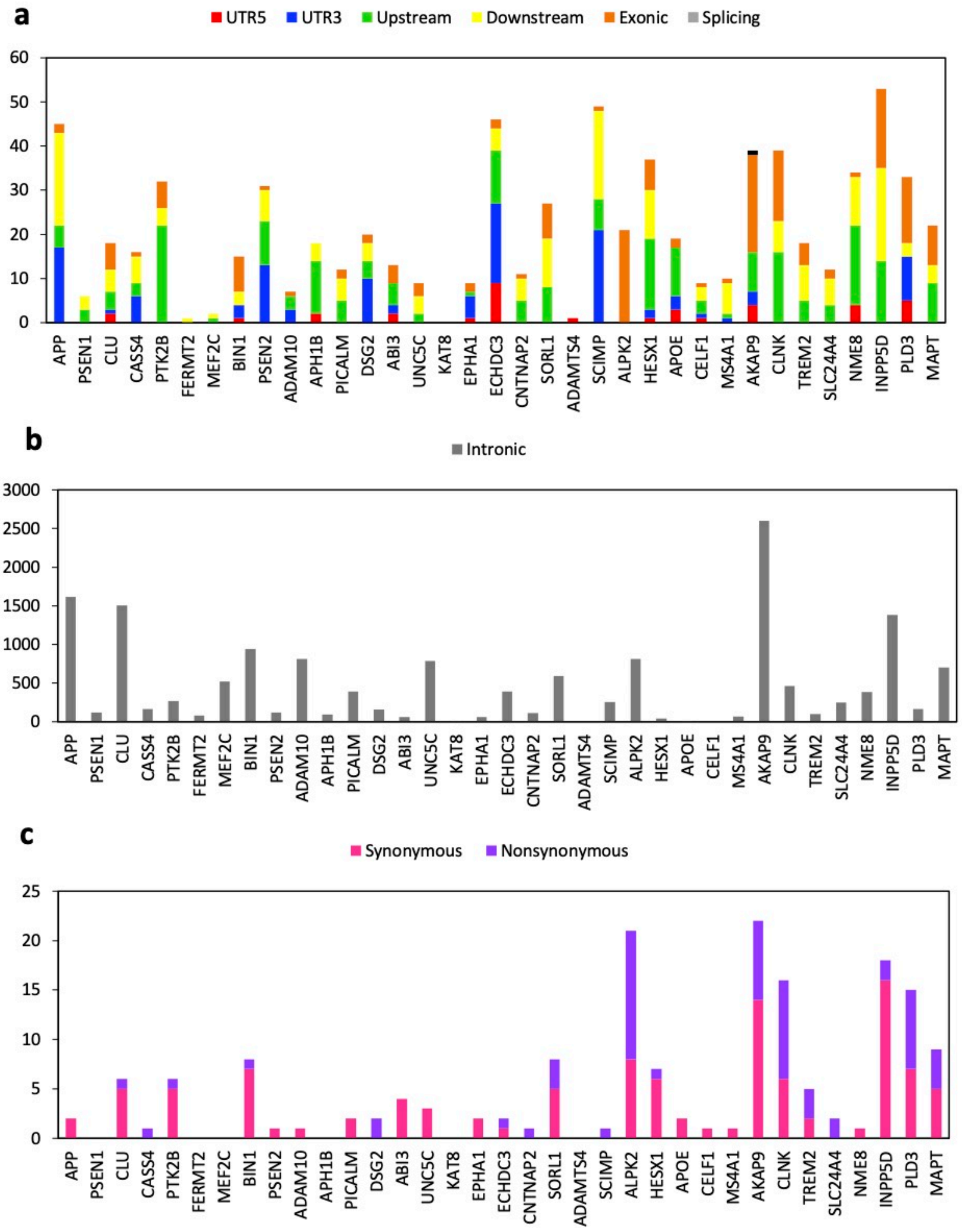


Figure 4.8 Number of each type of SNV associated with the target Alzheimer's-related genes in the antechinus. **a)** Numbers of SNVs present in the 5' UTR, 3' UTR, 1 kb upstream region, 1 kb downstream region, exons, and splice sites of each gene. **b)** Numbers of intronic SNVs present in each gene. **c)** Number of synonymous and nonsynonymous SNVs present in each gene.

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 *A. stuartii* and closely related species. Gene ontology analysis of the annotated brown 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 brown 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 brown antechinus genome as a springboard to study age-related neurodegeneration, as well as a model for extreme life history trade-offs like semelparity.

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

Altschul, SF, Gish, W, Miller, W, Myers, EW & Lipman, DJ 1990, 'Basic local alignment search tool', *Journal of Molecular Biology,* vol. 215, no. 3, pp. 403-410.

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

Balachandran, P & Beck, CR 2020, 'Structural variant identification and characterization', *Chromosome Research,* vol. 28, pp. 31-47.

Bidon, T, Schreck, N, Hailer, F, Nilsson, MA & Janke, A 2015, 'Genome-wide search identifies 1.9 Mb from the polar bear Y chromosome for evolutionary analyses', *Genome Biology and Evolution,* vol. 7, no. 7, pp. 2010-2022.

Bolger, AM, Lohse, M & Usadel, B 2014, 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics,* vol. 30, no. 15, pp. 2114-2120.

Bradley, A, McDonald, I & Lee, A 1980, 'Stress and mortality in a small marsupial (*Antechinus stuartii*, Macleay)', *General and Comparative Endocrinology,* vol. 40, no. 2, pp. 188-200.

Braithwaite, RW & Lee, AK 1979, 'A mammalian example of semelparity', *The American Naturalist,* vol. 113, no. 1, pp. 151-155.

Bryant, DM, Johnson, K, Ditommaso, T, Tickle, T, Couger, MB, Payzin-Dogru, D, Lee, TJ, Leigh, ND, Kuo, T-H & Davis, FG 2017, 'A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors', *Cell Reports,* vol. 18, no. 3, pp. 762-776.

Bushnell, B 2014, *BBTools*, viewed 23 August 2020, https://sourceforge.net/projects/bbmap/

Camacho, C, Coulouris, G, Avagyan, V, Ma, N, Papadopoulos, J, Bealer, K & Madden, TL 2009, 'BLAST+: architecture and applications', *BMC Bioinformatics,* vol. 10, no. 421, pp. 1-9.

Chen, WV & Maniatis, T 2013, 'Clustered protocadherins', *Development,* vol. 140, no. 16, pp. 3297-3302.

Cole, LC 1954, 'The population consequences of life history phenomena', *The Quarterly Review of Biology,* vol. 29, no. 2, pp. 103-137.

Cortez, D, Marin, R, Toledo-Flores, D, Froidevaux, L, Liechti, A, Waters, PD, Grützner, F & Kaessmann, H 2014, 'Origins and functional evolution of Y chromosomes across mammals', *Nature,* vol. 508, no. 7497, pp. 488-493.

Crowther, M & Braithwaite, RW 2008, 'Brown antechinus, *Antechinus stuartii*', in Van Dyck, S & Strahan, R (eds.), *The Mammals of Australia,* Third edn, Reed New Holland, Sydney, Australia.

Cuyvers, E & Sleegers, K 2016, 'Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond', *The Lancet Neurology,* vol. 15, no. 8, pp. 857-868.

De Bie, T, Cristianini, N, Demuth, JP & Hahn, MW 2006, 'CAFE: a computational tool for the study of gene family evolution', *Bioinformatics,* vol. 22, no. 10, pp. 1269-1271.

Deakin, JE 2018, 'Chromosome evolution in marsupials', *Genes,* vol. 9, no. 2, pp. 72.

Deakin, JE & Kruger-Andrzejewska, M 2016, 'Marsupials as models for understanding the role of chromosome rearrangements in evolution and disease', *Chromosoma,* vol. 125, no. 4, pp. 633-644.

Dillies, M-A, Rau, A, Aubert, J, Hennequet-Antier, C, Jeanmougin, M, Servant, N, Keime, C, Marot, G, Castel, D & Estelle, J 2013, 'A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis', *Briefings in Bioinformatics,* vol. 14, no. 6, pp. 671-683.

Eddy, SR 2018, *HMMER*, viewed 11 May 2020, http://hmmer.org

Edgar, RC 2004, 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research,* vol. 32, no. 5, pp. 1792-1797.

El-Gebali, S, Mistry, J, Bateman, A, Eddy, SR, Luciani, A, Potter, SC, Qureshi, M, Richardson, LJ, Salazar, GA & Smart, A 2019, 'The Pfam protein families database in 2019', *Nucleic Acids Research,* vol. 47, no. D1, pp. D427-D432.

Elder, GA, Gama Sosa, MA & De Gasperi, R 2010, 'Transgenic mouse models of Alzheimer's disease', *Mount Sinai Journal of Medicine,* vol. 77, no. 1, pp. 69-81.

Emms, DM & Kelly, S 2015, 'OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy', *Genome Biology,* vol. 16, no. 157, pp. 1-14.

Emms, DM & Kelly, S 2019, 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome Biology,* vol. 20, no. 238, pp. 1-14.

Faust, GG & Hall, IM 2014, 'SAMBLASTER: fast duplicate marking and structural variant read extraction', *Bioinformatics,* vol. 30, no. 17, pp. 2503-2505.

Feuk, L, Carson, AR & Scherer, SW 2006, 'Structural variation in the human genome', *Nature Reviews Genetics,* vol. 7, no. 2, pp. 85-97.

Götz, J, Streffer, J, David, D, Schild, A, Hoerndli, F, Pennanen, L, Kurosinski, P & Chen, F 2004, 'Transgenic animal models of Alzheimer's disease and related disorders: histopathology, behavior and therapy', *Molecular Psychiatry,* vol. 9, no. 7, pp. 664-683.

Grabherr, MG, Haas, BJ, Yassour, M, Levin, JZ, Thompson, DA, Amit, I, Adiconis, X, Fan, L, Raychowdhury, R & Zeng, Q 2011, 'Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data', *Nature Biotechnology,* vol. 29, no. 7, pp. 644.

Gray, EL, Baker, AM & Firn, J 2017, '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 Research,* vol. 62, no. 1, pp. 47-63.

Gremme, G, Steinbiss, S & Kurtz, S 2013, 'GenomeTools: a comprehensive software library for efficient processing of structured genome annotations', *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 10, no. 3, pp. 645-656.

Haas, BJ, Papanicolaou, A, Yassour, M, Grabherr, M, Blood, PD, Bowden, J, Couger, MB, Eccles, D, Li, B & Lieber, M 2013, 'De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nature Protocols,* vol. 8, no. 8, pp. 1494.

Hahn, MW, De Bie, T, Stajich, JE, Nguyen, C & Cristianini, N 2005, 'Estimating the tempo and mode of gene family evolution from comparative genomic data', *Genome Research,* vol. 15, no. 8, pp. 1153-1160.

Hayashi, S & Takeichi, M 2015, 'Emerging roles of protocadherins: from self-avoidance to enhancement of motility', *Journal of Cell Science,* vol. 128, no. 8, pp. 1455-1464.

Holleley, CE, Dickman, CR, Crowther, MS & Oldroyd, BP 2006, 'Size breeds success: multiple paternity, multivariate selection and male semelparity in a small marsupial, *Antechinus stuartii*', *Molecular Ecology,* vol. 15, no. 11, pp. 3439-3448.

Hou, Y, Lautrup, S, Cordonnier, S, Wang, Y, Croteau, DL, Zavala, E, Zhang, Y, Moritoh, K, O’Connell, JF & Baptiste, BA 2018, 'NAD+ supplementation normalizes key Alzheimer’s features and DNA damage responses in a new AD mouse model with introduced DNA repair deficiency', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 115, no. 8, pp. E1876-E1885.

Iqbal, K, Liu, F, Gong, C-X & Grundke-Iqbal, I 2010, 'Tau in Alzheimer disease and related tauopathies', *Current Alzheimer Research,* vol. 7, no. 8, pp. 656-664.

Johnson, S & Imai, SI 2018, 'NAD+ biosynthesis, aging, and disease', *F1000Research,* vol. 7, no. 132, pp. 1-10.

Kent, WJ, Sugnet, CW, Furey, TS, Roskin, KM, Pringle, TH, Zahler, AM & Haussler, D 2002, 'The human genome browser at UCSC', *Genome Research,* vol. 12, no. 6, pp. 996-1006.

King, A 2018, 'The search for better animal models of Alzheimer's disease', *Nature,* vol. 559, no. 7715, pp. S13.

Kurtz, S, Phillippy, A, Delcher, AL, Smoot, M, Shumway, M, Antonescu, C & Salzberg, SL 2004, 'Versatile and open software for comparing large genomes', *Genome Biology,* vol. 5, no. 2, pp. R12.

Kuzniar, A, Van Ham, RC, Pongor, S & Leunissen, JA 2008, 'The quest for orthologs: finding the corresponding gene across genomes', *Trends in Genetics,* vol. 24, no. 11, pp. 539-551.

Lagesen, K, Hallin, P, Rødland, EA, Stærfeldt, H-H, Rognes, T & Ussery, DW 2007, 'RNAmmer: consistent and rapid annotation of ribosomal RNA genes', *Nucleic Acids Research,* vol. 35, no. 9, pp. 3100-3108.

Lambert, J-C, Ibrahim-Verbaas, CA, Harold, D, Naj, AC, Sims, R, Bellenguez, C, Jun, G, Destefano, AL, Bis, JC & Beecham, GW 2013, 'Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease', *Nature Genetics,* vol. 45, no. 12, pp. 1452-1458.

Langergraber, KE, Prüfer, K, Rowney, C, Boesch, C, Crockford, C, Fawcett, K, Inoue, E, Inoue-Muruyama, M, Mitani, JC & Muller, MN 2012, 'Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 109, no. 39, pp. 15716-15721.

Langmead, B & Salzberg, SL 2012, 'Fast gapped-read alignment with Bowtie 2', *Nature Methods,* vol. 9, no. 4, pp. 357-359.

Lee, AK, Bradley, AJ & Braithwaite, RW 1977, 'Corticosteroid levels and male mortality in *Antechinus stuartii*', in Stonehouse, B & Gilmore, D (eds.), *The Biology of Marsupials. Studies in Biology, Economy and Society*, Palgrave, London.

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

Li, H & Durbin, R 2009, 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics,* vol. 25, no. 14, pp. 1754-1760.

Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G & Durbin, R 2009, 'The sequence alignment/map format and SAMtools', *Bioinformatics,* vol. 25, no. 16, pp. 2078-2079.

Li, Y, Chen, Z, Gao, Y, Pan, G, Zheng, H, Zhang, Y, Xu, H, Bu, G & Zheng, H 2017, 'Synaptic adhesion molecule Pcdh-γC5 mediates synaptic dysfunction in Alzheimer's disease', *Journal of Neuroscience,* vol. 37, no. 38, pp. 9259-9268.

Liu, C-C, Kanekiyo, T, Xu, H & Bu, G 2013, 'Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy', *Nature Reviews Neurology,* vol. 9, no. 2, pp. 106-118.

Mahmoud, M, Gobet, N, Cruz-Dávalos, DI, Mounier, N, Dessimoz, C & Sedlazeck, FJ 2019, 'Structural variant calling: the long and the short of it', *Genome biology,* vol. 20, no. 246, pp. 1-14.

Margulies, EH, Maduro, VV, Thomas, PJ, Tomkins, JP, Amemiya, CT, Luo, M, Green, ED & Program, NCS 2005, 'Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 9, pp. 3354-3359.

McAllan, B 2006, 'Dasyurid marsupials as models for the physiology of ageing in humans', *Australian Journal of Zoology,* vol. 54, no. 3, pp. 159-172.

McAllan, B, Hobbs, S & Norris, D 2006, 'Effects of stress on the neuroanatomy of a marsupial', *Journal of Experimental Zoology. Part A, Comparative Experimental Biology,* vol. 305A, no. 154.

Mendez, MF 2019, 'Early-onset Alzheimer disease and its variants', *Continuum (Minneap Minn),* vol. 25, no. 1, pp. 34-51.

Mikkelsen, T, Hillier, L, Eichler, E, Zody, M, Jaffe, D, Yang, S-P, Enard, W, Hellmann, I, Lindblad-Toh, K & Altheide, T 2005, 'Initial sequence of the chimpanzee genome and comparison with the human genome', *Nature,* vol. 437, no. 7055, pp. 69-87.

Mutton, TY, Phillips, MJ, Fuller, SJ, Bryant, LM & Baker, AM 2019, 'Systematics, biogeography and ancestral state of the Australian marsupial genus *Antechinus* (Dasyuromorphia: Dasyuridae)', *Zoological Journal of the Linnean Society,* vol. 186, no. 2, pp. 553-568.

Naylor, R, Richardson, S & McAllan, B 2008, 'Boom and bust: a review of the physiology of the marsupial genus *Antechinus*', *Journal of Comparative Physiology B Biochemical Systemic and Environmental Physiology,* vol. 178, no. 5, pp. 545-562.

Nielsen, H 2017, 'Predicting secretory proteins with SignalP', *Methods in Molecular Biology,* vol. 1611, pp. 59-73.

O'Brien, RJ & Wong, PC 2011, 'Amyloid precursor protein processing and Alzheimer's disease', *Annual Review of Neuroscience,* vol. 34, pp. 185-204.

O'Leary, NA, Wright, MW, Brister, JR, Ciufo, S, Haddad, D, McVeigh, R, Rajput, B, Robbertse, B, Smith-White, B & Ako-Adjei, D 2015, 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research,* vol. 44, no. D1, pp. D733-D745.

Patro, R, Duggal, G, Love, MI, Irizarry, RA & Kingsford, C 2017, 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods,* vol. 14, no. 4, pp. 417.

Pedersen, BS & Quinlan, AR 2017, 'Mosdepth: quick coverage calculation for genomes and exomes', *Bioinformatics,* vol. 34, no. 5, pp. 867-868.

Piovesan, A, Caracausi, M, Antonaros, F, Pelleri, MC & Vitale, L 2016, 'GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics', *Database,* vol. 2016, no. 2016, pp. baw153.

Pomaznoy, M, Ha, B & Peters, B 2018, 'GOnet: a tool for interactive Gene Ontology analysis', *BMC Bioinformatics,* vol. 19, no. 470, pp. 1-8.

Promislow, DEL & Harvey, PH 1990, 'Living fast and dying young: A comparative analysis of life-history variation among mammals', *Journal of Zoology,* vol. 220, no. 3, pp. 417-437.

Reardon, S 2018, 'Frustrated Alzheimer's researchers seek better lab mice', *Nature,* vol. 563, no. 7731, pp. 611-613.

Robinson, MD & Oshlack, A 2010, 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biology,* vol. 11, no. 3, pp. 1-9.

Rosenthal, SL & Kamboh, MI 2014, 'Late-onset Alzheimer's disease genes and the potentially implicated pathways', *Current Genetic Medicine Reports,* vol. 2, no. 2, pp. 85-101.

Salamov, AA & Solovyev, VV 2000, 'Ab initio gene finding in *Drosophila* genomic DNA', *Genome Research,* vol. 10, no. 4, pp. 516-522.

Schwab, C, Hosokawa, M & McGeer, PL 2004, 'Transgenic mice overexpressing amyloid beta protein are an incomplete model of Alzheimer disease', *Experimental Neurology,* vol. 188, no. 1, pp. 52-64.

Shen, W, Le, S, Li, Y & Hu, F 2016, 'SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation', *PloS One,* vol. 11, no. 10, pp. e0163962.

Simão, FA, Waterhouse, RM, Ioannidis, P, Kriventseva, EV & Zdobnov, EM 2015, 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics,* vol. 31, no. 19, pp. 3210-3212.

Sims, R, Hill, M & Williams, J 2020, 'The multiplex model of the genetics of Alzheimer's disease', *Nature Neuroscience,* vol. 23, no. 3, pp. 311-322.

Smit, A, Hubley, R & Green, P 2008-2015, *RepeatModeler Open-1.0*, viewed 19 December 2019, http://www.repeatmasker.org

Smit, A, Hubley, R & Green, P 2013-2015, *RepeatMasker Open-4.0*, viewed 19 December 2019, http://www.repeatmasker.org

Solovyev, V, Kosarev, P, Seledsov, I & Vorobyev, D 2006, 'Automatic annotation of eukaryotic genes, pseudogenes and promoters', *Genome Biology,* vol. 7, no. S1, pp. S10.

Solovyev, VV 2002, 'Finding genes by computer: probabilistic and discriminative approaches', in Tao Jiang, YX, Michael Q. Zhang (ed.), *Current Topics in Computational Molecular Biology*, MIT Press, Cambridge, MA, USA.

Stelzer, G, Rosen, N, Plaschkes, I, Zimmerman, S, Twik, M, Fishilevich, S, Stein, TI, Nudel, R, Lieder, I & Mazor, Y 2016, 'The GeneCards suite: from gene data mining to disease genome sequence analyses', *Current Protocols in Bioinformatics,* vol. 54, no. 1, pp. 1.30.1 - 1.30.33.

Sun, Q, Xie, N, Tang, B, Li, R & Shen, Y 2017, 'Alzheimer's disease: from genetic variants to the distinct pathological mechanisms', *Frontiers in Molecular Neuroscience,* vol. 10, no. 319, pp. 1-14.

Tábuas-Pereira, M, Santana, I, Guerreiro, R & Brás, J 2020, 'Alzheimer's disease genetics: review of novel loci associated with disease', *Current Genetic Medicine Reports,* vol. 8, no. 1, pp. 1-16.

Tasker, EM & Dickman, CR 2001, 'A review of Elliott trapping methods for small mammals in Australia', *Australian Mammalogy,* vol. 23, no. 2, pp. 77-87.

Toder, R, Wakefield, M & Graves, J 2000, 'The minimal mammalian Y chromosome– the marsupial Y as a model system', *Cytogenetic and Genome Research,* vol. 91, no. 1-4, pp. 285-292.

Ulep, MG, Saraon, SK & McLea, S 2018, 'Alzheimer Disease', *The Journal for Nurse Practitioners,* vol. 14, no. 3, pp. 129-135.

UniProt Consortium 2018, 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids Research,* vol. 47, no. D1, pp. D506-D515.

Van Dyck, S & Crowther, M 2000, 'Reassessment of northern representatives of the *Antechinus stuartii* complex (Marsupialia: Dasyuridae): *A subtropicus* sp. nov. and *A. adustus* new status', *Memoirs of the Queensland Museum,* vol. 45, no. 2, pp. 611-635.

Wang, K, Li, M & Hakonarson, H 2010, 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Research,* vol. 38, no. 16, pp. e164-e164.

Weisenfeld, NI, Kumar, V, Shah, P, Church, DM & Jaffe, DB 2017, 'Direct determination of diploid genome sequences', *Genome Research,* vol. 27, no. 5, pp. 757-767.

Wood, D 1970, 'An ecological study of *Antechinus stuartii* (Marsupialia) in a south-east Queensland rain forest', *Australian Journal of Zoology,* vol. 18, no. 2, pp. 185-207.

Woolley, P 1966, 'Reproduction in *Antechinus* spp. and other dasyurid marsupials', *Zoological Society of London,* vol. 15, pp. 281-294.

Wright, BR, Farquharson, KA, McLennan, EA, Belov, K, Hogg, CJ & Grueber, CE 2020, 'A demonstration of conservation genomics for threatened species management', *Molecular Ecology Resources,* vol. 00, pp. 1-16.

Xu, W, Tan, L & Yu, JT 2015, 'The role of PICALM in Alzheimer’s disease', *Molecular Neurobiology,* vol. 52, no. 1, pp. 399-413.

Yang, H & Wang, K 2015, 'Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR', *Nature Protocols,* vol. 10, no. 10, pp. 1556-1566.

Yang, Y & Sauve, AA 2016, 'NAD+ metabolism: Bioenergetics, signaling and manipulation for therapy', *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics,* vol. 1864, no. 12, pp. 1787-1800.

Zhao, Z, Sagare, AP, Ma, Q, Halliday, MR, Kong, P, Kisler, K, Winkler, EA, Ramanathan, A, Kanekiyo, T & Bu, G 2015, 'Central role for PICALM in amyloid-β blood-brain barrier transcytosis and clearance', *Nature Neuroscience,* vol. 18, no. 7, pp. 978-987.

Zheng, GX, Lau, BT, Schnall-Levin, M, Jarosz, M, Bell, JM, Hindson, CM, Kyriazopoulou-Panagiotopoulou, S, Masquelier, DA, Merrill, L & Terry, JM 2016, 'Haplotyping germline and cancer genomes with high-throughput linked-read sequencing', *Nature Biotechnology,* vol. 34, no. 3, pp. 303.

CHAPTER 5

A reference genome and use of next-generation sequencing data to inform long-term conservation management of Australia’s “Easter Bunny”; the greater bilby (*Macrotis lagotis*)

A REFERENCE GENOME AND USE OF NEXT-GENERATION SEQUENCING DATA TO INFORM LONG-TERM CONSERVATION MANAGEMENT OF AUSTRALIA'S "EASTER BUNNY"; THE GREATER BILBY (*MACROTIS LAGOTIS*)

5.1 BACKGROUND

Chapter 5 comprises an unpublished manuscript entitled "A reference genome and use of next-generation sequencing data to inform long-term conservation management of Australia's "Easter Bunny"; the greater bilby (*Macrotis lagotis*)". This chapter focuses on my contribution to the larger bilby genome consortium project which is currently developing a chromosome-length assembly which is due for completion in late 2021. The work presented in this chapter will form part of a larger research publication that brings together a range of other downstream analyses from a broad range of researchers to better understand and conserve Australia's greater bilby. Contributions to the body of work presented in this thesis are detailed at the end of this section.

Chapters 1 and 3 demonstrated how reference genomes and pre-existing genomic datasets add value to the conservation of the endangered Tasmanian devil, while Chapter 4 focused on creating reference genomes for non-threatened species as a resource for closely related threatened counterparts and as a model to investigate biological traits that may provide insights for other species. This chapter aims to showcase a complete end-to-end example of how a variety of genomic resources can be generated and amalgamated to inform the conservation management of a threatened species with limited pre-existing genomic data. Here we create a high-quality reference genome for the vulnerable greater bilby and employ whole genome resequencing (WGR) and reduced representation sequencing (RRS) techniques to answer a suite of conservation management questions. Specifically, we i) employ a hybrid approach to assemble a high-quality reference genome, ii) provide a summary of the current status of the bilby metapopulation using RRS data from all contemporary captive sites and compare this to monitored wild populations from the Pilbara region, iii) assess whether sufficient RRS data can be obtained from wild bilby scat samples for the monitoring of wild populations, iv) use WGR data to develop a panel of sex-linked markers for sexing of wild samples and v) determine whether estimates of genome-wide diversity from RRS data provide a good proxy for functional diversity estimated by WGR data. The results from this study will not only provide a baseline for general nation-wide metapopulation management of the greater bilby, but also provide crucial genomic resources for this species, which can be used for many additional species-specific downstream conservation applications.

The work presented in this thesis chapter would not have been possible without the valuable contributions from Katherine A. Farquharson, Emma Peel, Yuanyuan Cheng, Martin A. Dziminski, Fiona M. Carpenter, Shelley McArthur, Kym Ottewell, Katherine Belov and Carolyn J. Hogg. Katherine Belov and Carolyn J. Hogg conceived the study and obtained funding. Carolyn J. Hogg acquired samples and assisted Katherine A. Farquharson and Emma Peel with DNA extraction from captive samples. Emma Peel performed DNA and RNA extractions for the genome and transcriptome respectively. Yuanyuan Cheng performed DNA extraction for HiFi sequencing and assembled the global transcriptome. Katherine A. Farquharson and Carolyn J. Hogg assisted with WGR and population genetics analyses. Kym Ottewell coordinated analysis of wild samples and conceived the idea of the sex-linked markers with Martin A. Dziminski sourcing wild samples, Fiona M. Carpenter performing DNA extraction from wild samples, and Shelley McArthur performing the downstream testing of sex-linked markers. I performed all other work including genome assembly, genome annotation, alignment of RRS and WGR data, variant calling, variant annotation, sex-linked marker development and all downstream analyses. I wrote the manuscript with feedback and revisions from Katherine A. Farquharson, Carolyn J. Hogg, Catherine E. Grueber and Katherine Belov.

5.2 MAIN MANUSCRIPT

A reference genome and use of next-generation sequencing data to inform long-term conservation management of Australia's "Easter Bunny"; the greater bilby (*Macrotis lagotis*)

Abstract

Characterised by their rabbit-like ears, Australia's own "Easter Bunny", the greater bilby, is an iconic nocturnal marsupial that was once widespread across Australia but is now vulnerable to extinction. In 2019, an updated draft National Recovery Plan was released with a major goal of maintaining genetic diversity and adaptive potential of the greater bilby by managing all bilbies as an interconnected metapopulation. To achieve this goal, modern sequencing technologies, such as reduced representation sequencing, provides an efficient and cost-effective approach for the genetic monitoring of the bilby metapopulation. However, without an annotated reference genome, it is currently unknown whether standard management practices are truly conserving the adaptive potential of the species. This management uncertainty is further exacerbated by the difficulties of obtaining tissue samples from wild individuals due to the species' reticence to enter traps. As a result, there is currently limited data on the status of wild bilby populations across their range. Here we have generated a high-quality reference genome for the greater bilby. Combining this with whole genome resequencing (WGR) and reduced representation sequencing (RRS) data, we have developed a toolset that allows for the interpretation of genome-wide diversity in the species, in addition to developing sex-linked markers. We demonstrate the effectiveness of this toolset by characterising bilby genetic diversity at all contemporary fenced and captive sites within the national metapopulation, in addition to using non-invasive samples for the monitoring of wild bilby populations in the Pilbara region (north-western Australia). Here we clearly demonstrate the importance of foundational genomic resources and their value in developing downstream applications for conservation management of a threatened species.

Introduction

The greater bilby (*Macrotis lagotis*) is a nocturnal, omnivorous marsupial, native to Australia. Following the extinction of the lesser bilby (*Macrotis leucura*) in the mid-1900s (Burbidge et al., 1988), the greater bilby represents the only extant species in the Thylacomyidae family. They are characterised by their long snout, blue-grey fur and long ears. Their resemblance to rabbits encouraged greater bilbies to be dubbed as Australia's own "Easter Bunny" with chocolate bilbies often being sold at Easter in Australia to increase awareness for bilby conservation. The common name 'bilby' was derived from the indigenous Yuwaalayaay language name, 'bilbi' (Abbott, 2001). Bilbies hold a deep cultural significance to indigenous Australians (Commonwealth of Australia, 2019) as a totem for some communities and a food source for others. They are also known as "ecosystem engineers" (Jones, Lawton & Shachak, 1994; Manning, Eldridge & Jones, 2015) due to their burrowing behaviour and their vital role in the Australian ecosystem in maintaining soil health and vegetation growth, as well as their burrows providing shelter/habitat for other species (Dawson et al., 2019; Hofstede & Dziminski, 2017; Read et al., 2008). The greater bilby was once widespread across Australia (Johnson, 2002; Watts, 1969); however, competition and predation by introduced species following European settlement caused a significant (~80%) decline in the distribution of this species (Abbott, 2001; Kennedy, 1992; Southgate & Adams, 1994). Wild populations of the greater bilby are now restricted to a small region in south-west Queensland (Gordon, Hall & Atherton, 1990), and a low-density distribution across Western Australia (Dziminski, Carpenter & Morris, 2020b; Friend, 1990) and the Northern Territory (Johnson & Southgate, 1990). With currently decreasing population trends, the greater bilby is vulnerable to extinction (Burbidge & Woinarski, 2016). As a result, conservation initiatives are crucial to protecting the long-term survival of Australia's "Easter Bunny".

A National Recovery Plan for the greater bilby was published in 2006 with the overarching goal to improve, or at least maintain, the vulnerable conservation status of the bilby through a variety of recovery objectives including: reducing impact of predation, maintaining genetic diversity, establishing self-sustaining populations, monitoring population trends, assessing the impact of threatening processes and informing/involving the community and stakeholders in the recovery process (Pavey, 2006). In 2019, a new draft National Recovery Plan was released with the major objectives of: population growth, habitat maintenance/expansion, maintenance of genetic diversity and a greater role of indigenous people in bilby conservation, by 2029 (Commonwealth of Australia, 2019). For many years, captive populations of NT and WA bilbies were managed separately to QLD bilbies (Jodi Buchecker & Vere Nicolson, 2016). However, a major goal of the latest National Recovery Plan includes management of all captive or fenced bilby populations within Australia to be managed as a single, interconnected metapopulation (Commonwealth of Australia, 2019; Moritz et al., 1997). The metapopulation encompasses a range of management scenarios throughout Australia from intensive zoo facilities, to large predator-free fenced sites and island populations (Lott et al., 2020) (see Methods for more details). Ongoing genetic monitoring of the metapopulation is needed to ensure maintenance of genetic variability within the metapopulation and the long-term conservation of the greater bilby.

In 1997, a panel of nine microsatellites was developed for the greater bilby (Moritz et al., 1997) and was employed to monitor genetic diversity and population structure of the species nationally (Miller et al., 2015; Moritz et al., 1997; Smith, McRae & Hughes, 2009). More recently, reduced representation sequencing (RRS) approaches such as DArTSeq (Diversity Arrays Technology; Jaccoud et al., 2001) and double digest RADseq (ddRAD; Peterson et al., 2012) have been employed to better asses genetic diversity and population structure in captive bilby populations (Lott et al., 2020; Wright et al., 2019b). RRS approaches use restriction enzyme digestion to generate genome-wide high-throughput sequencing data for many individuals in a cost-effective manner and have shown to better reflect genetic diversity than other common marker types when monitoring threatened species (McLennan et al., 2019). Such techniques are often employed to monitor the genetic diversity of populations and assess how diversity can be maximised among populations through events such as targeted translocations between populations (e.g. McLennan et al., 2020). Current conservation genetics methodology assumes that maximising genome-wide diversity of populations will in turn result in maximising of functional diversity (i.e. genetic diversity at gene regions which may have functional implications to the organism and hence may have adaptive potential), giving populations the best chance of long-term survival (Forsman & Wennersten, 2016). Previous studies on the endangered Tasmanian devil (*Sarcophilus harrisii*) have shown that estimates of genome-wide diversity from RRS data provides an accurate representation of overall diversity from whole genome resequencing (WGR) data (Wright et al., 2020). However, it is currently unknown whether genome-wide diversity using RRS data is an effective proxy for functional diversity and whether the management technique of maximising genome-wide diversity based on RRS is enough to arm populations with future adaptive potential.

Bilbies are a cryptic nocturnal arid species that tend to be trap shy and not attracted to bait, making them a difficult species to obtain tissue samples from on a regular basis. Non-invasive sampling techniques are commonly used in wildlife species to assess a range of biological functions including hormonal analyses, microbiome analyses, dietary analyses, and population genetics (for examples see Grueber et al., 2020; Kersey & Dehnhard, 2014; Waits & Paetkau, 2005). Scat samples have been shown to be a viable option for collecting DNA from cryptic wildlife, including red foxes (*Vulpes vulpes*) (Vine et al., 2009), snow leopards (*Panthera uncia*) (Janečka et al., 2008), and koalas (*Phascolarctos cinereus*) (Schultz et al., 2018). Current surveys of wild bilby populations often rely on species specific microsatellite markers using DNA collected from scat samples (Dziminski, Carpenter & Morris, 2020a) but recent studies have shown the limited value of such marker sets as small numbers of loci can result in imprecise measures of genome-wide diversity (McLennan et al., 2019). As a result, our current understanding of both captive and wild bilby population genetics is limited by the use of a microsatellite marker set. This is further exacerbated by the current absence of sex-linked markers for the greater bilby, preventing the ability to assign sex of wild samples. Development of sex-linked markers and exploring whether sufficient RRS data can be obtained from bilby scat samples is crucial to the long-term monitoring of wild bilby populations and for meeting the goals of the National Recovery Plan.

The objective of this study was to develop and use a reference genome for this cryptic threatened marsupial and use this genomic resource to develop a suite of new tools for use in conservation management. Specifically we aimed to i) sequence and annotate a high-quality reference genome, ii) provide a summary of the current status of the bilby metapopulation using high-quality RRS data from all contemporary captive sites and compare this to monitored wild populations from the Pilbara region, iii) assess whether sufficient RRS data can be obtained from wild bilby scat samples, iv) develop a panel of sex-linked markers for sexing of wild samples and v) determine whether estimates of genome-wide diversity from RRS data provide a good proxy for functional diversity in the greater bilby by using whole genome resequencing.

Methods

Reference Genome

Sample Collection

Tissue samples from spleen, liver, lymph node, kidney, heart, tongue, ovary, uterus, pouch skin, mammary gland and salivary gland were harvested from a single female bilby at Perth Zoo that was euthanised due to medical reasons. Tissue samples were collected during post-mortem examination and stored both with and without RNAlater at -80˚C for DNA and RNA extraction respectively. Additionally, 500uL of peripheral blood from a single male bilby housed at Dreamworld QLD was collected into RNAprotect animal blood tube (Qiagen) during routine veterinary health checks and stored at -80˚C.

Genome Sequencing

To assemble a high-quality reference genome, a hybrid approach of 10x Genomics linked-read sequencing (Weisenfeld et al., 2017) and Pacific Biosciences (PacBio) HiFi sequencing (Wenger et al., 2019) was used. For 10x Genomics linked-read sequencing, high molecular weight (HMW) DNA was extracted from 25 mg of spleen tissue using the MagAttract HMW DNA kit (Qiagen). Sample quality control (QC) and library preparation was performed by the Ramaciotti Centre for Genomics (UNSW) prior to sequencing on a NovaSeq 6000 S1 flowcell using 150 bp PE reads and obtaining ~57× coverage. For PacBio HiFi sequencing, HMW DNA was extracted from 100mg of kidney tissue using the using the Nanobind tissue big DNA kit (Circulomics). Sample QC and HiFi library preparation was performed by the Australian Genome Research Facility (AGRF) using the SMRTbellTM Express Template Prep Kit 2.0. Sequencing was performed on the Pacific Biosciences Sequel II system across two SMRT Cells in circular consensus sequencing (CCS) mode obtaining ~10× coverage.

Genome Assembly

HiFi reads were generated using the CCS algorithm in SMRT Link v9.0.0.92188 and assembled using PacBio's Improved Phased Assembler (IPA) v1.1.2 (https://github.com/PacificBiosciences/pbipa). The Purge_dups v1.2.3 (Guan et al., 2020) pipeline was used to remove haplotigs and contig overlaps from both the primary and alternate assemblies. An interleaved linked reads file was created from the raw 10x Genomics reads by running the Long Ranger v2.2.2 (Zheng et al., 2016) basic pipeline and used for alignment to the draft assembly with Burrows–Wheeler Aligner (BWA) mem v0.7.17-r1188 (Li & Durbin, 2009). The output was sorted using samtools v1.9 (Li et al., 2009) and scaffolding was performed using ARCS v1.1.1 (Yeo et al., 2018) and LINKS v1.8.7 (Warren et al., 2015) with the *-D* option to estimate gap sizes. PBJelly v15.8.24 (English et al., 2012) was used for gap filling the scaffolded assembly with default parameters and Pilon v1.20 (Walker et al., 2014) was used to polish the final assembly using the 10x reverse reads that were quality trimmed (trimming parameters: ftl=10 trimq=20 qtrim=rl) using BBDuk v37.98 (Bushnell, 2014). BBTools v38.73 (RRID:SCR_016968) (Bushnell, 2014) was used to generate general assembly statistics and assembly completeness was assessed using mammalian Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2 (4,104 genes) and v4.0.6 (9,226 genes). Majority of the genome assembly pipelines were performed on a single cloud instance with polishing performed in a HPC environment (Table S1).

Transcriptome Sequencing

Total RNA was extracted from 25mg of each tissue (excluding blood) using the RNeasy Mini Kit (Qiagen) and from blood using the RNAprotect animal blood kit (Qiagen). Contaminating genomic DNA was removed through an on-column digestion using the RNase-free DNase I set (Qiagen). RNA was quantified using the Bioanalyzer RNA 6000 Nano Kit (Agilent Technologies) prior to TruSeq stranded total RNA library preparation, with ribosomal RNA depletion using the Illumina Ribo-zero gold kit at the Ramaciotti Centre for Genomics (UNSW). Tissue libraries (excluding blood) were sequenced as 150 bp PE reads across a single S1 flowcell on the Illumina NovaSeq6000. The blood library was sequenced as 75 bp PE reads across four lanes of a HO flowcell on the Illumina NextSeq 500.

Transcriptome Assembly

Raw RNA-seq reads (~100M reads per sample) underwent quality and length trimming using Trimmomatic v0.38 (Bolger, Lohse & Usadel, 2014) in paired-end mode, with the parameters ILLUMINACLIP:2:30:10, SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5 and MINLEN:25. FastQC v 0.11.8 (Andrews, 2014) was used to assess sequence quality of both raw and trimmed reads. Trimmed reads from each of the 12 tissues were then aligned to the genome using HISAT2 v2.1.0 (Kim et al., 2019) and alignments were converted and sorted using samtools. Transcripts were assembled using StringTie v2.1.3 (Pertea et al., 2015) and the resulting transcript models across the tissues were merged into a single global transcriptome using TAMA merge v0.0 (Kuo et al., 2020) with a splice junction threshold of 3, a 3-prime threshold of 500 and the option to merge duplicate transcript groups. The resultant transcripts were then filtered by removing transcripts with weak evidence (only found in one tissue and FPKM < 0.1) or single-exon transcripts with either low read support or low coding potential. Coding potentials of transcripts were determined using CPC2 v2.0 (Kang et al., 2017). Transcriptome completeness was assessed using BUSCO as above. All transcriptome analyses were performed in a HPC environment (Table S1).

Genome Annotation

RepeatModeler v2.0.1 (Smit, Hubley & Green, 2008-2015) was used to create a custom repeat database for the bilby and RepeatMasker v4.0.6 was used to mask repeats (excluding low complexity regions and simple repeats) (Smit, Hubley & Green, 2013-2015). Genome annotation was performed using Fgenesh++ v7.2.2 (Salamov & Solovyev, 2000; Solovyev et al., 2006; Solovyev, 2002) using general mammalian pipeline parameters and an optimised gene finding matrix for another marsupial species (*Sarcophilus harrisii*). Candidate coding regions within the transcriptome were identified with TransDECODER v2.0.1 and complete transcripts with the longest open reading frame per gene were extracted for mRNA-based gene predictions. A high-quality non-redundant metazoan protein dataset from NCBI was used for homology-based gene predictions using the “prot_map” method and *ab initio* gene predictions were performed in regions where no genes were predicted by other methods (i.e., mRNA mapping or protein homology). All genome annotation steps were performed on a single cloud instance (Table S1).

The predicted gene annotations were then filtered by removing any mRNA-based annotations where the location predicted by Fgenesh++ did not match the location of the reference-guided transcriptome assembly and removing any *ab initio* predictions that did not have strong BLAST hits to the non-redundant metazoan protein database. Final annotations were used in BLAST searches against the Swiss-Prot database and against the RefSeq proteome of the current highest quality marsupial assembly (*Sarcophilus harrisii*) with an e-value cut-off of 1e-5 to generate putative gene names. Completeness of the final gene set was assessed using BUSCO.

Population Genetics

Study Sites

As this study aimed to capture current levels of genetic diversity across the national greater bilby metapopulation and compare this to the wild, samples were obtained from all 18 contemporary captive sites (zoo-based through to semi-wild enclosures) across Australia (ranging from temperate to arid regions; Figure 5.1; Table 5.1) as well as from a number of monitored wild populations in the Pilbara region of Western Australia (Dziminski, Carpenter & Morris, 2020a) (Fig 1. Table 5.1). Nine of the captive zoo/wildlife park populations (Adelaide Zoo, Alice Springs Desert Park, Cleland Wildlife Park, Currumbin Wildlife Sanctuary, Dreamworld, Kanyana Wildlife Park, Monarto Zoo, Taronga Zoo and Taronga Western Plains Zoo) are members of the Zoo and Aquarium Association Australasia (ZAA) (Figure 5.1; Table 5.1) which established the first managed captive breeding colony in 1979 with 19 founders (Christie, 1991). The ZAA sites were originally managed as two separate management units, the WA/NT population and the QLD population. In 2016, these populations were amalgamated and managed as a single metapopulation with annual translocations between sites based on a pedigree-based mean kinship minimisation strategy (Ballou & Lacy, 1995; Jodi Buchecker & Vere Nicolson, 2016). The ZAA managed captive breeding population was subsequently used as a source population for the establishment of large, fenced enclosures such as Yookamurra, Arid Recovery and Venus Bay, as well as the Thistle Island population (Fig1; Table 5.1). The fenced Currawinya population is managed by the Save the Bilby Fund (STBF; Charleville, Queensland) and was recently re-founded by individuals from the ZAA population in 2019 and 2020; after a predator invasion in 2012 that reduced the population by 97% (Bradley et al., 2015). The remaining Australian Wildlife Conservancy (AWC) sites including the early established Scotia, and the more recently established Mt Gibson, Pilliga and Mallee Cliffs populations were founded by multiple source populations (Figure 5.1; Table 5.1).

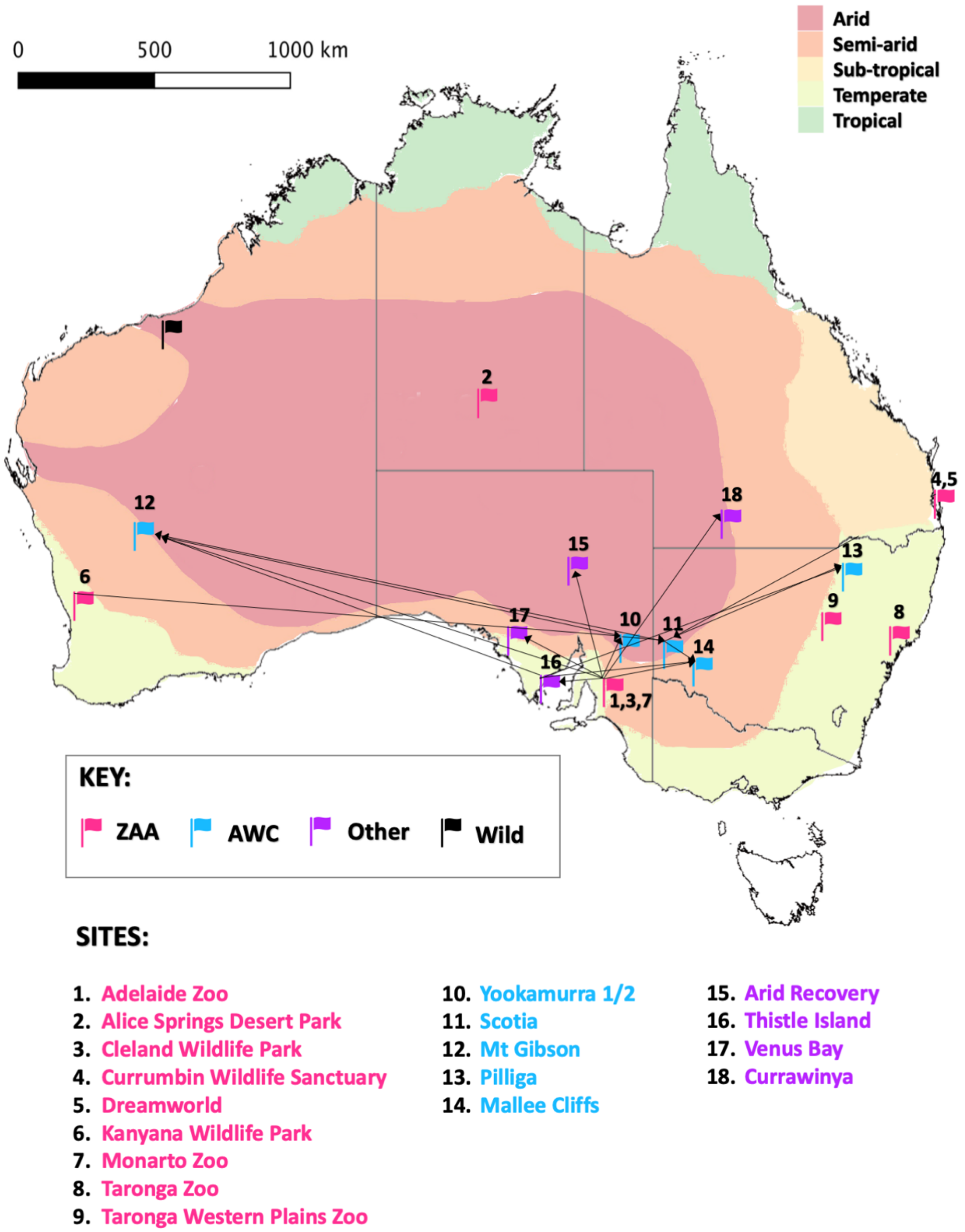


Figure 5.1 Map displaying the 18 captive sites and wild Pilbara location (black) where RRS and WGR bilby samples were sourced. Pink sites are managed by the Zoo and Aquarium Association Australasia (ZAA), Blue sites are managed by the Australian Wildlife Conservancy (AWC) and Purple sites are managed by other organisations such as state government and the Save the Bilby Fund. Black arrows indicate translocations between sites. Note that all ZAA sites have been managed as a single metapopulation since 2016 so translocations from the ZAA metapopulation are represented by a single arrow from the 1,3,7 site flag. Climate key based on (Doherty et al., 2019). See Table 5.1 for more information.

Table 5.1 Summary of bilby sites and respective samples included in the RRS and WGR analysis. Refer to Figure 5.1 for site locations.

Site Number	Population Name	Managed By*	Year Established	Source Population (no. trans-located)	Size (Hectares)	Climate	No. Samples	No. Technical Replicates	No. Plate Replicates
1-9	ZAA	ZAA	-	-	-	Arid/ Temperate	44	0	1
10	Yookamurra 1/2	AWC	1996	ZAA (8)	1,000	Semi-arid	20/3	0/2	4/0
11	Scotia	AWC	1997	ZAA (7) + Yookamurra (7)	65,000	Semi-arid	11	0	0
12	Mt Gibson	AWC	2016	Scotia (16) + Thistle Island (20) + Yookamurra (12) + ZAA (8)	131,812	Semi-arid	26	1	0
13	Pilliga	AWC	2018	Scotia (30) + Thistle Island (30)	5,800	Temperate	36	0	1
14	Mallee Cliffs	AWC	2019	Scotia (10) + Thistle Island (30) + ZAA (10)	480	Semi-arid	50 (5 WGR)	24	2
15	Arid Recovery	SA Gov + Others	2000	ZAA (32)	12,300	Arid	16	0	0
16	Thistle Island	SA Gov	1997	ZAA (21)	6,800	Temperate	29	18	5
17	Venus Bay	SA Gov	2001	ZAA (23)	6,300	Temperate	5	4	0
18	Currawinya	STBF	2019 (re-est.)	ZAA (26)	2,500	Arid	35 (1 WGR)	16	6
Wild	Pilbara	-	-	-	34,400,000	Arid/Semi-Arid	9 (13 scat)	4 (9 scat)	0 (3 scat)

*Sites managed by ZAA were grouped for analysis due to their long-term management as a single population and as such their year of establishment, source population and size is not included in the table.

Sample Collection and RRS Sequencing

A total of 298 bilbies sampled between 2011 and 2020 were included in the population genetics analysis. We utilised pre-existing RRS (DArTseq) data (Lott et al., 2020; Wright et al., 2019b) for 72 individuals that were previously sampled from the Yookamurrra Wildlife Sanctuary breeding enclosure in 2011 (Yooka1; N = 20), and a variety of captive sites in 2016 including Mt Gibson (N = 19; founders only), Scotia Wildlife Sanctuary (N = 11) and the ZAA metapopulation (N = 22 from sites including Adelaide Zoo, Alice Springs Desert Park, Cleland Wildlife Park, Currumbin Wildlife Sanctuary, Dreamworld, Kanyana Wildlife Rehabilitation Centre, Monarto Zoo and Taronga Zoo) (Figure 5.1, Table 5.1). An additional 212 ear biopsy samples were collected from a variety of pre-existing and newly established fenced reserves between 2016 and 2020 including Arid Recovery Reserve (N = 16), Yookamurra Wildlife Sanctuary main enclosure (Yooka2; N = 3), Currawinya (N = 35; founders + first generation), Mallee Cliffs (N = 50; founders only), Mt Gibson (N = 7; founders only), Pilliga (N = 36; founders only), Thistle Island (N = 29), Venus Bay (N = 5), Taronga Western Plains Zoo (N = 22; founders + first generation). DNA was extracted using the DNeasy blood and tissue kit (Qiagen) or the MagAttract HMW DNA Kit (Qiagen). To better understand the genetic diversity of wild bilby populations, ear-biopsy samples from nine individuals were collected from monitored wild populations in the Pilbara region of Western Australia (Dziminski, Carpenter & Morris, 2020b) (Figure 5.1, Table 5.1) with DNA extraction performed using a standard salting out extraction protocol (Sunnucks & Hales, 1996) with the addition of 3µL 10 mg/ml RNase to the TNES buffer to remove RNA contamination. As a pilot study, scat samples from 13 individuals were collected from the wild Pilbara region (Figure 5.1, Table 5.1) and DNA extraction was performed using the QIAamp DNA Stool Mini Kit (Qiagen) to determine the reliability of RRS data when using non-invasive sampling. All extracted DNA samples were sent to DArTseq Pty Ltd for RRS using a pstI-sphI enzyme combination (Jaccoud et al., 2001). Samples were uniquely barcoded and multiplexed and the resultant fragments were sequenced on a HiSeq 2500 as 77-bp single-end reads.

RRS Variant Calling

Variants from the RRS data were called and filtered using previously published methods (Wright et al., 2019a; Wright et al., 2019b) on a single cloud instance (Table

S1). Briefly, reads were cleaned and demultiplexed using process_radtags in Stacks v2.5.3 (Catchen et al., 2013; Catchen et al., 2011) and the first 5 bases of the raw DArTseq reads were trimmed using BBDuk. Resulting reads were aligned to the bilby reference genome using BWA samse with output converted and sorted using samtools. Variants were called using the Stacks ref_map pipeline outputting one random SNP per locus with a minimum minor allele frequency (MAF) of 0.01 and a minimum call rate of 30%. Additional filtering for minimum average allelic depth (2.5), allelic coverage difference (≤80%), locus heterozygosity (≤90%) and reproducibility (≥90% matching genotypes between technical replicates) was performed in R v 3.6.2 using a previously published pipeline (Wright et al., 2019a). Final variant positions were annotated based on the reference genome annotation using ANNOVAR v 20191024 (Wang, Li & Hakonarson, 2010; Yang & Wang, 2015).

WGR and Variant Calling

DNA samples from 12 individuals (7 males and 5 females, Table 5.1) that were previously extracted for the RRS analysis underwent sample QC, library preparation and WGR at the Ramaciotti Centre for Genomics (UNSW). Samples were sequenced as 150 bp PE reads across a single S2 flowcell on the Illumina NovaSeq6000 obtaining ~30× coverage per sample.

Due to the size of WGR datasets, all analyses were run in a HPC environment (Table S1) with data being split and run in parallel chunks where possible. Raw paired fastq files were split into smaller files of approximately 500,000 read pairs with fastp v0.20.0 (Chen et al., 2018) to improve computational efficiency. Split files were aligned to the bilby reference genome in parallel using BWA mem with split hits marked as secondary using the -M flag. Resulting BAM files were merged into sample-level BAMs using Sambamba v0.7.1 (Tarasov et al., 2015) and duplicate reads were marked with SAMBLASTER v0.1.26 (Faust & Hall, 2014) prior to standard sorting and indexing with Samtools.

As there is currently no “known variants” resource for the greater bilby that can be used to improve the accuracy of the variant calls using base quality score recalibration (BQSR), a bootstrapping approach was employed instead with the Genome Analysis Toolkit (GATK) v4.1.2.0 (DePristo et al., 2011; Poplin et al., 2017; Van der Auwera et al., 2013). First, GATK SplitIntervals was used to define 3,200 evenly sized genomic intervals. GATK HaplotypeCaller was then run in parallel with default settings on the duplicate-marked BAMs to perform an initial round of single nucleotide polymorphism (SNP) and insertion-deletion (indel) calling. Subsequent interval GVCFs were gathered into sample-level GVCFs using GATK GatherVCFs. Resulting GVCFs were consolidated with GATK GenomicsDBImport for joint-sample variant calling at the 3,200 pre-defined intervals using GATK GenotypeGVCFs. Multi-sample interval calls were then merged into a single genome-wide VCF with GatherVCFs.

SNPs and indels were extracted and filtered separately from the multi-sample genome-wide VCF following GATK best practices (DePristo et al., 2011; Van der Auwera et al., 2013) to select the most likely true positive variants. The filtered variants were used as “known variants” for base recalibration. Base recalibration with GATK BaseRecalibrator was run in parallel across 36 evenly sized genomic intervals of minimum 100 Mb created with GATK SplitIntervals. Recalibration results from each interval were merged for each sample using GATK GatherBQSRReports and were used to recalibrate the BAM files. Recalibrated sample BAM files (including unmapped reads) were generated in parallel by running GATK ApplyBQSR over discrete genomic intervals and then merging the interval BAMs using GATK GatherBamFiles. BQSR results were assessed with GATK AnalyseCovariates before performing a subsequent final round of short variant calling based on the recalibrated BAMs following the same process of parallel GATK HaplotypeCaller, GatherVCFs, parallel GenomicsDBImport and GenotypeGVCFs, then a final merge with GatherVCFs to produce the recalibrated multi-sample VCF file.

Final filtering and variant annotation were performed on a single cloud instance (Table S1). Variants were filtered according to GATK best practices (DePristo et al., 2011; Van der Auwera et al., 2013) including phred-scaled quality score ≥ 30, quality by depth ≥ 2, Fisher strand bias ≤ 60, strand odds ratio ≤3, root mean square mapping quality ≥ 40, mapping quality rank sum ≥ -12.5 and read position rank sum ≥ -8 for SNPs; and phred-scaled quality score ≥ 30, quality by depth ≥ 2, fisher strand bias ≤ 200 and read position rank sum ≥ -20 for indels. Bcftools v1.11 (Li et al., 2009) was used to split multi-allelic variant calls and to left-normalise the variants prior to variant annotation with ANNOVAR. Genotyping rates were calculated using PLINK.

Sex-linked Marker Development

To develop presence/absence sex-linked markers, putative Y-chromosome sequence data was obtained by extracting WGR reads that were unmapped to the female reference genome from the 7 male WGR samples using samtools and bedtools v2.29.2 (Quinlan & Hall, 2010). Velvet assembler v1.2.10 (Zerbino & Birney, 2008) was used to assemble these reads into longer contigs separately for each individual with a k-mer size of 31, min contig length of 100 bp and expected coverage calculated automatically. Twenty known marsupial Y-chromosome genes and their respective X homologs from a previous study (Cortez et al., 2014) were then used in BLAST searches against the resultant contigs. Contigs with strong BLAST matches (1e-10) to marsupial Y genes (but not the respective X chromosome homologs), were deemed as putative Y chromosome sequence. Intronic regions with no BLAST matches or very weak BLAST matches to the female reference genome were identified in in order to prevent amplification of non-target sequences and ensure sex could easily be assigned by presence/absence of the markers in males/females respectively. PCR primers were designed within these regions using Oligo v6 (Offerman & Rychlik, 2003), aiming for a product size of ~150-200 bp (Table 5.2). The panel of Y-linked markers were first tested on DNA extracted from tissue samples of 8 known sex individuals. PCRs were carried out using the Multiplex PCR Plus Kit (Qiagen). Thermocycling conditions followed a protocol of 15 min at 95°C, then 35 cycles of 30 s at 94°C, 90 s at 60˚C and 60 s at 72°C, with a final extension of 60°C for 30 min. Resultant fragments were run on a 3% agarose gel at 80V and stained with GelRed (Biotium) to be visualised under UV light. Presence of a distinct band indicated amplification of Y-chromosome sequence and therefore assigned males, while absence of a distinct band assigned females. Two of the best (i.e., clearest) markers (KDM5D.2 and HCFC1.2) were then tested on DNA extracted from scat samples of 10 known sex individuals using the same methods as above to confirm whether the markers could be implemented for non-invasive sex determination of wild individuals.

Table 5.2 Primer information for the 6 Y-linked markers.

Y Marker	Forward Primer (5'-3')	Reverse Primer (5'-3')	Product Size (bp)	Optimal Annealing Temperature (˚C)
HUWE1	ACATGGGCTAAGGGTGAATG	TACTTCCTCGCCTAAATAACAG	170	49.9
KDM5D.1	AGTTGGGATATGGAAACATTG	ATCTCCTGGATTGGCTTCTG	226	49.6
KDM5D.2	TTGTCCCAAATGTTCTAAGC	GTTGGCAATACAGAAAGAGG	155	47.2
HCFC1.1	TTGTTTGTGGAGCAGGAGAG	TTACCCTTCCCTATTCTTCC	152	48.5
HCFC1.2	ATCCTGCAATTATTGTTTATG	TATGGTTATAAACTAGCATGTG	132	46
HSFY	TAGGCAATAACAGAGCTGTC	ACTAACATAATGAAAGGTATTC	224	46.2

Population Genetics Analysis

The filtered variants for the 298 RRS samples (excluding scat samples) were converted to adegenet v2.1.3 (Jombart, 2008; Jombart & Ahmed, 2011) format and analysed in R using a standard laptop. All ZAA samples were grouped into a single population for further analyses due to their long-term management as a single population. All wild Pilbara samples were also classed as a single wild population for analysis due to the small sample size. To assess within population genetic diversity, observed heterozyosity (H_O), expected heterozygosity (H_E) and allelic richness (A_R) corrected for sample size (rarefied sample size: N = 3) were calculated using the hierfstat v0.5-7 package (Goudet, 2005) and the diveRsity v1.9.90 package (Keenan, 2017) was used to calculate population inbreeding coefficients (F_{IS}) with 95% confidence intervals from 1000 bootstrap iterations.

Principal Coordinate Analysis (PCoA) was performed using i) all SNPs and ii) only putatively "functional" SNPs (i.e. variants resulting in nonsynonymous mutations) to explore genetic variation among populations and determine whether functional diversity follows similar population-level patterns to that of whole-genome diversity. Divergence between sites was further examined by calculating pairwise fixation indexes (F_{ST}) between populations using 95% confidence intervals (CIs) from 2000 bootstrap iterations with the package StAMPP v1.6.1 (Pembleton, Cogan & Forster, 2013). Fixed alleles between populations were explored using the dartR v1.8.3 package (Gruber et al., 2018). The package related v1.0 (Pew et al., 2015) was used to calculate pairwise relatedness using TrioML, the triadic likelihood relatedness estimate which accounts for inbreeding (Wang, 2007). This estimator was chosen as it was found to have the smallest variance and highest correlation (Pearson correlation coefficient) with the simulated true mean of all estimators simulated in COANCESTRY v1.0 (Wang, 2011) for this dataset, including TrioML (Wang, 2007), Wang (Wang, 2002), Li & Lynch (Li, Weeks & Chakravarti, 1993), Lynch & Ritland (Lynch & Ritland, 1999), Ritland (Ritland, 1996), Queller & Goonight (Queller & Goodnight, 1989) and DyadML (Milligan, 2003) (see Blouin et al., 1996; Hogg et al., 2019) for further details).

Finally, to determine whether the approach of maximising genome-wide functional diversity based on RRS data also results in maximisation of functional diversity, RRS data was compared to WGR data across the 12 individuals that were sampled using both techniques. First, multilocus heterozygosity (MLH) was calculated across all 11,750 RRS SNPs for each individual by dividing the total number of heterozygous loci by the total number of loci typed in the individual. MLH was also calculated in the same way for the 12 WGR samples both at all variants and only nonsynonymous variants. A linear regression in R was used to assess the relationship between RRS MLH (predictor variable) and WGR MLH (response variable) at all variants across the 12 samples to confirm whether RRS data provides an accurate representation of individual genome-wide diversity overall. A second linear regression was then used to assess the relationship between MLH at functional (nonsynonymous) variants (predictor variable) and MLH at all variants (response variable) using the WGR data to determine whether there was a significant relationship between functional diversity and genome-wide diversity.

Results

Reference Genome

The bilby reference genome is 3.69 Gb in size which is larger than other marsupial genomes currently available e.g., Tasmanian devil (*Sarcophilus harrisii*): 3.087 Gb, koala (*Phascolarctos cinereus*): 3.193 Gb, brown antechinus (*Antechinus stuartii*): 3.187 Gb, gray short-tailed opossum (*Monodelphis domestica*): 3.598 Gb (O'Leary et al., 2015). The assembly is comprised of 3,348 scaffolds and 6,668 contigs with a scaffold N50 of 2.22 Mb and a contig N50 of 0.91 Mb (Table 5.3). The assembly showed high completeness with 0.87% gaps and 92% complete mammalian BUSCOs (Table 5.3). The global transcriptome was composed of 39,883 genes and 304,184 isoforms with an average transcript length of 6,870 bp and an N50 of 13.5kb (Table 5.3). 16,654 of these transcripts (representing isoforms which contained the longest complete ORF) were used as mRNA-based evidence for genome annotation. 51.67% of the genome was masked as repetitive and a total of 28,488 genes were annotated in the genome of which 16,277 were based on mRNA evidence, 1,479 were based on protein homology evidence and the remaining were made *ab initio* (Table 5.3).

Variant Calling

Variant calling and filtering of the RRS data resulted in a total of 11,750 high-confidence SNPs. Average genotyping rate was 91.2% for tissue samples and 53.5% for scat samples. Reproducibility was high, with an error rate between technical replicates of 0.95% for tissue samples and 1.73% for scat samples. Following variant annotation, 140 SNPs (1.2%) were found to be putatively functional (nonsynonymous) across 134 genes (Table 5.4, Figure S1a). Genotyping rate at these SNPs was 91.7% for tissue samples and 85.8% for scat samples.

Joint genotyping of the 12 WGR samples followed by BQSR and filtering resulted in 30,730,316 SNPs and 12,867,448 indels with an average genotyping rate of 88%. Variant annotation identified 92,375 (0.2%) nonsynonymous SNPs across 18,377 genes (Table 5.4, Figure S1b).

Table 5.3 Bilby reference genome and transcriptome statistics.

Statistic	Value
Genome	
Genome Size	3.69 Gb
No. Scaffolds	3,348
No. Contigs	6,668
Scaffold N50	2.22 Mb
Contig N50	0.91 Mb
Gaps	0.87%
Repeat Content*	51.67%
BUSCO V3	C:92.0%[S:88.7%,D:3.3%],F:3.3%,M:4.7%,n:4104
BUSCO V4	C:85.1%[S:82.3%,D:2.8%],F:2.0%,M:12.9%,n:9226
No. Protein-coding Genes	28,488
BUSCO V3 (proteins)	C:83.7%[S:80.6%,D:3.1%],F:12.0%,M:4.3%,n:4104
BUSCO V4 (proteins)	C:73.2%[S:70.5%,D:2.7%],F:9.3%,M:17.5%,n:9226
Global Transcriptome	
No. Transcripts	304,184
No. Genes	39,883
Average Transcript Length	6,870 bp
Transcript N50	13.5 Kb
BUSCO V3	C:93.4%[S:11.5%,D:81.9%],F:3.6%,M:3.0%,n:4104
BUSCO V4	C:82.1%[S:11.8%,D:70.3%],F:3.6%,M:14.3%,n:9226

*Excluding low complexity and simple repeats

Table 5.4 Summary of variants called from reduced representation sequencing (RRS) data vs whole genome resequencing (WGR) data.

Dataset (no. samples)	Average Genotyping Rate	Total Filtered Variants	Non-synonymous (ns) Variants*	No. Genes with ns Variants
RRS (298)	91.2%	11,750 SNPs	140	134
WGR (12)	88%	30,730,316 SNPs, 12,867,448 indels	92,375	18,377

*Nonsynonymous variants are more likely to have functional consequences on the species and therefore represent “functional” diversity.

Sex-linked Marker Development

Using the male WGR data and the female reference genome (see Methods), Y-chromosome contigs associated with 7 marsupial Y genes (*HUWE1*, *KDM5D*, *MECP2*, *RBM10*, *UBE1Y*, *HCFC1* and *HSFY*) were identified through BLAST searches, totalling ~20 kb worth of Y-chromosome sequence. Contigs associated with three genes (*MECP2*, *RBM10* and *UBE1Y*) showed potential for some non-specific binding when BLASTed against the female reference genome so were excluded from further analysis. Six primer sets were designed within intronic regions of the remaining four genes (*HUWE1*, *KDM5D*, *HCFC1* and *HSFY*; Table 5.2). The six markers were first tested on DNA extracted from tissue samples of eight known-sex individuals through PCR and gel electrophoresis (see Methods) and all six markers successfully identified sex with a strong band present in males and absent in females (Figure 5.2a). The two best markers (KDM5D.2 and HCFC1.2) were then tested on DNA extracted from scat samples of 10 known sex individuals and were found to successfully identify sex from these non-invasive samples (Figure 5.2b).

Population Genetics Analysis

Observed heterozygosity across the 12 populations ranged from 0.1 (Yooka1) to 0.17 (Currawinya) with majority of populations (except Yookamurra 1/2 and Venus Bay) exhibiting lower observed heterozygosity than expected under HWE (Table 5.5). In concordance with the observed excess of homozygosity, mean individual inbreeding (F_{IS}) was 0.034 (±0.032 SE) overall with a number of populations showing statistically significant F_{IS} values, particularly the Mallee Cliffs and Mt Gibson populations (Table 5.5). These results show some deviation to those found previously (Lott et al., 2020), likely due to the differences in sample size within populations, the larger SNP set (1,067 vs 11,750 SNPs here), and several of the recently established populations being from divergent source populations (see De Meeûs, 2018). Average A_R was 1.15 (±0.008 SE) overall (Table 5.5) with Currawinya and Mallee Cliffs showing the highest allelic richness (1.18), likely due to these sites being recently established by genetically diverse source populations (Figure 5.3a, Table 5.1). Average relatedness within populations was 0.31 (±0.07 SE) overall, ranging from 0.09 (ZAA and Mallee Cliffs) to 0.79 (Yooka1). Interestingly, the wild Pilbara samples showed the greatest deviation in observed vs expected heterozygosity and the highest estimate of inbreeding; however, relatedness was relatively low (Table 5.5).

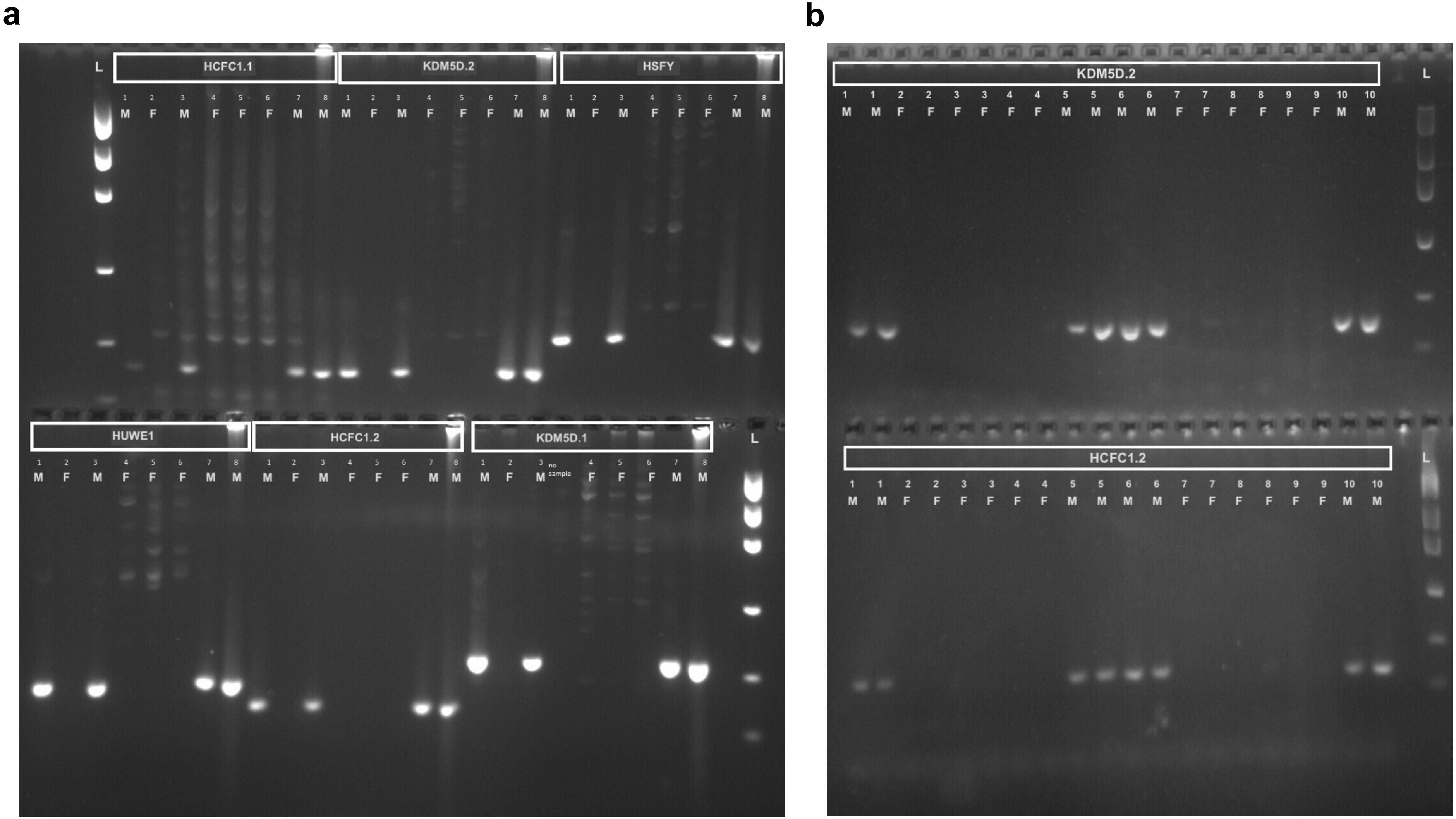

a
b
L
HCFC1.1
KDM5D.2
HSFY
1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8
M F M F F F M M M F M F F F M M M F M F F F M M
HUWE1
HCFC1.2
KDM5D.1
L
1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 1 2 3 no sample 4 5 6 7 8
M F M F F F M M M F M F F F M M M F M F F F M M
KDM5D.2
L
1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9 9 10 10
M M F F F F F F M M M M F F F F F F M M
HCFC1.2
L
1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9 9 10 10
M M F F F F F F M M M M F F F F F F M M

Figure 5.2 Agarose (3%, TBE) gel electrophoresis of the Y-linked markers (HUWE1, KDM5D.1, KDM5D.2, HCFC1.1, HCFC1.2, HSFY) tested on DNA extracted from a) tissue samples and b) scat samples of known sex individuals (M = male, F = female). Numbers indicate different individuals. L: Low DNA mass ladder (Invitrogen). Refer to Table 5.2 for primer information.

Table 5.5 Population genetic diversity statistics based on 11,750 SNPs including: sample size (N), average genotyping rate across samples (Geno), allelic richness (A_R) mean expected heterozygosity (H_E), mean observed heterozygosity (H_O), mean individual inbreeding coefficients (F_{IS}) with 95% confidence intervals (CI), and mean triadic likelihood relatedness estimate (Rel) within each population (within) and between the population and all other populations (between).

Population	**N**	**Geno**	**A_R**	**Mean H_E (SE)**	**Mean H_O (SE)**	**F_{IS}***	**95% CI**	**Rel within (SE)**	**Rel between (SE)**
Arid Recovery	16	95.4%	1.17	0.17 (0.046)	0.16 (0.047)	0.019	-0.02,0.06	0.21 (0.002)	0.03 (0.001)
Currawinya	35	95.8%	1.18	0.18 (0.027)	0.17 (0.027)	**0.050**	**0.02,0.08**	0.13 (0.002)	0.03 (0.001)
Mallee Cliffs	50	97.7%	1.18	0.18 (0.022)	0.15 (0.020)	**0.140**	**0.11,0.17**	0.09 (0.001)	0.06 (0.002)
Mt Gibson	26	83.3%	1.15	0.16 (0.034)	0.13 (0.031)	**0.159**	**0.1,0.19**	0.28 (0.003)	0.08 (0.003)
Pilbara	9	86.5%	1.16	0.16 (0.064)	0.12 (0.054)	**0.205**	**0.05,0.32**	0.19 (0.005)	0.04 (0.001)
Pilliga	36	97.7%	1.17	0.17 (0.027)	0.16 (0.027)	**0.084**	**0.04,0.12**	0.12 (0.001)	0.02 (0.001)
Scotia	11	75.5%	1.12	0.12 (0.057)	0.11 (0.058)	0.053	-0.03,0.16	0.48 (0.005)	0.05 (0.004)
Thistle Island	29	99.0%	1.17	0.17 (0.033)	0.16 (0.034)	**0.027**	**0,0.05**	0.17 (0.003)	0.02 (0.001)
Venus Bay	5	98.8%	1.12	0.12 (0.089)	0.12 (0.101)	-0.082	-0.44,0.19	0.74 (0.049)	0.04 (0.001)
Yooka1^	20	75.7%	1.10	0.10 (0.039)	0.10 (0.046)	-0.090	-0.15,-0.02	0.79 (0.006)	0.01 (0.001)
Yooka2	3	98.9%	1.14	0.14 (0.129)	0.14 (0.147)	-0.191	-0.64,-0.02	0.47 (0.019)	0.05 (0.001)
ZAA	44	90.8%	1.17	0.17 (0.024)	0.16 (0.024)	**0.034**	**0,0.06**	0.09 (0.002)	0.06 (0.001)

Standard errors (SE) are given in parentheses.

*Bold values indicate populations where inbreeding estimates and associated 95% CIs are positive, indicating statistically significant (α = 0.05) inbreeding within the population.

^Note that Yooka1 samples were collected in 2001 but likely reflect contemporary genetic variation due to the isolated management of this population

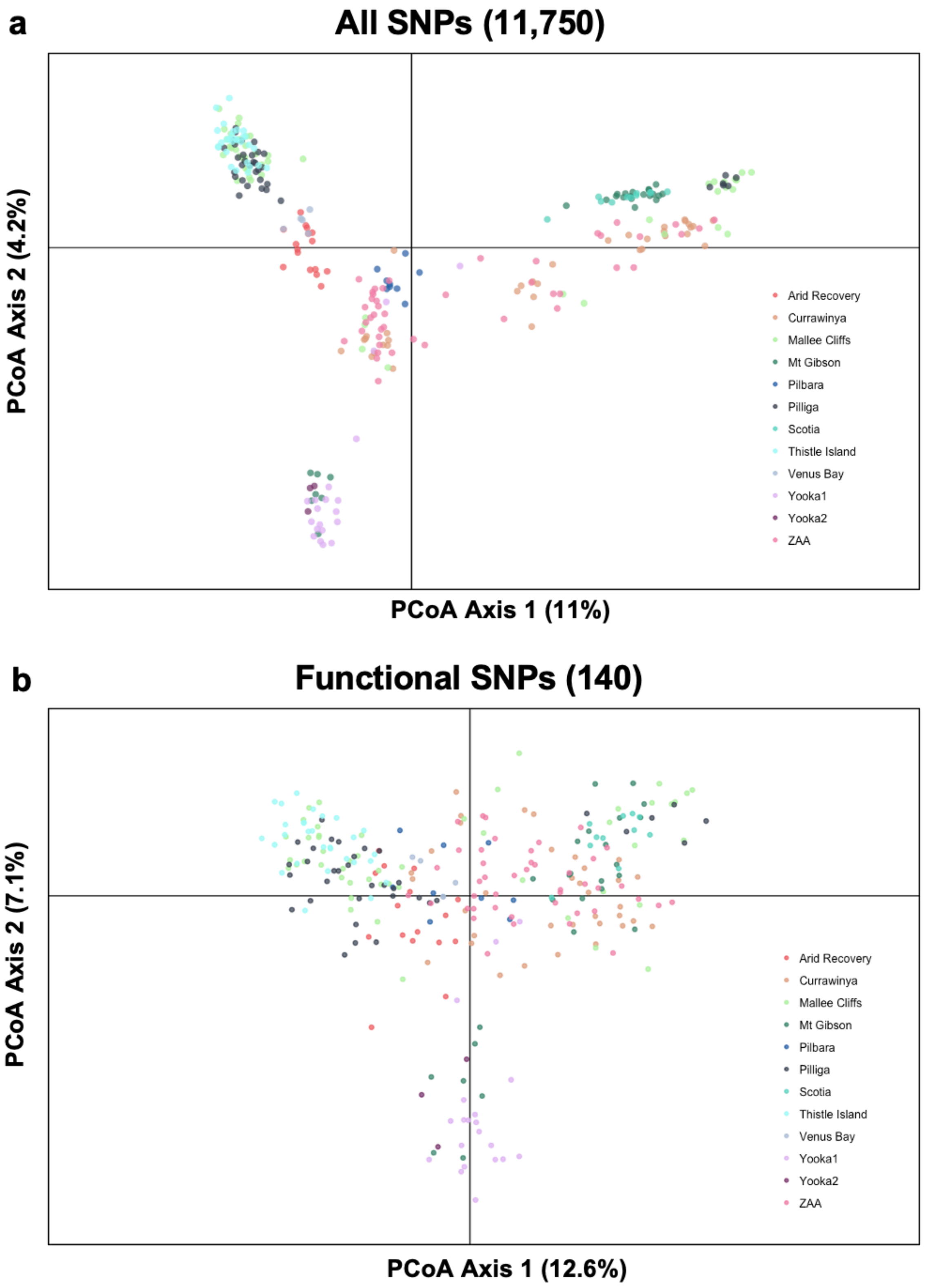


Figure 5.3 PCoAs using a) all 11,750 SNPs and b) 140 nonsynonymous (putatively functional) SNPs from RRS data showing genetic variation of 298 bilbies across 12 populations.

The PCoA showed evidence of population stratification in line with previous findings (Lott et al., 2020) and what would be expected based on the demographic and translocation history of the bilby populations (Figure 5.3a). For example, two of the earliest established sites Yookamurra and Scotia are genetically distinct as shown in previous research, likely due to majority of the founding individuals of Scotia originally being from QLD descent (Lott et al., 2020). The ZAA metapopulation shows high within-population genetic variation due to the NT/WA and QLD population amalgamation in 2016 and the long-term management of multiple sites across Australia as a single metapopulation (Figure 5.3a). ZAA also showed genetic similarity to most other populations due to ZAA being one of the main source populations for other populations nation-wide, and also due to the NT/WA and QLD population amalgamation in 2016 (Figure 5.3a). The Arid Recovery, Thistle Island and Venus Bay populations were all founded by individuals from the Monarto Zoo (ZAA population), so are genetically similar to one-another (Figure 5.3a). The Mt Gibson, Mallee Cliffs, Pilliga and Currawinya sites are similar to their respective source sites (Table 5.1, Figure 5.3a). The wild Pilbara (WA) samples show lower within population genetic diversity than other sites and are most genetically similar to the ZAA population (Table 5.6, Figure 5.3a). When including functional SNPs only, the patterns of differentiation between the populations in the PCoA remained similar but was less well defined due to the much lower number of SNPs (140 vs 11,750) (Figure 5.3b).

The Scotia, Venus Bay and Yookamurra breeding enclosure (Yooka1) populations were the most divergent based on F_{ST} (Table 5.6), in consonance with the PCoA (Figure 5.3a). Conversely, the Pilliga, Mallee Cliffs and ZAA populations showed the highest genetic similarity across sites based on F_{ST}, likely due to these locations containing individuals from both QLD and NT/WA descent (Table 5.6). Relatedness between populations was relatively low, except for relatedness between the two Yookamurra sites due to Yooka1 being the individuals in the breeding enclosure that were sampled in 2001 which founded the Yooka2 population, and between Scotia and Mt Gibson due to Scotia being used as a source population for Mt Gibson (Table 5.6). Scotia and Venus Bay showed the greatest number of fixed alleles (range 0-111) between multiple populations (Table 5.7). No fixed alleles were found at nonsynonymous SNPs.

Table 5.6 Pairwise fixation indexes (F_{ST}) (grey, below diagonal, all p<0.05) representing genetic differentiation between populations and mean triadic likelihood (TrioML) relatedness estimate showing average relatedness (±SE) within populations (white, on diagonal) and between populations (white, above diagonal).

	Y1	ZAA	MG	SC	TI	VB	Y2	AR	CW	PB	PG	MC
Y1	0.787 ($6e^{-03}$)	0.014 ($3e^{-04}$)	0.116 ($2e^{-03}$)	0.018 ($4e^{-04}$)	0.005 ($7e^{-05}$)	0.004 ($6e^{-05}$)	0.411 ($6e^{-03}$)	0.015 ($2e^{-04}$)	0.007 ($2e^{-04}$)	0.003 ($2e^{-05}$)	0.015 ($2e^{-04}$)	0.009 ($2e^{-04}$)
ZAA	0.183	0.091 ($2e^{-03}$)	0.039 ($9e^{-04}$)	0.051 ($1e^{-03}$)	0.002 ($6e^{-05}$)	0.011 ($8e^{-04}$)	0.006 ($4e^{-04}$)	0.011 ($4e^{-04}$)	0.086 ($1e^{-03}$)	0.030 ($6e^{-04}$)	0.011 ($3e^{-04}$)	0.031 ($7e^{-04}$)
MG	0.198	0.069	0.285 ($3e^{-03}$)	0.343 ($5e^{-03}$)	0.002 ($5e^{-05}$)	0.002 ($2e^{-04}$)	0.119 ($4e^{-03}$)	0.009 ($2e^{-04}$)	0.062 ($7e^{-04}$)	0.001 ($3e^{-05}$)	0.055 ($1e^{-03}$)	0.080 ($1e^{-03}$)
SC	0.362	0.105	0.035	0.478 ($5e^{-03}$)	0.000 ($0e^{+00}$)	0.000 ($0e^{+00}$)	0 .000 ($0e^{+00}$)	0.000 ($0e^{+00}$)	0.082 ($1e^{-03}$)	0.001 ($5e^{-05}$)	0.074 ($2e^{-03}$)	0.104 ($2e^{-03}$)
TI	0.209	0.096	0.145	0.203	0.167 ($3e^{-03}$)	0.053 ($1e^{-03}$)	0.007 ($6e^{-04}$)	0.051 ($1e^{-03}$)	0.001 ($2e^{-05}$)	0.004 ($1e^{-04}$)	0.122 ($1e^{-03}$)	0.097 ($1e^{-03}$)
VB	0.400	0.199	0.257	0.352	0.185	0.736 ($5e^{-02}$)	0.002 ($0e^{+00}$)	0.069 ($3e^{-03}$)	0.008 ($6e^{-04}$)	0.018 ($2e^{-03}$)	0.033 ($2e^{-03}$)	0.023 ($1e^{-03}$)
Y2	0.155	0.133	0.120	0.290	0.149	0.331	0.467 ($2e^{-02}$)	0.050 ($1e^{-03}$)	0.005 ($3e^{-05}$)	0.004 ($4e^{-04}$)	0.015 ($1e^{-03}$)	0.009 ($4e^{-04}$)
AR	0.228	0.088	0.139	0.207	0.075	0.200	0.141	0.214 ($2e^{-03}$)	0.006 ($2e^{-04}$)	0.012 ($4e^{-04}$)	0.042 ($6e^{-04}$)	0.028 ($4e^{-04}$)
CW	0.205	0.014	0.068	0.088	0.121	0.210	0.147	0.112	0.126 ($2e^{-03}$)	0.025 ($6e^{-04}$)	0.017 ($4e^{-04}$)	0.043 ($7e^{-04}$)
PB	0.261	0.064	0.125	0.186	0.110	0.235	0.160	0.108	0.080	0.193 ($5e^{-03}$)	0.002 ($1e^{-04}$)	0.005 ($2e^{-04}$)
PG	0.177	0.061	0.088	0.131	0.014	0.175	0.119	0.063	0.080	0.083	0.115 ($1e^{-03}$)	0.098 ($1e^{-03}$)
MC	0.168	0.034	0.058	0.087	0.030	0.169	0.109	0.065	0.045	0.065	0.005	0.092 ($9e^{-04}$)

Y1 = Yooka1, ZAA = Zoo and Aquarium Association Australasia, MG = Mt Gibson, SC = Scotia, TI = Thistle Island, VB = Venus Bay, Y2 = Yooka2, AR = Arid Recovery, CW = Currawinya, PB = Pilbara, PG = Pilliga, MC = Mallee Cliffs

Table 5.7 Identification of fixed alleles (across 11,750 SNPs) between pairwise populations.

Population 1	Population 2	No. of Fixed Alleles
Arid Recovery	Pilbara	1
Arid Recovery	Scotia	37
Arid Recovery	Venus Bay	3
Arid Recovery	Yooka1	7
Arid Recovery	Yooka2	4
Currawinya	Scotia	3
Currawinya	Venus Bay	7
Currawinya	Yooka1	3
Currawinya	Yooka2	6
Mallee Cliffs	Venus Bay	2
Mallee Cliffs	Yooka2	1
Mt Gibson	Pilbara	1
Mt Gibson	Scotia	2
Mt Gibson	Thistle Island	1
Mt Gibson	Venus Bay	17
Pilbara	Scotia	44
Pilbara	Thistle Island	1
Pilbara	Venus Bay	16
Pilbara	Yooka1	13
Pilbara	Yooka2	16
Pilbara	ZAA	1
Pilliga	Venus Bay	1
Pilliga	Yooka1	1
Pilliga	Yooka2	2
Scotia	Thistle Island	39
Scotia	Venus Bay	111
Scotia	Yooka1	71
Scotia	Yooka2	100
Thistle Island	Venus Bay	6
Thistle Island	Yooka1	4
Thistle Island	Yooka2	6
Venus Bay	Yooka1	79
Venus Bay	Yooka2	69
Venus Bay	ZAA	6
Yooka1	Yooka2	6
Yooka1	ZAA	3
Yooka2	ZAA	3

MLH estimates based on RRS data significantly predicted WGR MLH (N = 12, F = 42.51, p < 0.001, R^2 = 0.79, β = 1.11 ± 0.17, intercept = 0.03 ± 0.03; Figure 5.4a). Additionally, there was a significant relationship between MLH based on nonsynonymous variants only (putatively functional diversity) and MLH based on all variants (genome-wide diversity) using the WGR data (N = 12, F = 191.1, p < 0.001, R^2 = 0.95, β = 1.08 ± 0.08, intercept = -0.01 ± 0.02; Figure 5.4b).

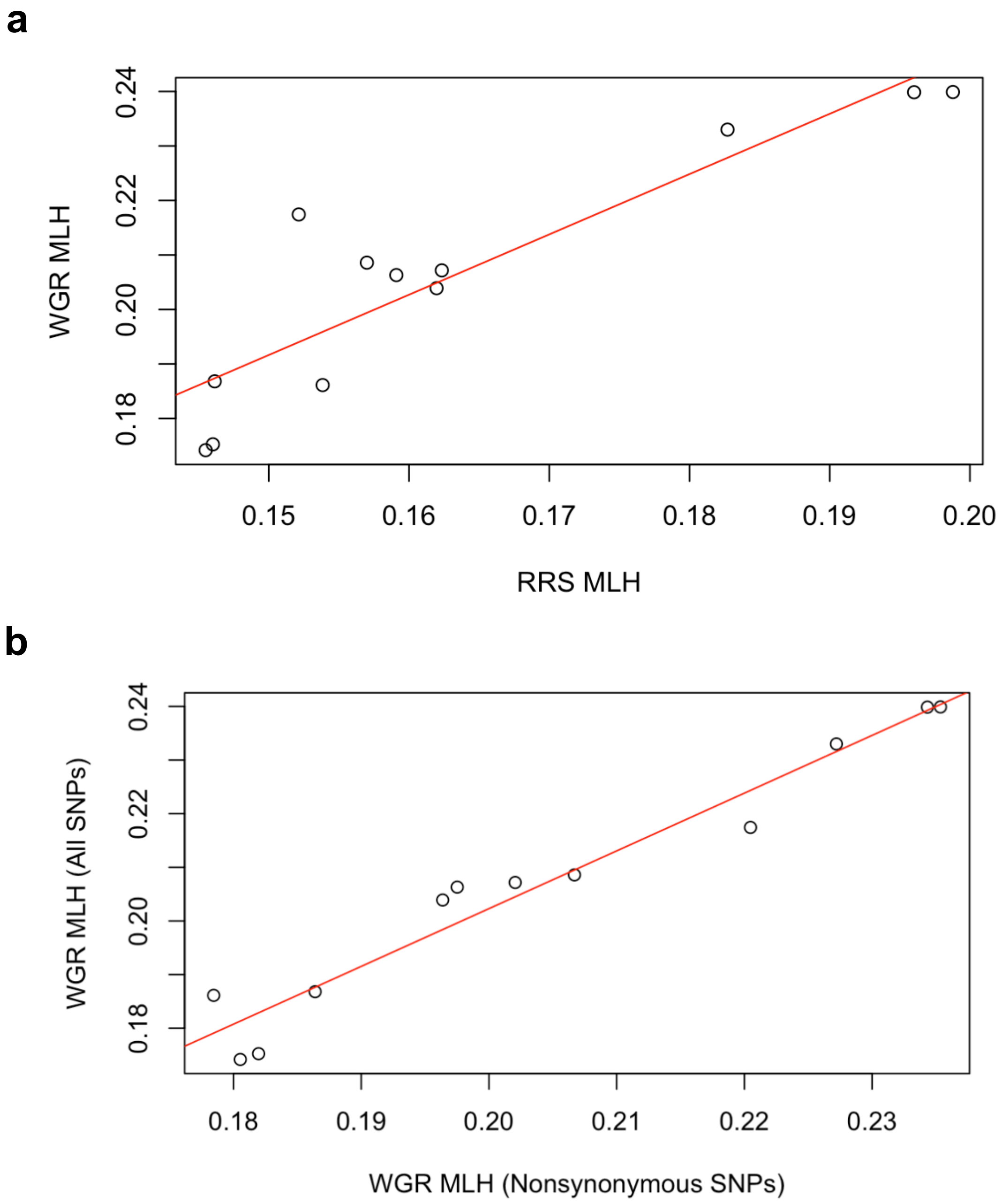


Figure 5.4 Linear models showing the relationship between a) multilocus heterozygosity (MLH) estimates from RRS (11,750 SNPs) vs WGR (45,103,325 variants) data and b) MHL estimates at nonsynonymous variants (i.e. functional diversity; 92,375 variants) vs all variants (i.e. genome-wide diversity; 45,103,325 variants) from WGR data across 12 individuals. Trend Lines are plotted in red.

Discussion

Crucial to the survival of many species living in a fragmented landscape is an effective genetic management strategy. Having methods and tools to assess and maintain genetic diversity at functional regions of the genome are needed to ensure species maintain long-term adaptive potential. Without these tools, conservation practitioners are constrained in their decision making when managing populations. Here we have generated the first annotated reference genome for the greater bilby and demonstrated a suite of downstream applications which can be used for the management of the national metapopulation. Using our WGR data we demonstrate that maximising genetic diversity across the metapopulation using RRS data should in turn maximise genome-wide (including putatively functional) diversity for the species. We also characterised genetic diversity at all contemporary managed fenced and captive populations and compared this to samples from monitored wild populations in the Pilbara region of north-western Australia. Finally, we demonstrated that reliable RRS data can be obtained from wild scat samples and developed a suite of sex-linked markers to aid in the sex identification of wild samples. The tools developed here will be vital for monitoring both wild bilby populations as well as those housed behind fences. This study provides resources for continued genetic monitoring and population management of the greater bilby metapopulation in line with the National Recovery Plan.

Reference genomes are a key starting point to aid in the development of downstream applications that can assist in the conservation of threatened species (Brandies et al., 2019). Here we assembled a high-quality reference genome for the greater bilby using one of the latest sequencing technologies, PacBio's HiFi long reads. The reference genome was 3.69 Gb in size and showed high contiguity and completeness with 28,488 protein-coding genes annotated based on 12 tissue transcriptomes. Conserving functional genetic diversity, and consequently conserving adaptative potential, is predicted to give populations the best chance of survival (Forsman & Wennersten, 2016; Hoelzel, Bruford & Fleischer, 2019). An annotated reference genome is crucial in understanding what gene regions may have functional consequences on the species and hence may hold adaptive potential (Hoelzel, Bruford & Fleischer, 2019). Many conservation management strategies rely on a combination of measures such as population differentiation (F_{ST}), inbreeding (F_{IS}), heterozygosity and estimates of relatedness in order to maintain levels of genetic diversity and reduce inbreeding within populations (Frankham, Ballou & Briscoe, 2010; Frankham et al., 2017). Until 2014, the genetic management of captive bilby populations relied on traditional pedigree-based analysis (zoos) and the wild populations relied on microsatellite analysis, which may not provide precise measures of genetic diversity due to the small number of loci (McLennan et al., 2019). More recent employment of RRS approaches for bilby population management (Lott et al., 2020) have allowed for more accurate genome-wide diversity measures (McLennan et al., 2019). Though without an annotated reference genome, it was still unknown whether the management strategy of maximising overall genetic diversity based on RRS data would in turn maximise diversity at functional regions within the bilby metapopulation. Our study employed the use of WGR of 12 samples that had previously undergone RRS sequencing and used the newly created reference genome to annotate variants and assess functional diversity in the greater bilby. Our results show that estimates of genome-wide diversity (based on multi-locus heterozygosity at all SNPs) from RRS data were strongly correlated with actual genome-wide diversity based on WGR data, a similar pattern to that shown previously in the Tasmanian devil (Wright et al., 2020). Additionally, the WGR data showed a significant relationship between functional diversity (based on multi-locus heterozygosity at nonsynonymous SNPs) and genome-wide diversity. These results together demonstrate that maximising genetic diversity based on RRS data should result in the maximisation of overall genome-wide diversity (including putatively functional diversity) in greater bilby populations. Future work should investigate the fitness implications of increased functional diversity of populations with improved genetic diversity (e.g., due to admixture of genetically dissimilar source populations) compared with control populations (e.g., individuals from source populations) to better understand how functional variation may contribute to population viability.

In order to effectively manage the bilby metapopulation nation-wide and meet the genetic goals of the National Recovery Plan (Commonwealth of Australia, 2019; Pavey, 2006), it is essential that genetic data is collected and analysed frequently (every 5 years or ~10 generation intervals is recommended for bilby populations; Lott et al., 2020). However, when loss of diversity is a concern, translocations from genetically diverse populations should be implemented every 1-2 generations (Lott et al., 2020). Previous studies have shown that the greater bilby is an ideal candidate for genetic rescue due to strong evidence of recent gene flow between populations (Moritz et al., 1997), and any divergence among bilby populations is likely being driven by random genetic drift rather than local adaptation (Lott et al., 2020; Weeks, Stoklosa & Hoffmann, 2016). The results from our study provide further evidence for these previous findings as patterns of population stratification based on functional diversity were found to be consistent with genome-wide diversity, and none of the fixed alleles identified between populations were found to occur in functional regions. Together, these results suggest that genetic differences among populations are likely due to random genetic drift rather than local adaptation due to natural selection at functional loci. To maximise genetic diversity across the metapopulation and meet the goals of the National Bilby Recovery Plan (Commonwealth of Australia, 2019; Pavey, 2006), it is crucial that the population genetic diversity results presented in the current study are interpreted collectively. For example, when deciding where to source individuals for future translocations it is important that efforts are made to source individuals from the least similar population (as inferred from both the F_{ST} and between population relatedness values) to the receiving population to maximise genetic diversity and adaptive potential across the metapopulation (Weeks, Stoklosa & Hoffmann, 2016). Mixing populations that are genetically different and where one or both populations show high within-population relatedness can reduce the risk of inbreeding depression and improve the fitness of future generations through genetic rescue (Frankham, 2015; Frankham, 2016; Hoffmann, Miller & Weeks, 2021). However, mixing populations that are genetically similar (i.e., show low F_{ST} values and high between population relatedness) will provide limited improvements to genetic diversity so should be avoided where possible, particularly when standing genetic variation of the populations is already low (Hoffmann, Miller & Weeks, 2021). It is important to note that sample size can affect the accuracy of estimating such population genetic statistics, so results from populations with low sample sizes (e.g., less than six individuals) should be treated with caution as they may not provide an accurate representation of the whole population (Li et al., 2020). While genomic data can be used to inform population management strategies for species recovery, it is also important to consider non-genetic factors that may affect population management decisions. For instance, it is important to evaluate whether the desired source populations are large enough to provide individuals for translocations or assess whether there are any other demographic or logistical barriers that may impact translocation success (e.g., the age structure of individuals within the population or the feasibility of transporting individuals long distances). The population genetic diversity statistics employed here have previously been utilised to inform translocation recommendations for the establishment of the more recent bilby populations at Mt Gibson, Pilliga and Mallee Cliffs. Future sampling of the offspring born at these recently established sites will determine whether admixture of individuals from genetically divergent source populations will improve genetic diversity across the bilby metapopulation.

Our study is the first to perform RRS on samples from wild bilby populations. Using the RRS data, we found that individuals from the wild Pilbara region exhibit low levels of genetic diversity. These results are likely due to the reduced range and low density of wild bilby populations in this region, with low relatedness between individuals due to the isolated nature of these populations across a wide geographic area (Dziminski, Carpenter & Morris, 2020a; Dziminski, Carpenter & Morris, 2020b). Despite the small sample size of wild individuals (due to the difficulty in obtaining tissue samples), the diversity estimates from the current study are concerning and provide further evidence in line with a previous study that used microsatellite analysis of 800 wild scat samples along with spatially explicit capture-recapture data to show that wild bilby populations within the Pilbara are small, isolated and likely vulnerable to threats of extinction (Dziminski, Carpenter & Morris, 2020a). Future reintroductions and translocations of individuals from the managed bilby metapopulation to these wild sites based on genetic data may improve the long-term viability of these populations (Dziminski, Carpenter & Morris, 2020a). The current study also demonstrated that reliable RRS data can be obtained from scat samples, providing a tool for fast and effective genetic monitoring of these wild populations into the future. We note that the genotyping rate was significantly reduced when working with scat samples versus tissue samples, however a genotyping rate of 53.5% across 11,750 high-quality SNPs, will still provide sufficient data to allow for non-invasive genetic monitoring of the wild populations (Schultz et al., 2018). Additionally, we have developed a panel of six Y-linked markers which can be successfully employed to assign sex of wild samples from either tissue samples or non-invasive scat samples, facilitating more comprehensive monitoring of wild populations.

Overall, the results from this study will be used as a foundation for continued genetic monitoring and management of the national bilby metapopulation. We have shown that management strategies aiming to maximise genetic diversity across the national metapopulation based on RRS data should result in maximal functional diversity of populations, in turn conserving adaptive potential and fostering conservation success of the bilby metapopulation. Our study provides genetic diversity measures for all contemporary captive bilby populations which will act as a baseline for continued monitoring and genetic management of the metapopulation. Additionally, we have demonstrated that sufficient, reliable RRS data can be obtained from scat samples, and have developed a suite of sex-linked markers, allowing more comprehensive monitoring of wild bilby populations. The availability of a bilby reference genome will not only provide benefits for continued data analysis using reference aligned RRS approaches (Brandies et al., 2019), but also provide a key resource for future research to explore crucial questions such as whether functional diversity may be associated with important phenotypic traits that may aid in species conservation (Wright et al., 2015; Wright et al., 2020). Future research should continue to assess genetic diversity across the bilby metapopulation and from the remaining wild populations regularly in order to make informed conservation management decisions and meet the genetic goals of the National Recovery Plan (Commonwealth of Australia, 2019; Pavey, 2006), giving the greater bilby the best chance of long-term survival.

Acknowledgements

There are a large number of people who helped make this work possible. Thanks to Perth Zoo for the provision of samples for the reference genomes, and the Australian Wildlife Conservancy, Arid Recovery, the Department of Biodiversity, Conservation and Attractions (Government of Western Australia), the Department for Environment and Water (Government of South Australia), the Save the Bilby Fund and the ZAA member institutions holding bilbies for their assistance with sample collection from the bilby metapopulation. We also acknowledge the technical support and access to infrastructure of Sydney Informatics Hub Core Research Facility, the University of Sydney and the Australian Biocommons which is enabled by the Australian Government National Collaborative Research Infrastructure Strategy (NCRIS). WGR computation was performed on NCI Gadi with support through the National Computational Merit Allocation Scheme 2020. The authors wish to thank NCI support staff, in particular Dr Andrey Bliznyuk and Dr Ben Menadue for their assistance with scaling the WGR workflow. Cloud support was provided by Microsoft Azure and

Pawsey Supercomputing Centre for genome assembly, annotation and RRS data analysis. The 10x Genomics sequencing was provided as part of the Oz Mammals Genomics (OMG) initiative through funding support from Bioplatforms Australia and NCRIS. This work was supported by funding from the Australian Wildlife Conservancy, the Australian Research Council (DP180102465 to KB), Toledo Zoo and Aquarium, and the University of Sydney.

References

Abbott, I 2001, 'The bilby *Macrotis lagotis* (Marsupialia: Peramelidae) in south-western Australia: original range limits, subsequent decline, and presumed regional extinction', *Records-Western Australian Museum,* vol. 20, no. 3, pp. 271-306.

Andrews, S 2014, *FastQC A Quality Control tool for High Throughput Sequence Data*, viewed 21 September 2020, http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Ballou, JD & Lacy, RC 1995, 'Identifying genetically important individuals for management of genetic variation in pedigreed populations', in Ballou, J, Gilpin, M & Foose, T (eds.), *Population Management for Survival and Recovery: Analytical Methods and Strategies in Small Population Conservation*, Columbia University Press, New York, USA.

Blouin, M, Parsons, M, Lacaille, V & Lotz, S 1996, 'Use of microsatellite loci to classify individuals by relatedness', *Molecular Ecology,* vol. 5, no. 3, pp. 393-401.

Bolger, AM, Lohse, M & Usadel, B 2014, 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics,* vol. 30, no. 15, pp. 2114-2120.

Bradley, K, Lees, C, Lundie-Jenkins, G, Copley, P, Paltridge, R, Dziminski, M, Southgate, R, Nally, S & Kemp, L 2015, 'Greater bilby conservation summit and interim conservation plan: an initiative of the Save the Bilby Fund', IUCN SSC Conservation Breeding Specialist Group, Apple Valley, MN.

Brandies, PA, Peel, E, Hogg, CJ & Belov, K 2019, 'The value of reference genomes in the conservation of threatened species', *Genes,* vol. 10, no. 11, pp. 846.

Burbidge, A & Woinarski, J 2016, '*Macrotis lagotis*', *The IUCN Red List of Threatened Species*, pp. 2016-2.

Burbidge, AA, Johnson, KA, Fuller, PJ & Southgate, R 1988, 'Aboriginal knowledge of the mammals of the central deserts of Australia', *Wildlife Research,* vol. 15, no. 1, pp. 9-39.

Bushnell, B 2014, *BBTools*, viewed 23 August 2020, https://sourceforge.net/projects/bbmap/

Catchen, J, Hohenlohe, PA, Bassham, S, Amores, A & Cresko, WA 2013, 'Stacks: an analysis tool set for population genomics', *Molecular Ecology,* vol. 22, no. 11, pp. 3124-3140.

Catchen, JM, Amores, A, Hohenlohe, P, Cresko, W & Postlethwait, JH 2011, 'Stacks: building and genotyping loci de novo from short-read sequences', *G3: Genes, Genomes, Genetics,* vol. 1, no. 3, pp. 171-182.

Chen, S, Zhou, Y, Chen, Y & Gu, J 2018, 'fastp: an ultra-fast all-in-one FASTQ preprocessor', *Bioinformatics,* vol. 34, no. 17, pp. i884-i890.

Christie, PM 1991, 'The Australasian species management plan for zoo populations of the greater bilby, *Macrotis lagotis*', Species Management Co-ordinating Council (SMCC), Dubbo, Australia.

Commonwealth of Australia 2019, 'Recovery Plan for the Greater Bilby-DRAFT'.

Cortez, D, Marin, R, Toledo-Flores, D, Froidevaux, L, Liechti, A, Waters, PD, Grützner, F & Kaessmann, H 2014, 'Origins and functional evolution of Y chromosomes across mammals', *Nature,* vol. 508, no. 7497, pp. 488-493.

Dawson, SJ, Broussard, L, Adams, PJ, Moseby, KE, Waddington, KI, Kobryn, HT, Bateman, PW & Fleming, PA 2019, 'An outback oasis: the ecological importance of bilby burrows', *Journal of Zoology,* vol. 308, no. 3, pp. 149-163.

De Meeûs, T 2018, 'Revisiting F_{IS}, F_{ST}, Wahlund effects, and null alleles', *Journal of Heredity,* vol. 109, no. 4, pp. 446-456.

Depristo, MA, Banks, E, Poplin, R, Garimella, KV, Maguire, JR, Hartl, C, Philippakis, AA, Del Angel, G, Rivas, MA & Hanna, M 2011, 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics,* vol. 43, no. 5, pp. 491-498.

Doherty, TS, Davis, NE, Dickman, CR, Forsyth, DM, Letnic, M, Nimmo, DG, Palmer, R, Ritchie, EG, Benshemesh, J & Edwards, G 2019, 'Continental patterns in the diet of a top predator: Australia's dingo', *Mammal Review,* vol. 49, no. 1, pp. 31-44.

Dziminski, MA, Carpenter, FM & Morris, F 2020a, 'Monitoring the abundance of wild and reintroduced bilby populations', *The Journal of Wildlife Management*, pp. 1-14.

Dziminski, MA, Carpenter, FM & Morris, F 2020b, 'Range of the greater bilby (*Macrotis lagotis*) in the Pilbara Region, Western Australia', *Journal of the Royal Society of Western Australia,* vol. 103, pp. 97-102.

English, AC, Richards, S, Han, Y, Wang, M, Vee, V, Qu, J, Qin, X, Muzny, DM, Reid, JG & Worley, KC 2012, 'Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology', *PloS One,* vol. 7, no. 11, pp. e47768.

Faust, GG & Hall, IM 2014, 'SAMBLASTER: fast duplicate marking and structural variant read extraction', *Bioinformatics,* vol. 30, no. 17, pp. 2503-2505.

Forsman, A & Wennersten, L 2016, 'Inter-individual variation promotes ecological success of populations and species: Evidence from experimental and comparative studies', *Ecography,* vol. 39, no. 7, pp. 630-648.

Frankham, R 2015, 'Genetic rescue of small inbred populations: Meta-analysis reveals large and consistent benefits of gene flow', *Molecular Ecology,* vol. 24, no. 11, pp. 2610-2618.

Frankham, R 2016, 'Genetic rescue benefits persist to at least the F3 generation, based on a meta-analysis', *Biological Conservation,* vol. 195, pp. 33-36.

Frankham, R, Ballou, JD & Briscoe, DA 2010, *Introduction to Conservation Genetics*, 2 edn, Cambridge University Press, Cambridge, UK.

Frankham, R, Ballou, JD, Ralls, K, Eldridge, M, Dudash, MR, Fenster, CB, Lacy, RC & Sunnucks, P 2017, *Genetic Management of Fragmented Animal and Plant Populations*, Oxford University Press, New York, USA.

Friend, J 1990, 'Status of bandicoots in Western Australia', in Seebeck, JH, Brown, PR, Wallis, RL & Kemper, CM (eds.), *Bandicoots and Bilbies*, Surrey Beatty & Sons in association with Australian Mammal Society, Chipping Norton, N.S.W.

Gordon, G, Hall, L & Atherton, R 1990, 'Status of bandicoots in Queensland', in Seebeck, JH, Brown, PR, Wallis, RL & Kemper, CM (eds.), *Bandicoots and Bilbies*, Surrey Beatty & Sons in association with Australian Mammal Society, Chipping Norton, N.S.W.

Goudet, J 2005, 'Hierfstat, a package for R to compute and test hierarchical F-statistics', *Molecular Ecology Notes,* vol. 5, no. 1, pp. 184-186.

Gruber, B, Unmack, PJ, Berry, OF & Georges, A 2018, 'dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing', *Molecular Ecology Resources,* vol. 18, no. 3, pp. 691-699.

Grueber, CE, Chong, R, Gooley, RM, McLennan, EA, Barrs, VR, Belov, K & Hogg, CJ 2020, 'Genetic analysis of scat samples to inform conservation of the Tasmanian devil', *Australian Zoologist,* vol. 40, no. 3, pp. 492-504.

Guan, D, McCarthy, SA, Wood, J, Howe, K, Wang, Y & Durbin, R 2020, 'Identifying and removing haplotypic duplication in primary genome assemblies', *Bioinformatics,* vol. 36, no. 9, pp. 2896-2898.

Hoelzel, AR, Bruford, MW & Fleischer, RC 2019, 'Conservation of adaptive potential and functional diversity', *Conservation Genetics,* vol. 20, pp. 1-5.

Hoffmann, AA, Miller, AD & Weeks, AR 2021, 'Genetic mixing for population management: From genetic rescue to provenancing', *Evolutionary Applications,* vol. 14, no. 3, pp. 634-652.

Hofstede, L & Dziminski, MA 2017, 'Greater bilby burrows: important structures for a range of species in an arid environment', *Australian Mammalogy,* vol. 39, no. 2, pp. 227-237.

Hogg, CJ, Wright, B, Morris, KM, Lee, AV, Ivy, JA, Grueber, CE & Belov, K 2019, 'Founder relationships and conservation management: empirical kinships reveal the effect on breeding programmes when founders are assumed to be unrelated', *Animal Conservation,* vol. 22, no. 4, pp. 348-361.

Jaccoud, D, Peng, K, Feinstein, D & Kilian, A 2001, 'Diversity arrays: a solid state technology for sequence information independent genotyping', *Nucleic Acids Research,* vol. 29, no. 4, pp. e25-e25.

Janečka, JE, Jackson, R, Yuquang, Z, Diqiang, L, Munkhtsog, B, Buckley-Beason, V & Murphy, W 2008, 'Population monitoring of snow leopards using noninvasive collection of scat samples: a pilot study', *Animal Conservation,* vol. 11, no. 5, pp. 401-411.

Jodi Buchecker & Vere Nicolson 2016, *Annual Report and Recommendations - Greater Bilby (Macrotis lagotis) Conservation Program*, Zoo and Aquarium Association, Sydney, Australia.

Johnson, K 2002, 'Subfamily *Thylacomyinae*: bilbies', in Van Dyck, S & Strahan, R (eds.), *The Mammals of Australia,* 2nd edn, Reed New Holland, Sydney, Australia.

Johnson, K & Southgate, R 1990, 'Present and former status of bandicoots in the Northern Territory', in Seebeck, JH, Brown, PR, Wallis, RL & Kemper, CM (eds.), *Bandicoots and Bilbies*, Surrey Beatty & Sons in association with Australian Mammal Society, Chipping Norton, N.S.W, Australia.

Jombart, T 2008, 'adegenet: a R package for the multivariate analysis of genetic markers', *Bioinformatics,* vol. 24, no. 11, pp. 1403-1405.

Jombart, T & Ahmed, I 2011, 'adegenet 1.3-1: new tools for the analysis of genome-wide SNP data', *Bioinformatics,* vol. 27, no. 21, pp. 3070-3071.

Jones, CG, Lawton, JH & Shachak, M 1994, 'Organisms as ecosystem engineers', in Samson, FB & Knopf, FL (eds.), *Ecosystem Management*, Springer, New York, USA.

Kang, Y-J, Yang, D-C, Kong, L, Hou, M, Meng, Y-Q, Wei, L & Gao, G 2017, 'CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features', *Nucleic Acids Research,* vol. 45, no. W1, pp. W12-W16.

Keenan, K 2017, *DiveRsity: A comprehensive, general purpose population genetics analysis package*, viewed 16 March 2021, https://github.com/kkeenan02/diveRsity

Kennedy, M 1992, *Australasian marsupials and monotremes: an action plan for their conservation*, World Conservation Union, Gland, Switzerland.

Kersey, DC & Dehnhard, M 2014, 'The use of noninvasive and minimally invasive methods in endocrinology for threatened mammalian species conservation', *General and Comparative Endocrinology,* vol. 203, pp. 296-306.

Kim, D, Paggi, JM, Park, C, Bennett, C & Salzberg, SL 2019, 'Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype', *Nature Biotechnology,* vol. 37, no. 8, pp. 907-915.

Kuo, RI, Cheng, Y, Zhang, R, Brown, JWS, Smith, J, Archibald, AL & Burt, DW 2020, 'Illuminating the dark side of the human transcriptome with long read transcript sequencing', *BMC Genomics,* vol. 21, no. 751, pp. 1-22.

Li, C, Weeks, D & Chakravarti, A 1993, 'Similarity of DNA fingerprints due to chance and relatedness', *Human Heredity,* vol. 43, no. 1, pp. 45-52.

Li, H & Durbin, R 2009, 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics,* vol. 25, no. 14, pp. 1754-1760.

Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G & Durbin, R 2009, 'The sequence alignment/map format and SAMtools', *Bioinformatics,* vol. 25, no. 16, pp. 2078-2079.

Li, H, Qu, W, Obrycki, JJ, Meng, L, Zhou, X, Chu, D & Li, B 2020, 'Optimizing sample size for population genomic study in a global invasive lady beetle, *Harmonia Axyridis*', *Insects,* vol. 11, no. 5, pp. 290.

Lott, MJ, Wright, BR, Kemp, LF, Johnson, RN & Hogg, CJ 2020, 'Genetic management of captive and reintroduced bilby populations', *The Journal of Wildlife Management,* vol. 84, no. 1, pp. 20-32.

Lynch, M & Ritland, K 1999, 'Estimation of pairwise relatedness with molecular markers', *Genetics,* vol. 152, no. 4, pp. 1753-1766.

Manning, AD, Eldridge, DJ & Jones, CG 2015, 'Policy implications of ecosystem engineering for multiple ecosystem benefits', in Armstrong, DP, Hayward, MW & Seddon, PJ (eds.), *Advances in Reintroduction Biology of Australian and New Zealand Fauna*, CSIRO Publishing, Clayton, Australia.

McLennan, E, Grueber, CE, Wise, P, Belov, K & Hogg, CJ 2020, 'Mixing genetically differentiated populations successfully boosts diversity of an endangered carnivore', *Animal Conservation,* vol. 23, no. 6, pp. 700-712.

McLennan, EA, Wright, BR, Belov, K, Hogg, CJ & Grueber, CE 2019, 'Too much of a good thing? Finding the most informative genetic data set to answer conservation questions', *Molecular Ecology Resources,* vol. 19, no. 3, pp. 659-671.

Miller, EJ, Eldridge, MDB, Morris, K, Thomas, N & Herbert, CA 2015, 'Captive management and the maintenance of genetic diversity in a vulnerable marsupial, the greater bilby', *Australian Mammalogy,* vol. 37, no. 2, pp. 170-181.

Milligan, BG 2003, 'Maximum-likelihood estimation of relatedness', *Genetics,* vol. 163, no. 3, pp. 1153-1167.

Moritz, C, Heideman, A, Geffen, E & McRae, P 1997, 'Genetic population structure of the Greater Bilby *Macrotis lagotis*, a marsupial in decline', *Molecular Ecology,* vol. 6, no. 10, pp. 925-936.

O'Leary, NA, Wright, MW, Brister, JR, Ciufo, S, Haddad, D, McVeigh, R, Rajput, B, Robbertse, B, Smith-White, B & Ako-Adjei, D 2015, 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research,* vol. 44, no. D1, pp. D733-D745.

Offerman, JD & Rychlik, W 2003, 'Oligo primer analysis software', in Krawetz, SA & Womble, DD (eds.), *Introduction to Bioinformatics: A Theoretical and Practical Approach*, Humana Press, New York, USA.

Pavey, C 2006, *National recovery plan for the greater bilby*, Northern Territory Department of Natural Resources, Environment and the Arts, Alice Springs, Australia.

Pembleton, LW, Cogan, NO & Forster, JW 2013, 'St AMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations', *Molecular Ecology Resources,* vol. 13, no. 5, pp. 946-952.

Pertea, M, Pertea, GM, Antonescu, CM, Chang, T-C, Mendell, JT & Salzberg, SL 2015, 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nature Biotechnology,* vol. 33, no. 3, pp. 290-295.

Peterson, BK, Weber, JN, Kay, EH, Fisher, HS & Hoekstra, HE 2012, 'Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species', *PloS One,* vol. 7, no. 5, pp. e37135.

Pew, J, Wang, J, Muir, P & Frasier, T 2015, *related: an R package for analyzing pairwise relatedness data based on codominant molecular markers*, viewed 7 January 2021, https://r-forge.r-project.org/projects/related/

Poplin, R, Ruano-Rubio, V, Depristo, MA, Fennell, TJ, Carneiro, MO, Van Der Auwera, GA, Kling, DE, Gauthier, LD, Levy-Moonshine, A & Roazen, D 2017, 'Scaling accurate genetic variant discovery to tens of thousands of samples', *BioRxiv*, pp. 201178.

Queller, DC & Goodnight, KF 1989, 'Estimating relatedness using genetic markers', *Evolution,* vol. 43, no. 2, pp. 258-275.

Quinlan, AR & Hall, IM 2010, 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics,* vol. 26, no. 6, pp. 841-842.

Read, J, Carter, J, Moseby, K & Greenville, A 2008, 'Ecological roles of rabbit, bettong and bilby warrens in arid Australia', *Journal of Arid Environments,* vol. 72, no. 11, pp. 2124-2130.

Ritland, K 1996, 'Estimators for pairwise relatedness and individual inbreeding coefficients', *Genetics Research,* vol. 67, no. 2, pp. 175-185.

Salamov, AA & Solovyev, VV 2000, 'Ab initio gene finding in *Drosophila* genomic DNA', *Genome Research,* vol. 10, no. 4, pp. 516-522.

Schultz, AJ, Cristescu, RH, Littleford-Colquhoun, BL, Jaccoud, D & Frère, CH 2018, 'Fresh is best: Accurate SNP genotyping from koala scats', *Ecology and Evolution,* vol. 8, no. 6, pp. 3139-3151.

Smit, A, Hubley, R & Green, P 2008-2015, *RepeatModeler Open-1.0*, viewed 19 December 2019, http://www.repeatmasker.org

Smit, A, Hubley, R & Green, P 2013-2015, *RepeatMasker Open-4.0*, viewed 19 December 2019, http://www.repeatmasker.org

Smith, S, McRae, P & Hughes, J 2009, 'Faecal DNA analysis enables genetic monitoring of the species recovery program for an arid-dwelling marsupial', *Australian Journal of Zoology,* vol. 57, no. 2, pp. 139-148.

Solovyev, V, Kosarev, P, Seledsov, I & Vorobyev, D 2006, 'Automatic annotation of eukaryotic genes, pseudogenes and promoters', *Genome Biology,* vol. 7, no. S1, pp. S10.

Solovyev, VV 2002, 'Finding genes by computer: probabilistic and discriminative approaches', in Tao Jiang, YX, Michael Q. Zhang (ed.), *Current Topics in Computational Molecular Biology*, MIT Press, Cambridge, MA, USA.

Southgate, R & Adams, M 1994, 'Genetic variation in the greater bilby (Macrotis lagotis)', *Pacific Conservation Biology,* vol. 1, no. 1, pp. 46-52.

Sunnucks, P & Hales, DF 1996, 'Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Sitobion* (Hemiptera: Aphididae)', *Molecular Biology and Evolution,* vol. 13, no. 3, pp. 510-524.

Tarasov, A, Vilella, AJ, Cuppen, E, Nijman, IJ & Prins, P 2015, 'Sambamba: fast processing of NGS alignment formats', *Bioinformatics,* vol. 31, no. 12, pp. 2032-2034.

Van Der Auwera, GA, Carneiro, MO, Hartl, C, Poplin, R, Del Angel, G, Levy-Moonshine, A, Jordan, T, Shakir, K, Roazen, D & Thibault, J 2013, 'From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline', *Current Protocols in Bioinformatics,* vol. 43, no. 1, pp. 11.10. 1-11.10. 33.

Vine, S, Crowther, M, Lapidge, S, Dickman, CR, Mooney, N, Piggott, M & English, A 2009, 'Comparison of methods to detect rare and cryptic species: a case study using the red fox (*Vulpes vulpes*)', *Wildlife Research,* vol. 36, no. 5, pp. 436-446.

Waits, LP & Paetkau, D 2005, 'Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection', *The Journal of Wildlife Management,* vol. 69, no. 4, pp. 1419-1433.

Walker, BJ, Abeel, T, Shea, T, Priest, M, Abouelliel, A, Sakthikumar, S, Cuomo, CA, Zeng, Q, Wortman, J & Young, SK 2014, 'Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PloS One,* vol. 9, no. 11, pp. e112963.

Wang, J 2002, 'An estimator for pairwise relatedness using molecular markers', *Genetics,* vol. 160, no. 3, pp. 1203-1215.

Wang, J 2007, 'Triadic IBD coefficients and applications to estimating pairwise relatedness', *Genetics Research,* vol. 89, no. 3, pp. 135-153.

Wang, J 2011, 'COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients', *Molecular Ecology Resources,* vol. 11, no. 1, pp. 141-145.

Wang, K, Li, M & Hakonarson, H 2010, 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Research,* vol. 38, no. 16, pp. e164-e164.

Warren, RL, Yang, C, Vandervalk, BP, Behsaz, B, Lagman, A, Jones, SJ & Birol, I 2015, 'LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads', *GigaScience,* vol. 4, no. 1, pp. s13742-015-0076-3.

Watts, C 1969, 'Distribution and habits of the rabbit bandicoot', *Transactions of the Royal Society of South Australia,* vol. 93, pp. 135-141.

Weeks, AR, Stoklosa, J & Hoffmann, AA 2016, 'Conservation of genetic uniqueness of populations may increase extinction likelihood of endangered species: the case of Australian mammals', *Frontiers in Zoology,* vol. 13, no. 1, pp. 1-9.

Weisenfeld, NI, Kumar, V, Shah, P, Church, DM & Jaffe, DB 2017, 'Direct determination of diploid genome sequences', *Genome Research,* vol. 27, no. 5, pp. 757-767.

Wenger, AM, Peluso, P, Rowell, WJ, Chang, P-C, Hall, RJ, Concepcion, GT, Ebler, J, Fungtammasan, A, Kolesnikov, A & Olson, ND 2019, 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature Biotechnology,* vol. 37, no. 10, pp. 1155-1162.

Wright, B, Farquharson, KA, McLennan, EA, Belov, K, Hogg, CJ & Grueber, CE 2019a, 'From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species', *BMC Genomics,* vol. 20, no. 453, pp. 1-10.

Wright, B, Grueber, C, Lott, M, Belov, K, Johnson, R & Hogg, C 2019b, 'Impact of reduced-representation sequencing protocols on detecting population structure in a threatened marsupial', *Molecular Biology Reports,* vol. 46, no. 5, pp. 5575-5580.

Wright, B, Morris, K, Grueber, CE, Willet, CE, Gooley, R, Hogg, CJ, O'Meally, D, Hamede, R, Jones, M & Wade, C 2015, 'Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population', *BMC Genomics,* vol. 16, no. 791, pp. 1-11.

Wright, BR, Farquharson, KA, McLennan, EA, Belov, K, Hogg, CJ & Grueber, CE 2020, 'A demonstration of conservation genomics for threatened species management', *Molecular Ecology Resources,* vol. 00, pp. 1-16.

Yang, H & Wang, K 2015, 'Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR', *Nature Protocols,* vol. 10, no. 10, pp. 1556-1566.

Yeo, S, Coombe, L, Warren, RL, Chu, J & Birol, I 2018, 'ARCS: scaffolding genome drafts with linked reads', *Bioinformatics,* vol. 34, no. 5, pp. 725-731.

Zerbino, DR & Birney, E 2008, 'Velvet: algorithms for de novo short read assembly using de Bruijn graphs', *Genome research,* vol. 18, no. 5, pp. 821-829.

Zheng, GX, Lau, BT, Schnall-Levin, M, Jarosz, M, Bell, JM, Hindson, CM, Kyriazopoulou-Panagiotopoulou, S, Masquelier, DA, Merrill, L & Terry, JM 2016, 'Haplotyping germline and cancer genomes with high-throughput linked-read sequencing', *Nature Biotechnology,* vol. 34, no. 3, pp. 303-311.

5.3 SUPPLEMENTARY

Table S1 Summary of computational resources used for bioinformatic analyses.

Bioinformatic Pipeline	Main Software	Computing Platform (Specifications)	Walltime (hr)	Disk Space (Gb)
Genome Assembly				
Assemble PacBio HiFi Reads	IPA	Cloud Instance (64 vCPUs, 256 GB RAM)	4	100
Remove Duplicates	purge_dups	Cloud Instance (64 vCPUs, 256 GB RAM)	2	15
Convert 10x Genomics reads	longranger	Cloud Instance (64 vCPUs, 256 GB RAM)	17	2000
Scaffolding with 10x data	BWA + ARCS	Cloud Instance (64 vCPUs, 256 GB RAM)	28	750
Polishing with 10x data	Pilon	HPC (129 jobs with 2CPUs, 30 GB RAM each)	0.5	10
Transcriptome Assembly				
Align and assemble transcripts	HISAT2 + Stringtie	HPC (12 jobs with 6 CPUs, 48 GB RAM each)	4	2000
Merge transcripts	TAMA merge	HPC (1 job with 1 CPU, 10 GB RAM each)	4	10
Genome Annotation				
Build repeat database	RepeatModeler	Cloud Instance (64 vCPUs, 256 GB RAM)	23	50
Mask Repeats	RepeatMasker	Cloud Instance (64 vCPUs, 256 GB RAM)	2	15
Annotate genome	FGENESH++	Cloud Instance (64 vCPUs, 256 GB RAM)	32	50
Popgen Analysis				
WGR Alignment and SNP Calling	BWA + GATK	HPC (Max 7,200 CPUs, 30 TB RAM)	24	5000
RRS Alignment and SNP Calling	BWA + STACKS	Cloud Instance (64 vCPUs, 256 GB RAM)	10	150
Variant filtering and annotation	GATK + ANNOVAR	Cloud Instance (64 vCPUs, 256 GB RAM)	12	200

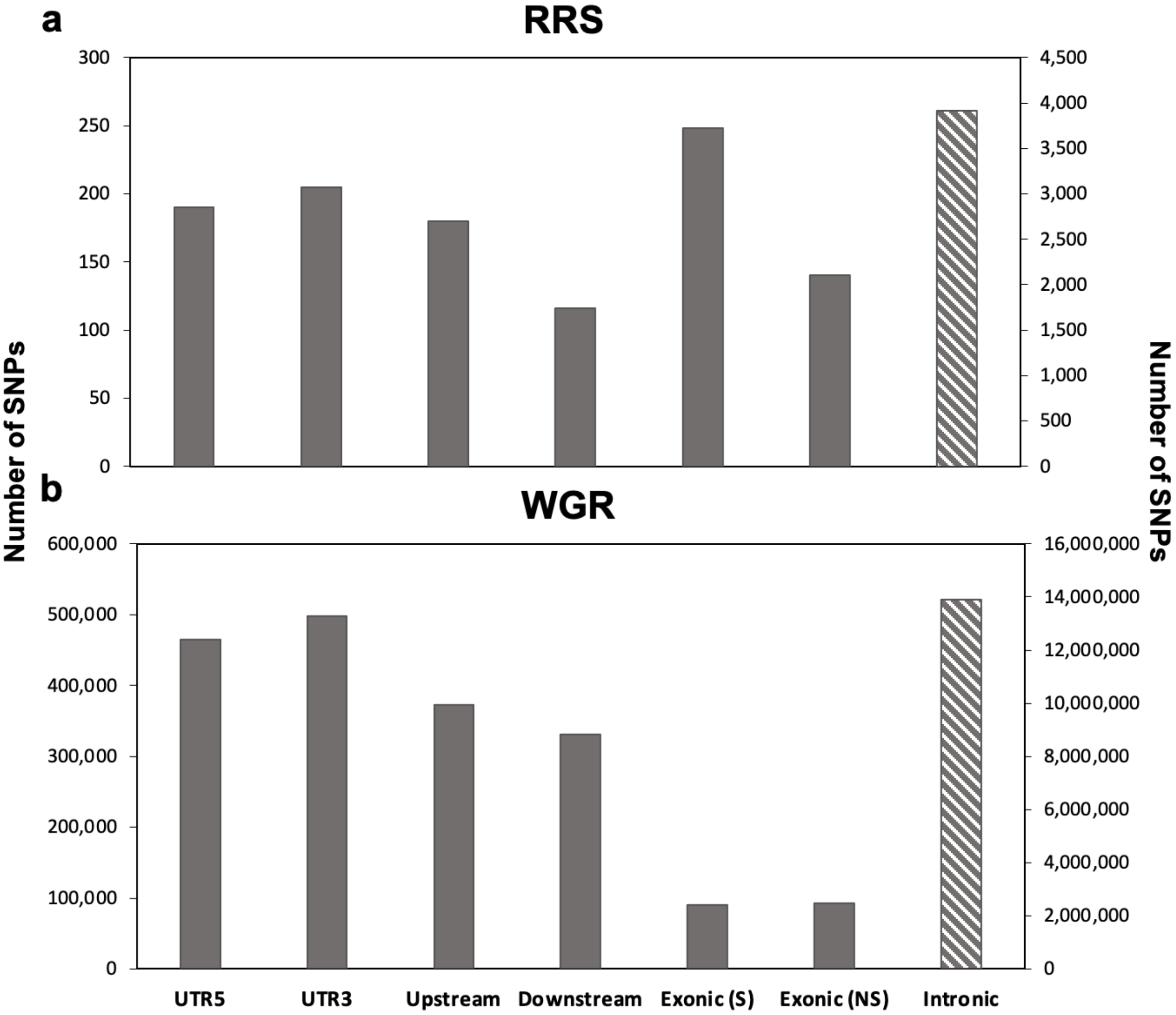


Figure S1 Number of each type of SNP annotated from a) RRS data and b) WGR data including: 3' untranslated region (UTR3), 5' untranslated region (UTR5), 1 kb downstream of gene (downstream), 1 kb upstream of gene (upstream), synonymous (S) and nonsynonymous (NS) exonic, and intronic SNPs. Intronic SNPs (striped) are plotted on the secondary Y axis.

CHAPTER 6

Final discussion, future directions and conclusions

FINAL DISCUSSION, FUTURE DIRECTIONS AND CONCLUSION

Genomic data is a valuable tool to assist in the conservation of threatened species (Chapter 1) (Fuentes-Pardo & Ruzzante, 2017; Khan et al., 2016; Larsen & Matocq, 2019; McMahon, Teeling & Höglund, 2014; Supple & Shapiro, 2018). With numbers of species at risk of extinction growing every year, successful integration of genomic data into conservation initiatives is crucial for arming populations with the best chance of long-term survival (Ballou & Lacy, 1995; Frankham, Ballou & Briscoe, 2010; Frankham et al., 2017). Conservationists are enthusiastic about the potential of genomic data as a tool for conservation (Taylor, Dussex & van Heezik, 2017). However, a lack of understanding of exactly how genomic data can assist with conservation efforts and how it can be successfully employed toward species management is a major driver of the research-implementation gap (Britt et al., 2018; Shafer et al., 2015; Taylor, Dussex & van Heezik, 2017). Additionally, the bioinformatic expertise now required to work with modern large next-generation sequencing datasets is posing another major challenge, further driving the gap between genomics and conservation (Chapter 2).

In this thesis I have taken a step-by-step approach to address these barriers and made crucial steps to bridging the research-implementation gap. With a focus on Australian marsupials, my research provides a number of key examples that demonstrate the value of genomic data in species conservation across a range of contexts and provides researchers with the bioinformatic knowledge and tools needed to generate and utilise genomic datasets for conservation. Specifically, I 1). showcased the value of reference genomes and accompanying genomic data in threatened species management using the Tasmanian devil as a model, 2). employed a range of sequencing technologies and novel bioinformatic approaches to create a variety of new genomic resources for three Australian species and demonstrated the diverse ways that modern genomic data types can be utilised to inform conservation management, and 3). provided ten simple rules to help researchers get started with applying the bioinformatic approaches used throughout this thesis to other threatened species. Below I summarise how the research presented in this thesis will facilitate others to harness the power of genomics for the conservation of threatened species and help close the research-implementation gap.

6.1 GENERATING GENOMIC RESOURCES FOR SPECIES CONSERVATION

Throughout this thesis I have demonstrated a number of ways that reference genomes and accompanying genomic datasets can be used to answer a variety of questions with implications for species management. When planning conservation genetic studies, it is important for researchers to first work with conservation teams to determine what key questions need to be addressed to assist species recovery and how genomic data may contribute to answering such questions (Hogg et al., 2017; Hohenlohe, Funk & Rajora, 2021). Once key conservation questions have been identified for the species of interest, it is important for researchers to assess both the quality and type of genomic data needed to answer such questions. A number of factors that contribute to this decision-making process include: i) the research question at hand, ii) the cost of obtaining particular data types, iii) the availability of samples and their respective quality, and iv) the downstream bioinformatic requirements. Below I discuss how this thesis has addressed these factors with respect to reference genomes and accompanying genomic datasets.

Reference Genomes

In Chapter 1, I used the Tasmanian devil as a model to demonstrate how reference genomes can be employed to answer a multitude of questions that can inform threatened species management. Some conservation questions have broad applicability to threatened species and concern the general management of captive and wild populations, such as an understanding of the genetic diversity within populations, or resolving parentage to make more informed translocation decisions and breeding recommendations (Ballou & Lacy, 1995; Ballou et al., 2010; Frankham, Ballou & Briscoe, 2010; Frankham et al., 2017). Other questions may be highly species-specific and relate to particular threatening processes such as disease (Gupta, Robin & Dharmarajan, 2020), or issues relating to adaptation to captivity (Frankham, 2008; Frankham et al., 1986). The research presented throughout this thesis allowed me to conclude that in almost all cases where genomic data is being employed, a reference genome will be needed, or at least provide several advantages for downstream tools and analysis. For example, in Chapter 1 I discussed how reference genomes can enable fast and cost-effective development of common conservation genetic tools such as microsatellites for general population monitoring and resolving parentage (e.g., Gooley et al., 2017), as well as targeted SNP panels to better understand and manage genetic variation at important gene families such as immune genes (e.g., Morris et al., 2015). Other genomic resources such as RRS data have vastly improved our capacity to manage many threatened species through measures of predominantly neutral diversity. While a reference genome is not essential for the common conservation applications of RRS data, such as general population management and parentage, Chapter 1 described several advantages for the conjunction of reference genomes with such datasets, including: more reliable genotype calls (Torkamaneh, Laroche & Belzile, 2016); lower required sequencing coverage (Davey et al., 2011); identification of more variants (Shafer et al., 2017); and the ability to annotate variants and explore their associations with important traits such as disease resistance (Margres et al., 2018). Additionally, my research in Chapter 5 showed that pairing RRS data with a reference genome and other genomic datasets can assist in answering additional conservation questions that could not be answered with RRS data alone. For example, to determine whether population management decisions based on RRS data are likely to result in the desired outcome of maximising functional diversity and hence conserving adaptive potential across populations. As sequencing costs continue to decline, the availability of WGR data for more individuals will become a reality. Pairing WGR data with a reference genome enables a multitude of additional applications for species conservation such as discovering genetic variation that may have important functional consequences on the species (Chapter 3), characterising sex chromosomes to facilitate research on reproduction (Chapter 4), developing sex-linked markers for non-invasive monitoring of populations (Chapter 5), and assessing the efficacy of other genomic data types for population management (Chapter 5) (see Accompanying Genomic Data section below). Overall, this thesis has shown that reference genomes are a valuable genomic resource for species conservation due to their versatility to aid in the development of new conservation tools and be employed in conjunction with other genomic datasets to answer a wide variety of conservation questions.

When wanting to employ a reference genome to assist with species conservation, it is important to determine whether a reference genome for the target species (or closely related counterpart) may already exist as in Chapter 3, or whether a reference genome will need to be generated as in Chapters 4 and 5. Researchers should first search through common genome repositories such as NCBI (O'Leary et al., 2015) and/or Ensembl (Howe et al., 2020), as well as search the literature, to determine whether a suitable reference genome is already available for their species of interest. While there is currently a limited number of reference genomes available for threatened species (as discussed in Chapter 1), the recent establishment of many national and international sequencing consortia such as the Earth Biogenome Project (Lewin et al., 2018), the Vertebrate Genomes Project (Rhie et al., 2021) and DNA Zoo (Aiden Lab, 2018; Dudchenko et al., 2017) are creating a whole suite of reference genomes for species around the world. Many of these projects are making the assembled reference genomes available to the public prior to publication through common genome repositories or their own independent online databases. While some genomes are placed under embargo until publication, the rapid availability of these genomes will greatly facilitate the downstream use of genomic data in conservation contexts by providing one of the major genomic resources that is often one of the most difficult to obtain (both due to the cost and expertise required to generate a reference genome). If a reference genome for a species of interest is not publicly available, connecting with sequencing consortia and other researchers is useful for determining whether a reference genome may already be planned or is currently in progress. The greater availability of reference genomes in the coming years will enable researchers to take advantage of such resources and associated genomic datasets to explore novel research questions related to species conservation (e.g. Chapter 3) (Rhie et al., 2021).

In cases where a reference genome for the specific target species is not available or in progress, it is important to consider whether a reference genome for a related species may be sufficient. Previous studies in birds have shown that while reference genomes from more closely related species (i.e., species within the same genus) produce the most accurate results, reference genomes from more distantly related species (i.e., species within the same family or even the same order) can still provide accurate diversity measures for conservation recommendations (Galla et al., 2019). Creating reference genomes for non-threatened species can therefore provide a resource to explore and monitor genetic variation in related threatened species. For example, the brown antechinus (*Antechinus stuartii*) reference genome generated in Chapter 4 is currently being employed to align RRS data and analyse population structure and conservation units across a variety of antechinus species such as *A stuartii*, *A. subtropicus*, *A. agilis* and the recently described and endangered *A. argentus* (Baker, Mutton & Hines, 2013). Being able to employ reference genomes from a related species also has advantages when sample collection from a related non-threatened species may be much easier than from the threatened target species e.g., when opportunistic sampling from a critically endangered species is highly unlikely, but sample acquisition for a less threatened congeneric species is more feasible. This method could be useful to apply to assist in the conservation of other species such as the critically endangered Gilbert's potoroo (*Potorous gilbertii*), whereby genomic samples have previously been collected for general population monitoring of the closely related but lesser threatened congener the Long-nosed potoroo (*Potorous tridactylus*) (Mulvena et al., 2020). In situations where sample acquisition for the species of interest may be difficult, it is important to determine whether a reference genome for a suitable related species already exists, or whether generating a reference genome for a related non-threatened species (rather than the species of interest) may be a more viable option for the conservation genomic questions at hand.

Finally, in situations where a reference genome is needed but a related genome does not exist or will not be suitable, researchers may need to create a reference genome for the species of interest. In Chapter 5, I demonstrated an end-to-end example of generating a high-quality reference genome for a threatened species and utilising the genome along with a variety of other genomic data types to answer a variety of questions with direct implications for species management. The approaches employed in Chapter 5 could be applied to other threatened species where limited genomic data currently exists but in-depth genetic monitoring of populations is crucial to species recovery e.g., the critically endangered woylie (*Bettongia penicillata*) (Pacioni et al., 2020) or the endangered numbat (*Myrmecobius fasciatus*) (Hayward et al., 2015). Throughout my PhD I worked on the creation of reference genomes for these two threatened species using the methodologies described in Chapters 4 and 5 of this thesis. These genomes are currently being employed to explore a range of biological questions pertaining to marsupial immune function and conservation planning (in the case of the woylie).

When generating a reference genome, it is important to determine the quality of reference genome that is needed to answer the conservation questions at hand. The desired quality of the reference genome will determine: i) what sequencing technology should be employed, ii) what sample quality and quantity is required for the respective sequencing technology, iii) what sort of bioinformatic experience, time and resources are required and iv) the overall cost to create the reference genome. Throughout this thesis I have generated and/or utilised genomes of varying quality that were created using differing sequencing technologies (Table 6.1). In Chapter 3, I used the pre-existing Tasmanian devil genome that was sequenced in 2011 using short-read technologies (Murchison et al., 2012) (Table 6.1). By today's standards, this genome would now be considered quite poor quality and yet was still sufficient to answer a wide range of conservation applications as discussed in Chapter 1.

Table 6.1 Comparison of genome sequencing technologies employed for the marsupial reference genomes used throughout this thesis

Sequencing Type	Sequencing Technology	Thesis Chapter	Est. 2021 Sequencing Cost (AUD)	Required Sample Quality	Required DNA Quantity	Reference Genome quality
Short-read	Illumina & Roche	1 & 3	$2,500	Low	Low (0.5μg)	Low
Linked-read	10x Genomics	4 & 5	$7,500	High	Low (0.5μg)	Med
Long-read	PacBio HiFi	5	$12,000*	High	High (12μg)	High

*Note: Prior to the PacBio Sequel II system and HiFi sequencing being introduced, the average cost to sequence a mammalian genome with PacBio long-reads was ~$75,000.

In Chapter 4, I used 10x Genomics linked read sequencing to generate the brown antechinus reference genome. At the time (2019) when this genome was generated, 10x Genomics linked read sequencing was one of the most popular whole genome sequencing technologies available as it provided an intermediate option between the low-quality but affordable short-read sequencing (e.g. Illumina) and the high-quality but expensive long-read sequencing options (e.g. PacBio CLR). This technology enabled the generation of reference genomes of a suitable quality to facilitate downstream applications in a cost-effective manner (Weisenfeld et al., 2017). One limitation of 10x Genomics sequencing is that the quality of the resultant reference genome was dependent on the integrity of the input DNA. This relies on obtaining high-quality tissue samples that have been preserved in a such a way (usually by flash freezing) as to prevent any DNA degradation, though the required DNA quantity was still low (Table 6.1). Another limitation is that the computational requirements for genome assembly were quite high with some mammalian-sized genomes requiring more than 512GB of RAM and up to one week of walltime with 64 CPUs.

By the time the 10x Genomics whole genome sequencing service was removed from the market in mid-2020, long-read technologies had become much more affordable with the release of PacBio's Sequel II system which enabled ~8x more sequencing data from a single sequencing cell, along with the release of HiFi reads, which enabled more accurate long-read data than ever before (Wenger et al., 2019). I took advantage of this latest technology to generate a reference genome for the greater bilby (*Macrotis lagotis*) (Chapter 5). Paired with 10x Genomics data that had already been generated in 2019, the long HiFi reads enabled the creation of a high-quality reference genome that facilitated a variety of population management analyses. This reference genome is currently being coupled with HiC sequencing (see below) and will be used to answer a whole suite of additional evolutionary and demographic questions as part of the larger bilby genome project being led by our research group. The required computational resources for assembling the HiFi reads were also relatively low (see Supplementary Table 1 of Chapter 5) when compared with the 10x Genomics assembly requirements described above. However, there are still some limitations when wanting to employ current long-read technologies, namely the high sample quality and DNA quantity input requirements (Table 6.1). These limitations are problematic when working with threatened species since most samples are collected opportunistically and hence may not have been optimally preserved. Clear communication between researchers and those who are most likely to collect samples, such as conservation managers and veterinarians, is needed to make sure best practices for sample collection and preservation are used. This is crucial in ensuring high quality samples are available when opportunistic situations arise for threatened species.

In recent years, Hi-C sequencing has also become a popular technology used in reference genome creation as the chromatin conformation capture protocol can be used to link genomic regions that are in close proximity and assist in genome scaffolding. Hi-C is therefore often employed after initial draft genome assembly to create chromosome-length reference assemblies (Burton et al., 2013). Whilst Hi-C can greatly improve the contiguity of reference assemblies which may be important for particular analyses such as assessing runs of homozygosity, exploring large structural variants, undertaking comparative evolutionary analyses, or studying chromosomal organisation, this technology does come at an additional cost (~$20,000 AUD for a mammalian sized genome as of 2021) and also requires high-quality tissue or blood samples. Researchers should liaise with conservation managers to determine what genome quality is required to answer the conservation questions at hand and also take into consideration the cost and sample input requirements when deciding what sequencing technologies are most suitable for reference genome creation.

This thesis has shown that a variety of conservation questions can be answered with reference genomes of varying quality. Lower quality reference genomes (such as the antechinus genome generated in Chapter 4, or the Tasmanian devil genome employed in Chapter 3) are often simpler, faster, and cheaper to create and are usually sufficient for questions relating to general population monitoring and management. In Chapter 1, I demonstrated that such genomes can also be employed to answer a wide range of additional species-specific conservation questions, though there can be limitations that result from the reduced sequence contiguity of lower quality genomes. For instance, only being able to obtain partial gene sequences for some of the genes targeted in Chapters 3 and 4. High quality genomes often result in better gene annotation (particularly for large diverse gene families such as immune genes) which is important when wanting to understand the relationship between genes and specific traits that may be important for species conservation such as disease resistance or susceptibility. The effectiveness of generating high-quality genomes to explore complex gene families in disease-threatened species has successfully been demonstrated in the koala which is threatened by chlamydia. The high-quality koala genome enabled comprehensive characterisation of immune gene clusters which enabled me to identify variation in immune genes that may be involved in differential immune responses to chlamydia vaccine (Johnson et al., 2018; see Appendix 2). Obtaining a high-quality reference genome may therefore be beneficial to other vulnerable species that are threatened by disease such as the southern corroboree frog (*Pseudophryne corroboree*) which is critically endangered due to the amphibian chytrid fungus *Batrachochytrium dendrobatidis* (IUCN, 2020). Previous research on the southern corroboree frog has used experimental methods to characterise a subset of immune genes (Kosch et al., 2017) and employed targeted sequencing approaches such as PCR, along with RRS techniques, to better understand how genetic variation (both genome-wide and at particular genes) is associated with resistance to chytrid fungus (Kosch et al., 2019). A high-quality reference genome for this species (currently being generated by the Vertebrate Genomes Project) will facilitate fast and effective characterisation of all immune gene families and enable the identification of additional functional variation which may be involved in chytrid fungus resistance. As sequencing technologies continue to improve, obtaining high quality reference genomes will become simpler and more affordable than ever, enabling more researchers to have access to high quality genomic resources for a wide range of conservation applications.

Accompanying Genomic Data

Once a plan has been set for the reference genome, it is important to consider what associated data type will be most suitable for answering the conservation questions at hand. Microsatellites were once the post popular genetic tool for informing general population management (Selkoe & Toonen, 2006). However, the rise of the genomics era has seen a shift towards the use of next generation sequencing technologies as these data types can provide a higher resolution of genome-wide diversity (McLennan et al., 2019; Narum et al., 2013). In Chapter 1, I introduced the three main genomic data types that are often employed to inform conservation management, namely reduced representation sequencing (RRS), whole genome resequencing (WGR), and targeted sequencing/targeted SNP panels, and described how each technology has successfully been employed to assist conservation efforts for the Tasmanian devil. Each genomic data type has its own advantages and limitations and sometimes a single data type may not be sufficient for answering all questions (Fuentes-Pardo & Ruzzante, 2017). Which technology to use is dependent on the same factors that need to be considered when creating a reference genome including sample availability, bioinformatic requirements and budget, though the most important consideration is which data type/s will be most informative for the conservation questions at hand. Below I discuss how this thesis has shown how common genomic datasets can be applied in different ways across several species to answer a variety of questions that can be used to inform conservation management.

One of the major goals included in almost every Australian species recovery plan is the maintenance of genetic diversity within and/or across populations (Frankham, Ballou & Briscoe, 2010; Frankham et al., 2017). Maintaining genetic diversity is predicted to provide populations the best chance of long-term survival by conserving functional variation (i.e., variation in genes that have functional consequences on the individual) and hence preserving the potential for populations to adapt to future change (Hoelzel, Bruford & Fleischer, 2019; Holderegger, Kamm & Gugerli, 2006). For this goal to be met, effective monitoring of populations is required. This involves obtaining genomic data across a number of individuals to identify genetic variation within a population or across multiple populations. To achieve this, RRS data is often sufficient as it provides a simple, reliable, cost effective way of collecting genome-wide data from a large number of individuals (Andrews et al., 2016; Fuentes-Pardo & Ruzzante, 2017). My research in Chapter 5 demonstrated how RRS can successfully be employed to a large number of individuals from both captive and wild populations to monitor and manage genetic diversity of a threatened species at a national scale. Furthermore, in line with previous research on the koala (Schultz et al., 2018), my research showed that reliable RRS data could be obtained from non-invasive scat samples in the bilby. This further validation shows that the RRS approach used in Chapter 5 could be applied to other species where high-quality tissue samples may be difficult to obtain but regular cost-efficient genetic data is needed for population monitoring. For example, RRS could be applied to scat samples from the critically endangered hairy-nosed wombat (*Lasiorhinus krefftii*). Current genetic monitoring of this extremely rare threatened species relies on microsatellite analysis of non-invasively collected hair samples (White et al., 2014), though a recent review has highlighted the limitations of this approach and describes the benefits for future work to employ next generation sequencing approaches for more reliable population monitoring (Martin & Carver, 2020). The methodologies used in Chapter 5 have broad applicability to other species and should be used as a model to demonstrate the potential of incorporating multiple genomic resources together to answer questions that have direct implications on species management.

More complex questions relating to diversity at particular gene families and their association with particular traits will usually require WGR or targeted sequencing (Fuentes-Pardo & Ruzzante, 2017). For example, my research presented in Chapter 3 showed how WGR can be used to identify variants in genes that may have functional consequences on a species. Specifically, I identified nonsynonymous variants that may have implications on reproduction in the Tasmanian devil and hence could be contributing to the previously identified declines in reproductive success across generations in captivity (Farquharson, Hogg & Grueber, 2017). Understanding how genetics may underpin species-specific processes that are limiting the success of conservation efforts can provide insights into possible management solutions. For example, if an association between particular alleles and reproductive success is identified, specific breeding and/or translocation recommendations can be implemented to ensure functional diversity is maximised and beneficial alleles are maintained within the population of interest.

WGR enables exploration of all loci across the genome but is usually only performed on a small number of individuals due to the cost (~$1,250 per WGR sample versus ~$53 per RRS sample in 2021), whereas targeted sequencing using either target capture approaches or targeted SNP panels enables exploration of a subset of target loci across many individuals (Jones & Good, 2016). Often large numbers of individuals are required to gain enough statistical power to investigate the relationship between genetic variation and particular species traits (Fuentes-Pardo & Ruzzante, 2017), though it is important to first identify which loci may be informative. In Chapter 3 I showed that WGR data can first be employed to identify a subset of target genes that may be informative for species conservation and targeted sequencing approaches can then be used to test the association between target genes and desired traits. The results of my chapter have formed the foundation for future research to test this hypothesis by incorporating the identified polymorphic reproductive genes into a targeted gene panel which is currently being deployed to genotype a large number of individuals across the Tasmanian devil insurance population, as well as animals in wild populations. The research presented in Chapter 3 therefore serves as a model for demonstrating how genomic data can be utilised to investigate and address species-specific conservation questions in a threatened species. This approach could be applied to other species such as the critically endangered helmeted honeyeater (*Lichenostomus melanops cassidix*) where previous research showed that pairings between genetically dissimilar mates can improve fitness in this species (Harrisson et al., 2019). Future studies could extend this research by employing WGR to enable a better understanding of how inbreeding affects functional diversity and then use targeted sequencing to determine whether genetic diversity at particular gene regions is associated with improved fitness measures. Such information could assist in the long-term conservation management of the helmeted honeyeater by providing more informed breeding and/or translocation recommendations to maximise functional diversity and conserve adaptive potential, hence improving the long-term viability of the population.

WGR data can also provide the opportunity to address a wide range of other conservation questions that have been highlighted in Chapter 1 and a number of other reviews (see Fuentes-Pardo & Ruzzante, 2017; Hohenlohe, Funk & Rajora, 2021; Khan et al., 2016). Some of the most common examples of WGR applications in conservation (other than the exploration of the genetic basis of phenotypic traits and adaptations) include: i) a better understanding of population size, population structure and the demographic history of populations, ii) resolving species phylogeny to better determine species resolution and identify species hybridisation or conservation units, and iii) assessing the genetic implications of inbreeding on the viability of populations. The WGR datasets generated in this thesis can be employed in future studies to address such questions. For instance, the WGR data generated for the greater bilby in Chapter 5 is currently being used to better understand the effective population size and demographic history of bilby populations. While these are some of the most common applications of WGR data in species conservation, throughout this thesis I have demonstrated a number of additional ways WGR data can be employed to provide new genomic resources and develop new conservation management tools for threatened species. For example, obtaining Y-chromosome information is important both when wanting to investigate specific questions relating to the genetics of male reproductive traits (Chapters 3 and 4), or when wanting to develop population monitoring tools to identify the sex of individuals from non-invasive samples (Chapter 5). Despite the growing number of genomic resources available for species across the phylogenetic tree of life, there is currently still limited Y-chromosome information available for many non-model species, particularly marsupials as marsupial Y-chromosomes are small and can be difficult to sequence (Toder, Wakefield & Graves, 2000). This was particularly evident in Chapter 3 whereby some male reproductive genes were unable to be characterised or investigated in the Tasmanian devil genome due to the absence of Y chromosome data for this species. However, this thesis has shown how WGR data can be used to bioinformatically characterise the Y-chromosome sequence (Chapter 4) and to develop sex-linked markers for wildlife (Chapter 5).

In the case of the antechinus, one of the major aims for creating the first *Antechinus* reference genome in Chapter 4 was to enable future studies to explore the genetic basis of male semelparity (Braithwaite & Lee, 1979). Since a number of important male reproductive genes exist on the Y-chromosome, it is important for future studies examining semelparity in the antechinus to have access to Y-chromosome sequence data. Previous studies have used transcriptomic data from a range of tissues, in conjunction with genomic data, to detect the coding sequence of Y-chromosome genes in marsupials (Cortez et al., 2014). However, this approach is limited as it does not allow for characterisation of complete gene sequences (including introns, UTRs etc), or the organisation of genes along the chromosome (which is important for evolutionary analyses). The bioinformatic method employed in Chapter 4 utilises average read depth information from male and female genomic data to assign male scaffolds as Y-chromosome (Bidon et al., 2015). This method allowed me to identify 0.78Mb of Y-chromosome sequence in the antechinus genome and enabled complete characterisation of a variety of key Y-chromosome genes which will facilitate future studies to investigate the genetic interplay between stress, reproduction and immunity in this species. For example, the Y-chromosome sequence data can be used to design RT-qPCR assays to monitor changes in gene expression of male reproductive genes across the breeding season. Using the antechinus as a model to better understand the genetic mechanisms behind their extreme life history trade-offs could have broad implications for threatened species where a balance between immunity and reproduction is key to species conservation. With high-quality whole-genome sequence data becoming more obtainable, the bioinformatic Y-chromosome assignment method employed in Chapter 4 could be applied to explore conservation questions related to male reproduction in threatened species. For example, significant male reproductive skew and a low proportion of male reproductive success in captive populations of the endangered eastern black rhinoceros (*Diceros bicornis michaeli*) are limiting the viability of ex situ conservation initiatives for this species (Edwards et al., 2015). Y-chromosome information for this species could be used to explore whether variation at male reproductive genes might be associated with observed differences in male reproductive success and hence assist with making more informed captive management decisions.

Y-chromosome information can also be used to develop sex-specific markers for non-invasive sex determination of individuals, which is important in monitoring wild populations of threatened species, particularly those that are cryptic and/or difficult to trap. In Chapter 5, I developed sex-linked markers for the greater bilby using Y-chromosome gene sequences. Furthermore, I showed how Y-chromosome information could be obtained from WGR data in species where a male reference genome is unavailable (and hence the previous approach used in Chapter 4 would not be suitable). In addition to general demographic monitoring of wild populations, development of non-invasive sex-linked markers could also be applied to monitor sex-biased dispersal of threatened species, particularly those in human-dominated landscapes such as the endangered southern brown bandicoot (*Isoodon obesulus obesulus*) (Maclagan et al., 2020). Overall, this thesis has demonstrated how a variety of bioinformatic methods can be applied to common genomic datasets to provide important genomic resources that can assist with species conservation and develop new genomic tools for threatened species management.

Like reference genomes, it is imperative to consider the required quality of accompanying genomic datasets prior to commencement of the study. Sequencing samples to a higher coverage improves the accuracy of results and also enables rare variants to be detected (Sims et al., 2014); however, obtaining high-coverage sequencing data is more costly and requires higher DNA input so may reduce the number of individuals that can be sequenced (Fuentes-Pardo & Ruzzante, 2017). High-coverage data (~30×) is important for accurate variant identification (Chapters 3, 4 and 5), but low-coverage datasets (~5×) can be useful to explore known variants across a larger number of individuals (Chapter 3), or for general population genetic monitoring (Benjelloun et al., 2019). In Chapter 3, I demonstrated how high and low coverage datasets can be used together by first performing initial variant identification with high coverage datasets and then incorporating low coverage data to further explore the identified variants across a greater number of samples. This method is useful for combining pre-existing genomic datasets with differing coverages, or when wanting to balance the accuracy of the data with the cost of the number of individuals genotyped. As sequencing costs continue to fall, high coverage WGR data across many individuals will become more achievable, enabling greater accuracy and power for a variety of broad applications to assist in the conservation threatened species.

6.2 TACKLING THE BIOINFORMATICS OF LARGE GENOMIC DATASETS

For a non-bioinformatician, knowing how to start using large genomic datasets is a significant challenge. The analysis of next-generation sequencing data often requires significant computational power (particularly for organisms with large genomes such as mammals), which means most bioinformatic analyses need to be conducted on the command-line of HPC or cloud-based infrastructure (see Chapter 5, Supplementary Table 1). Without a strong background in bioinformatics, understanding how to tackle the analysis of large-scale genomic data is one of the major hurdles that researchers are currently facing. To resolve this, in Chapter 2 I present ten simple rules which provide researchers with the background knowledge needed to get started with command-line bioinformatics. The scientific literature is a valuable resource that can provide summaries of the bioinformatic methods employed and the software used to analyse genomic datasets across a wide range of species and contexts. However, determining which analysis to use, what computational resources are required, and how to run the analysis on a new dataset is no small feat, and may result in researchers being hesitant to make use of the latest genomic technologies. My ten simple rules can assist researchers to overcome these bioinformatic hurdles, which should facilitate greater development and use of genomic resources in conservation contexts. One of the main points raised throughout Chapter 2 is that reaching out to other researchers to determine: i) what software may work best for the species of interest or chosen data type, ii) what compute resources are required to run such software; and iii) whether that software has already been optimised or made available on a particular platform, is one of the best ways for researchers to get started running large bioinformatic analyses. Another solution is for genomic publications to publish (in the supplementary section at the very least) the types of compute resources required to undertake their analysis. While it is currently not standard practice to report the specific details surrounding computational requirements for any bioinformatic analyses conducted, this information is helpful in informing other researchers of some general guidelines so that they can determine where to run such analyses and how much it may cost when wanting to apply the same bioinformatic methods on similar species or similar genomic data. In Chapter 5 I provided an example of how such information could be presented in Supplementary

Table 1. As the use of genomic data becomes more common place it is important for researchers to continue to publish how they produce their data beyond just stating the software used.

Throughout Chapters 3 to 5 I employed a variety of bioinformatic techniques to answer specific conservation questions for threatened species. In particular, I showed how approaches from previous studies can be employed in novel ways to answer key conservation questions in other species or using new datasets. For instance, the approach that was previously used to assess immune gene diversity in the endangered Tasmanian devil (Morris et al., 2015) was employed to explore reproductive gene diversity in the same species (Chapter 3). Similarly, a technique previously used to characterise Y-chromosome sequence in the polar bear (*Ursus maritimus*) (Bidon et al., 2015) was used to identify Y-chromosome scaffolds in the brown antechinus (Chapter 4). Additionally, I have shown that novel bioinformatic approaches can be employed to answer specific research questions or conservation needs. For example, using WGR on a small number of individuals to determine whether population management based on RRS data is enough to maximise functional diversity across a metapopulation, or to develop sex-linked markers for monitoring wild populations (Chapter 5). The ten simple rules provided in Chapter 2 will enable researchers to employ the bioinformatic techniques presented throughout this thesis to other species, facilitating the use of next generation sequencing technologies as a tool to assist in species conservation.

6.3 AREAS FOR FUTURE WORK

Currently, of the 15,500 animal species listed as threatened by the IUCN, approximately only 100 of these species have reference genomes available. Encouraging the creation of reference genomes for more threatened species is important in facilitating genomic research for conservation purposes (Chapter 1). Funding limitations, sample input requirements and bioinformatic expertise are some of the main barriers preventing the generation of these valuable resources for many threatened species. Here I have addressed a number of these limitations and expanded the current knowledge base by: i) demonstrating that affordable genomes generated from short-read technologies are sufficient for answering a large variety of conservation questions (Chapters 1 and 3); ii) exhibiting how reference genomes for non-threatened species (where sample acquisition may be more feasible) can be created as a resource for related threatened species (Chapter 4); iii) showing that opportunistic sample acquisition and preservation can enable high quality reference genomes to be obtained for threatened species (Chapter 5); and iv) providing ten simple rules to assist with the bioinformatic knowledge required for reference genome creation (Chapter 2). As DNA extraction techniques and sequencing technologies continue to improve, generation of reference genomes for species of interest should become more accessible and cost effective. The establishment of large national and international sequencing consortia will also facilitate the availability of reference genomes for threatened species or their closely related counterparts. Conservation researchers should take advantage of these high-quality genomic resources and work with conservation practitioners to identify how genomic data can assist in reaching management objectives.

Secondly, bioinformatics support for researchers wanting to use next generation sequencing data is vital in preventing further widening of the research-implementation gap. Encouraging more transparency with computational requirements for bioinformatic analyses will greatly assist researchers in understanding what infrastructure is required for working with genomic datasets and what costs may be involved. Future studies should include details regarding the required computing resources for bioinformatic analyses (e.g., Supplementary Table 1 of Chapter 5) to facilitate this. Developing other ways for researchers to share bioinformatic expertise and work together to solve complex bioinformatics problems will also permit greater use of genomic datasets in conservation management. Throughout my PhD I worked closely with the recently established Australian BioCommons (Australian BioCommons, 2019; Lonie & Francis, 2020) which aims to support life science research in Australia by providing researchers with the tools, methods and training required to undertake bioinformatic analyses. As part of this collaborative effort, I developed documentation for a range of complex bioinformatic pipelines (such as genome assembly and annotation) for the Australian BioCommons and the Australian genome community (Appendix 4). I have also been a member of several bioinformatics working groups that aim to facilitate the sharing and collaboration of information related to bioinformatic analyses among Australian researchers. Such initiatives are crucial in fostering and enabling researchers to utilise genomic resources for important downstream applications and to prevent further barriers between genomics and conservation.

Finally, future work should focus on ensuring genomic research findings are more accessible to conservation managers. One important future step towards this goal is for more publications to provide clear implications of conservation genomic research findings on threatened species management (Britt et al., 2018). Describing how research findings have or could be used to assist conservation efforts will not only inspire other researchers to see what's possible with genomic datasets but also provide a library of examples to encourage greater incorporation of genomic data into species conservation worldwide. However, just including management recommendations in academic publications is not enough, researchers need to build strong relationships with the conservation industry and communicate their results directly with management teams to explore how findings can be implemented into conservation efforts (Britt et al., 2018; Galla et al., 2016; Hohenlohe, Funk & Rajora, 2021; Shafer et al., 2015; Taylor, Dussex & van Heezik, 2017). Ensuring conservation teams also understand the value of genomic data and that researchers are providing genomic resources that are in line with the goals of species recovery is also vital. For example, we are already working with the National Bilby Recovery Team and metapopulation management group to ensure such findings and tools from Chapter 5 are implemented into the current bilby metapopulation management practice. Towards the end of my PhD, the Threatened Species Initiative (TSI) was launched which aims to generate genomic resources to assist in the conservation management of threatened species across Australia (Threatened Species Initiative, 2020; Hogg et al, in press). I have been involved with the pilot phase of this program which involves the creation of a user-friendly web portal where standard genomic data types can be submitted, analysed and translated into a single standardised report that conservation managers can use to inform their current population management strategies. Initiatives such as this are momentous in bringing researchers and conservation agencies together and bridging the research-implementation gap.

6.4 CONCLUSION

In the United Nations Decade on Restoration 2021-2030 (United Nations Environment Programme, 2021), being able to harness the power of genomic data to inform threatened species management is fundamental to conserving the world's biodiversity. With the rapid progression of sequencing technologies, the ability to obtain reference genomes and WGR data for threatened species will become commonplace, arming conservation researchers with the tools they need to better estimate the size, structure and demographic history of populations, explore the genetic basis of specific traits, and detect adaptive variation within populations (Fuentes-Pardo & Ruzzante, 2017; Hohenlohe, Funk & Rajora, 2021; Supple & Shapiro, 2018). These invaluable insights will enable conservation practitioners to make more informed population management decisions to promote, monitor and maintain the adaptive potential of threatened populations and ensure species have the best chance of long-term survival.

This thesis not only provides a variety of valuable genomic resources that will assist conservation efforts for a number of threatened Australian marsupial species, but also provides researchers and conservation managers with the bioinformatic background and tools needed to successfully employ genomic data into species conservation. The examples, tools, recommendations, and resources I have provided herein will aid researchers and conservation managers to overcome many of the barriers driving the current gap between genomics and conservation and encourage the implementation of genomic data into threatened species management globally.

As we continue to build a catalogue of genomic resources for both threatened and non-threatened species worldwide, it is vital that researchers and conservation managers work together to harness the power of this data for species conservation. Future research should continue to take steps towards bridging the research-implementation gap by extending the work presented in this thesis to other threatened organisms. Amid a global biodiversity crisis, and the rising era of genomics, researchers and conservationists together hold the key to preserving and protecting our planet's biodiversity.

6.4 REFERENCES

Aiden Lab 2018, *DNA Zoo*, viewed 5 May 2021, https://www.dnazoo.org

Andrews, KR, Good, JM, Miller, MR, Luikart, G & Hohenlohe, PA 2016, 'Harnessing the power of RADseq for ecological and evolutionary genomics', *Nature Reviews Genetics,* vol. 17, no. 2, pp. 81-92.

Australian BioCommons 2019, *Enhancing Australia’s digital life science research through world class collaborative distributed infrastructure*, viewed 26 March 2021, https://www.biocommons.org.au

Baker, AM, Mutton, TY & Hines, HB 2013, 'A new dasyurid marsupial from Kroombit Tops, south-east Queensland, Australia: the silver-headed antechinus, *Antechinus argentus* sp. nov.(Marsupialia: Dasyuridae)', *Zootaxa,* vol. 3746, no. 2, pp. 201-239.

Ballou, JD & Lacy, RC 1995, 'Identifying genetically important individuals for management of genetic variation in pedigreed populations', in Ballou, J, Gilpin, M & Foose, T (eds.), *Population Management for Survival and Recovery: Analytical Methods and Strategies in Small Population Conservation*, Columbia University Press, New York, USA.

Ballou, JD, Lees, C, Faust, LJ, Long, S, Lynch, C, Bingaman Lackey, L & Foose, TJ 2010, 'Demographic and genetic management of captive populations', *Wild Mammals in Captivity: Principles and Techniques for Zoo Management,* 2nd edn, The University of Chicago Press, Chicago, IL, USA.

Benjelloun, B, Boyer, F, Streeter, I, Zamani, W, Engelen, S, Alberti, A, Alberto, FJ, Benbati, M, Ibnelbachyr, M & Chentouf, M 2019, 'An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity', *Molecular Ecology Resources,* vol. 19, no. 6, pp. 1497-1515.

Bidon, T, Schreck, N, Hailer, F, Nilsson, MA & Janke, A 2015, 'Genome-wide search identifies 1.9 Mb from the polar bear Y chromosome for evolutionary analyses', *Genome Biology and Evolution,* vol. 7, no. 7, pp. 2010-2022.

Braithwaite, RW & Lee, AK 1979, 'A mammalian example of semelparity', *The American Naturalist,* vol. 113, no. 1, pp. 151-155.

Britt, M, Haworth, SE, Johnson, JB, Martchenko, D & Shafer, AB 2018, 'The importance of non-academic coauthors in bridging the conservation genetics gap', *Biological Conservation,* vol. 218, pp. 118-123.

Burton, JN, Adey, A, Patwardhan, RP, Qiu, R, Kitzman, JO & Shendure, J 2013, 'Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions', *Nature Biotechnology,* vol. 31, no. 12, pp. 1119-1125.

Cortez, D, Marin, R, Toledo-Flores, D, Froidevaux, L, Liechti, A, Waters, PD, Grützner, F & Kaessmann, H 2014, 'Origins and functional evolution of Y chromosomes across mammals', *Nature,* vol. 508, no. 7497, pp. 488-493.

Dudchenko, O, Batra, SS, Omer, AD, Nyquist, SK, Hoeger, M, Durand, NC, Shamim, MS, Machol, I, Lander, ES & Aiden, AP 2017, 'De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds', *Science,* vol. 356, no. 6333, pp. 92-95.

Edwards, KL, Walker, SL, Dunham, AE, Pilgrim, M, Okita-Ouma, B & Shultz, S 2015, 'Low birth rates and reproductive skew limit the viability of Europe’s captive eastern black rhinoceros, *Diceros bicornis michaeli*', *Biodiversity and Conservation,* vol. 24, no. 11, pp. 2831-2852.

Farquharson, KA, Hogg, CJ & Grueber, CE 2017, 'Pedigree analysis reveals a generational decline in reproductive success of captive Tasmanian devil

(*Sarcophilus harrisii*): implications for captive management of threatened species', *Journal of Heredity,* vol. 108, no. 5, pp. 488-495.

Frankham, R 2008, 'Genetic adaptation to captivity in species conservation programs', *Molecular Ecology,* vol. 17, no. 1, pp. 325-333.

Frankham, R, Ballou, JD & Briscoe, DA 2010, *Introduction to Conservation Genetics*, 2 edn, Cambridge University Press, Cambridge, UK.

Frankham, R, Ballou, JD, Ralls, K, Eldridge, M, Dudash, MR, Fenster, CB, Lacy, RC & Sunnucks, P 2017, *Genetic Management of Fragmented Animal and Plant Populations*, Oxford University Press, New York, USA.

Frankham, R, Hemmer, H, Ryder, O, Cothran, E, Soulé, M, Murray, N & Snyder, M 1986, 'Selection in captive populations', *Zoo Biology,* vol. 5, no. 2, pp. 127-138.

Fuentes-Pardo, AP & Ruzzante, DE 2017, 'Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations', *Molecular Ecology,* vol. 26, no. 20, pp. 5369-5406.

Galla, SJ, Buckley, TR, Elshire, R, Hale, ML, Knapp, M, McCallum, J, Moraga, R, Santure, AW, Wilcox, P & Steeves, TE 2016, 'Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances', *Molecular Ecology,* vol. 25, no. 21, pp. 5267-5281.

Galla, SJ, Forsdick, NJ, Brown, L, Hoeppner, M, Knapp, M, Maloney, RF, Moraga, R, Santure, AW & Steeves, TE 2019, 'Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management', *Genes,* vol. 10, no. 9, pp. 1-19.

Gooley, R, Hogg, CJ, Belov, K & Grueber, CE 2017, 'No evidence of inbreeding depression in a Tasmanian devil insurance population despite significant variation in inbreeding', *Scientific Reports,* vol. 7, no. 1830, pp. 1-11.

Gupta, P, Robin, V & Dharmarajan, G 2020, 'Towards a more healthy conservation paradigm: integrating disease and molecular ecology to aid biological conservation', *Journal of Genetics,* vol. 99, no. 1, pp. 1-26.

Harrisson, KA, Magrath, MJ, Yen, JD, Pavlova, A, Murray, N, Quin, B, Menkhorst, P, Miller, KA, Cartwright, K & Sunnucks, P 2019, 'Lifetime fitness costs of inbreeding and being inbred in a critically endangered bird', *Current Biology,* vol. 29, no. 16, pp. 2711-2717. e4.

Hayward, MW, Poh, ASL, Cathcart, J, Churcher, C, Bentley, J, Herman, K, Kemp, L, Riessen, N, Scully, P & Diong, CH 2015, 'Numbat nirvana: conservation ecology of the endangered numbat (*Myrmecobius fasciatus*)(Marsupialia: Myrmecobiidae) reintroduced to Scotia and Yookamurra Sanctuaries, Australia', *Australian Journal of Zoology,* vol. 63, no. 4, pp. 258-269.

Hoelzel, AR, Bruford, MW & Fleischer, RC 2019, 'Conservation of adaptive potential and functional diversity', *Conservation Genetics,* vol. 20, pp. 1-5.

Hogg, CJ, Grueber, CE, Pemberton, D, Fox, S, Lee, AV, Ivy, JA & Belov, K 2017, '"Devil Tools & Tech": a synergy of conservation research and management practice', *Conservation Letters,* vol. 10, no. 1, pp. 133-138.

Hogg, CJ, Ottewell, K, Latch, P, Rossetto, M, Biggs, J, Gilbert, A, Richmond, S & Belov, K In press, 'Threatened Species Initiative – empowering conservation action using genomic resources'. *Cell Genomics*

Hohenlohe, PA, Funk, WC & Rajora, OP 2021, 'Population genomics for wildlife conservation and management', *Molecular Ecology,* vol. 30, no. 1, pp. 62-82.

Holderegger, R, Kamm, U & Gugerli, F 2006, 'Adaptive vs. neutral genetic diversity: implications for landscape genetics', *Landscape Ecology,* vol. 21, no. 6, pp. 797-807.

Howe, KL, Contreras-Moreira, B, De Silva, N, Maslen, G, Akanni, W, Allen, J, Alvarez-Jarreta, J, Barba, M, Bolser, DM & Cambell, L 2020, 'Ensembl Genomes 2020—enabling non-vertebrate genomic research', *Nucleic Acids Research,* vol. 48, no. D1, pp. D689-D695.

IUCN 2020, *The IUCN Red List of Threatened Species. Version 2020-3*, viewed 16 March 2021, http://www.iucnredlist.org

Johnson, RN, O'Meally, D, Chen, Z, Etherington, GJ, Ho, SYW, Nash, WJ, Grueber, CE, Cheng, Y, Whittington, CM, Dennison, S, Peel, E, Haerty, W, O'Neill, RJ, Colgan, D, Russell, TL, Alquezar-Planas, DE, Attenbrow, V, Bragg, JG, Brandies, PA, Chong, AY-Y, Deakin, JE, Di Palma, F, Duda, Z, Eldridge, MDB, Ewart, KM, Hogg, CJ, Frankham, GJ, Georges, A, Gillett, AK, Govendir, M, Greenwood, AD, Hayakawa, T, Helgen, KM, Hobbs, M, Holleley, CE, Heider, TN, Jones, EA, King, A, Madden, D, Graves, JaM, Morris, KM, Neaves, LE, Patel, HR, Polkinghorne, A, Renfree, MB, Robin, C, Salinas, R, Tsangaras, K, Waters, PD, Waters, SA, Wright, B, Wilkins, MR, Timms, P & Belov, K 2018, 'Adaptation and conservation insights from the koala genome', *Nature Genetics,* vol. 50, no. 8, pp. 1102-1111.

Jones, MR & Good, JM 2016, 'Targeted capture in evolutionary and ecological genomics', *Molecular Ecology,* vol. 25, no. 1, pp. 185-202.

Khan, S, Nabi, G, Ullah, MW, Yousaf, M, Manan, S, Siddique, R & Hou, H 2016, 'Overview on the role of advance genomics in conservation biology of endangered species', *International Journal of Genomics,* vol. 2016, pp. 1-8.

Kosch, T, Silva, C, Brannelly, L, Roberts, A, Lau, Q, Marantelli, G, Berger, L & Skerratt, L 2019, 'Genetic potential for disease resistance in critically endangered amphibians decimated by chytridiomycosis', *Animal Conservation,* vol. 22, no. 3, pp. 238-250.

Kosch, TA, Eimes, JA, Didinger, C, Brannelly, LA, Waldman, B, Berger, L & Skerratt, LF 2017, 'Characterization of MHC class IA in the endangered southern corroboree frog', *Immunogenetics,* vol. 69, no. 3, pp. 165-174.

Larsen, PA & Matocq, MD 2019, 'Emerging genomic applications in mammalian ecology, evolution, and conservation', *Journal of Mammalogy,* vol. 100, no. 3, pp. 786-801.

Lewin, HA, Robinson, GE, Kress, WJ, Baker, WJ, Coddington, J, Crandall, KA, Durbin, R, Edwards, SV, Forest, F & Gilbert, MTP 2018, 'Earth BioGenome Project: Sequencing life for the future of life', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 115, no. 17, pp. 4325-4333.

Lonie, A & Francis, R 2020, 'Australian BioCommons Strategic Plan 2019-2023', Zenodo

Maclagan, S, Coates, T, Hradsky, B, Butryn, R & Ritchie, E 2020, 'Life in linear habitats: the movement ecology of an endangered mammal in a peri-urban landscape', *Animal Conservation,* vol. 23, no. 3, pp. 260-272.

Margres, MJ, Jones, ME, Epstein, B, Kerlin, DH, Comte, S, Fox, S, Fraik, AK, Hendricks, SA, Huxtable, S & Lachish, S 2018, 'Large-effect loci affect survival in Tasmanian devils (*Sarcophilus harrisii*) infected with a transmissible cancer', *Molecular Ecology,* vol. 27, no. 21, pp. 4189-4199.

Martin, AM & Carver, S 2020, 'Ecology and conservation of the critically endangered northern hairy-nosed wombat (*Lasiorhinus krefftii*): past, present and future', *Australian Mammalogy,* vol. 43, pp. 10-21.

McLennan, EA, Wright, BR, Belov, K, Hogg, CJ & Grueber, CE 2019, 'Too much of a good thing? Finding the most informative genetic data set to answer conservation questions', *Molecular Ecology Resources,* vol. 19, no. 3, pp. 659-671.

McMahon, BJ, Teeling, EC & Höglund, J 2014, 'How and why should we implement genomics into conservation?', *Evolutionary Applications,* vol. 7, no. 9, pp. 999-1007.

Morris, KM, Wright, B, Grueber, CE, Hogg, C & Belov, K 2015, 'Lack of genetic diversity across diverse immune genes in an endangered mammal, the Tasmanian devil (*Sarcophilus harrisii*)', *Molecular Ecology,* vol. 24, no. 15, pp. 3860-3872.

Mulvena, SR, Pierson, JC, Farquharson, KA, McLennan, EA, Hogg, CJ & Grueber, CE 2020, 'Investigating inbreeding in a free-ranging, captive population of an Australian marsupial', *Conservation Genetics,* vol. 21, pp. 665-675.

Murchison, EP, Schulz-Trieglaff, OB, Ning, Z, Alexandrov, LB, Bauer, MJ, Fu, B, Hims, M, Ding, Z, Ivakhno, S & Stewart, C 2012, 'Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer', *Cell,* vol. 148, no. 4, pp. 780-791.

Narum, SR, Buerkle, CA, Davey, JW, Miller, MR & Hohenlohe, PA 2013, 'Genotyping-by-sequencing in ecological and conservation genomics', *Molecular Ecology,* vol. 22, no. 11, pp. 2841-3190.

O'Leary, NA, Wright, MW, Brister, JR, Ciufo, S, Haddad, D, McVeigh, R, Rajput, B, Robbertse, B, Smith-White, B & Ako-Adjei, D 2015, 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research,* vol. 44, no. D1, pp. D733-D745.

Pacioni, C, Atkinson, A, Trocini, S, Rafferty, C, Morley, K & Spencer, PB 2020, 'Is supplementation an efficient management action to increase genetic diversity in translocated populations?', *Ecological Management & Restoration,* vol. 21, no. 2, pp. 123-130.

Rhie, A, McCarthy, SA, Fedrigo, O, Damas, J, Formenti, G, Koren, S, Uliano-Silva, M, Chow, W, Fungtammasan, A & Kim, J 2021, 'Towards complete and error-free genome assemblies of all vertebrate species', *Nature,* vol. 592, no. 7856, pp. 737-746.

Schultz, AJ, Cristescu, RH, Littleford-Colquhoun, BL, Jaccoud, D & Frère, CH 2018, 'Fresh is best: Accurate SNP genotyping from koala scats', *Ecology and Evolution,* vol. 8, no. 6, pp. 3139-3151.

Selkoe, KA & Toonen, RJ 2006, 'Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers', *Ecology Letters,* vol. 9, no. 5, pp. 615-629.

Shafer, AB, Wolf, JB, Alves, PC, Bergström, L, Bruford, MW, Brännström, I, Colling, G, Dalén, L, De Meester, L & Ekblom, R 2015, 'Genomics and the challenging translation into conservation practice', *Trends in Ecology & Evolution,* vol. 30, no. 2, pp. 78-87.

Sims, D, Sudbery, I, Ilott, NE, Heger, A & Ponting, CP 2014, 'Sequencing depth and coverage: key considerations in genomic analyses', *Nature Reviews Genetics,* vol. 15, no. 2, pp. 121-132.

Supple, MA & Shapiro, B 2018, 'Conservation of biodiversity in the genomics era', *Genome Biology,* vol. 19, no. 131, pp. 1-12.

Taylor, HR, Dussex, N & Van Heezik, Y 2017, 'Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners', *Global Ecology and Conservation,* vol. 10, pp. 231-242.

Threatened Species Initiative 2020, viewed 26 March 2021, https://threatenedspeciesinitiative.com

Toder, R, Wakefield, M & Graves, J 2000, 'The minimal mammalian Y chromosome–the marsupial Y as a model system', *Cytogenetic and Genome Research,* vol. 91, no. 1-4, pp. 285-292.

United Nations Environment Programme 2021, *United Nations Decade on Ecosystem Restoration 2021-2030*, viewed 16 March 2021, https://www.decadeonrestoration.org

Weisenfeld, NI, Kumar, V, Shah, P, Church, DM & Jaffe, DB 2017, 'Direct determination of diploid genome sequences', *Genome Research,* vol. 27, no. 5, pp. 757-767.

Wenger, AM, Peluso, P, Rowell, WJ, Chang, P-C, Hall, RJ, Concepcion, GT, Ebler, J, Fungtammasan, A, Kolesnikov, A & Olson, ND 2019, 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature Biotechnology,* vol. 37, no. 10, pp. 1155-1162.

White, LC, Horsup, A, Taylor, AC & Austin, JJ 2014, 'Improving genetic monitoring of the northern hairy-nosed wombat (*Lasiorhinus krefftii*)', *Australian Journal of Zoology,* vol. 62, no. 3, pp. 246-250.

APPENDICES

APPENDIX 1: PDF VERSIONS OF PUBLISHED MANUSCRIPTS

A1.1 THE VALUE OF REFERENCE GENOMES IN THE CONSERVATION OF THREATENED SPECIES

The PDF version of the review titled “The Value of Reference Genomes in the Conservation of Threatened Species” published in *Genes* (2019; 10 (11), 846), which was included in Chapter 1 of this thesis, is presented on the following pages.

genes

Review

The Value of Reference Genomes in the Conservation of Threatened Species

Parice Brandies, Emma Peel, Carolyn J. Hogg and Katherine Belov *

School of Life & Environmental Sciences, The University of Sydney, Sydney 2006, Australia; parice.brandies@sydney.edu.au (P.B.); emma.peel@sydney.edu.au (E.P.); carolyn.hogg@sydney.edu.au (C.H.)

* Correspondence: kathy.belov@sydney.edu.au

Received: 18 October 2019; Accepted: 23 October 2019; Published: 25 October 2019

Abstract: Conservation initiatives are now more crucial than ever—over a million plant and animal species are at risk of extinction over the coming decades. The genetic management of threatened species held in insurance programs is recommended; however, few are taking advantage of the full range of genomic technologies available today. Less than 1% of the 13505 species currently listed as threated by the International Union for Conservation of Nature (IUCN) have a published genome. While there has been much discussion in the literature about the importance of genomics for conservation, there are limited examples of how having a reference genome has changed conservation management practice. The Tasmanian devil (*Sarcophilus harrisii*), is an endangered Australian marsupial, threatened by an infectious clonal cancer devil facial tumor disease (DFTD). Populations have declined by 80% since the disease was first recorded in 1996. A reference genome for this species was published in 2012 and has been crucial for understanding DFTD and the management of the species in the wild. Here we use the Tasmanian devil as an example of how a reference genome has influenced management actions in the conservation of a species.

Keywords: conservation; genomes; Tasmanian devil

1. Introduction

We are currently in the midst of a global sixth mass extinction event, with biodiversity rapidly declining around the world [1], and extinction rates are accelerating [2]. Australia has the worst mammal extinction rate of any country, with 25 mammals declared extinct since European settlement and almost 20% of current mammalian species listed as vulnerable [2–5]. This significant decline is concerning as Australia is one of seventeen "megadiverse" countries that comprises a large proportion of the Earth's biological diversity [6]. Megadiverse countries have at least 5,000 endemic plant species and have marine ecosystems within its borders [6]. In addition to this, 87% of Australian mammals, 93% of Australian reptiles, and 94% of Australian frogs are endemic to Australia [7]. Therefore, conservation initiatives that protect and maintain Australia's biodiversity are now more crucial than ever.

Only 39% of the 1,890 Australian species (517 animals; 1373 plants), listed as threatened under the Environment Protection and Biodiversity Conservation Act (EPBC Act), have a recovery plan in place to improve their threat status [8]. These recovery plans set out management and research actions to slow population decline and promote recovery of threatened species and communities. This is achieved by providing a framework for key interest groups and government agencies to coordinate their efforts to improve the plight of threatened species [8]. Management actions range from mitigating threatening processes such as predation, habitat loss, or change, in addition to research into basic species biology, ecosystem integration, and genetics. The main goal of recovery plans is to maintain the long-term viability of a chosen population/community. Maintaining genetic diversity is an important component of population viability as it assists with mitigating negative effects associated with inbreeding and arms populations with the potential to adapt to future environmental change [9–11]. As such, understanding a populations' inherent genetic diversity, in addition to their historical diversity and future potential, is of utmost importance in species conservation. For this reason, more than 80% of the current 200 Australian national vertebrate recovery plans have some form of genetic action listed in the species' recovery plan. Yet, less than 15% of these recovery plans have any form of genetic or genomic data available, either in existence or currently in development. Here we refer to genetic data as information based on specific, limited regions of the genome (e.g., targeted gene sequencing, microsatellite analysis, etc.), whilst genomic data is information based on the whole genome (e.g., whole genome sequencing/resequencing, whole-genome single nucleotide polymorphism (SNP) analysis/reduced representation sequencing, etc.).

Advances in sequencing technologies and the reduction in sequencing costs have given rise to the era of genomics, whereby holistic genome-wide approaches are rapidly replacing traditional genetic marker approaches in many non-model species [12–14]. Although recent reviews have highlighted the importance of implementing genomic data into conservation initiatives [13,15,16], the application of such powerful advances in sequencing technologies is lacking in the current literature. This limited use in conservation may be due to a number of reasons including: costs, a lack of understanding of the potential of new genomics approaches, lack of expertise in developing and utilizing the data, and the absence of a reference genome for the species of interest (or a closely-related species) [13,15,17]. The latter is an important concern as the generation of a reference genome requires considerable expertise, funds, computational resources, and time that are not often accessible by wildlife managers and conservation teams [15,18].

Of the 13505 animal species that are listed as threatened (Lower Risk/Conservation Dependent or worse) on the International Union for Conservation of Nature (IUCN) Red List [2], 108 (< 1%) have published genomes on NCBI [19]. This equates to only 6% of the 1842 animal genomes currently available on NCBI [19]. Creating high-quality reference genomes that can provide insights into species evolution and biology is a costly task (~$30,000 for an average eukaryotic genome size of 2.5 Gbp [20]), and also requires large collaborative groups to provide expertise from varying fields (e.g., [21–23]). Fortunately, in recent years a number of national and international consortia and genome projects have been formed with the aim of creating high-quality reference genomes for species spanning the phylogenetic tree of life including: the Earth Biogenome Project (EBP) [20], the Genome 10K Project (G10K) [24,25], the Vertebrate Genomes Project (VGP) [26], the Bird 10K Project (B10K) [27], the Bat 1K Project (Bat1K) [28], the Global Invertebrate Genomics Alliance (GIGA) [29,30], and the Oz Mammal Genomics initiative (OMG) [31], to name a few. The goal of many of these consortia is to bring together the required expertise to generate reference genomes of a sufficient quality, which are publicly available to the science community, thereby providing the vital resources required to implement genomics into conservation management better [13,15,18]. However, just providing the reference genomes or genomic data is not enough to improve conservation outcomes. Geneticists need to continually communicate how genomic techniques can be utilized in a cost-effective manner to assist species conservation better [17,32]. As highlighted by Taylor et al. [33], targeted education and training is also required to teach conservation managers how to interpret and utilize genomic data. To better assist conservation managers, a number of groups and communities have already been established to assist in providing conservation genetics advice for threated species management. These include the IUCN/SSC (Species Survival Commission) Conservation Genetics Specialist Group (CGSG), the Genetic Composition Working Group of GEO BON (Group on Earth Observations Biodiversity Observation Network), and the pan-European COST (Cooperation in Science and Technology) action ConGRESS (Conservation Genetic Resources for Effective Species Survival) (for further information and examples from these groups, see Holderegger et al. [34]). Conservationists in their respective countries can get in touch with these groups to obtain the contact details of geneticists who work in their region who may be able to assist them with their management needs.

While a number of papers have reviewed current genomic techniques and the way they can, or have been, applied to assist in conservation decisions across species [15,17], questions are still raised as to whether reference genomes are necessary for species conservation. Reference genomes hold the key to investigate a number of paradigms that are essential for species conservation, including: demography, inbreeding, hybridization, disease susceptibility, behavioral ecology, and adaptation [12,13,15,16,18]. Here we demonstrate the value of a reference genome to the conservation effort of an endangered species, the Tasmanian devil (*Sarcophilus harrisii*), and how this information has been applied in real-time management practice [35].

The Tasmanian devil, an endangered Australian marsupial, is often used in the literature as an example of how genetics/genomics approaches could be used in conservation [12,13,36]. However, something that is not often discussed is that having a reference genome for this species is one of the key factors that contributed to using genomics in management practice. Although this species has a unique conservation issue, low genetic diversity coupled with an infectious clonal cancer, the methods described herein apply to many other threatened species. Here we show how the reference genome has allowed a range of conservation questions to be answered in a timely, cost-effective manner and enabled conservation researchers to adapt to the rapid advances in genomic technologies.

2. The Tasmanian Devil and Its Genome

The Tasmanian devil is the largest extant carnivorous marsupial, native to mainland Tasmania, Australia. The emergence of transmissible cancer, devil facial tumor disease (DFTD) in the mid-1990s has led to a rapid population decline of up to 80% across their range [37]. In 2003, the Tasmanian and Australian governments responded to the disease threat by establishing the Save the Tasmanian Devil Program (STDP). Since then, researchers, wildlife managers, and the zoo industry have worked closely with the STDP to ensure that Tasmanian devils have a sustainable ecological function in the Tasmanian ecosystem and landscape [35,38]. This work has included a range of activities such as monitoring of wild populations, developing an insurance population, describing and characterizing the disease, and developing new genomic tools to understand the disease and the Tasmanian devil [38].

Prior to the publication of a reference genome for the Tasmanian devil, traditional genetic approaches such as MHC (major histocompatibility complex) typing and microsatellite analysis were used to explore genetic diversity at specific genes as well as general genetic diversity in the species [39–41]. These techniques were able to show that the Tasmanian devil had low genetic diversity [39–42]. However, the low rates of polymorphism for most of these markers did not have high enough resolution to assist in answering crucial conservation questions such as determining founder relatedness within the insurance population [43,44], identifying high-resolution population substructure [45], or to better understand the origin and evolution of DFTD [46]. In instances such as these, further genomic data was required to improve resolution. For other threatened species, where there may be moderate to high genome-wide diversity, microsatellite markers may be highly polymorphic, and so these markers have value as a continuing genetic management tool.

To overcome this knowledge gap, the Tasmanian devil genome was sequenced independently by two different research groups in 2011 [45,46]. Miller et al. [45] sequenced the nuclear genome of two individuals (originating from extreme northwest and southeast Tasmania), as well as the tumor from one individual, using both Roche and Illumina sequencing platforms. The analysis of genome-wide SNPs confirmed low genetic diversity across the Tasmanian devil genome, as well as enabling the construction of genotyping arrays, which revealed a new population substructure and the identification of tumor-specific SNPs. However, the low contiguity of this reference genome assembly (148,891 scaffolds, scaffold N50 147 kb) limited the applicability of the data in downstream research. In 2012, a more contiguous, annotated nuclear genome (35974 scaffolds, scaffold N50 1.85 Mb), and tumor genome was published by Murchison et al. [46], resulting in the primary reference genome used today. This higher quality assembly facilitated an enormous effort in downstream genetic and genomic research. It should be noted that as of August 2019, the 2012 Tasmanian devil reference genome paper [46] has been cited over 200 times (Google Scholar Citation Search), highlighting the value of this reference genome to the research community. It is not possible to cover all of the research that has stemmed from the sequencing of the 2012 genome here. Rather, here, we present key examples of how having a reference genome has contributed to conservation decisions and outcomes for the Tasmanian devil. We also note that at the time of this publication, an updated Tasmanian devil genome assembly has been released [47]. This assembly utilized an in vitro proximity ligation technique to further improve the scaffolding of the 2012 assembly (10010 scaffolds, N50 7.75 Mb); however, chromosome assignment and annotation have not been performed at this stage.

3. Conservation Applications as a Result of a Reference Genome

3.1. Basic Conservation Management

3.1.1. Microsatellite Analysis

Traditionally population genetic measures to answer basic questions regarding population structure, population size, population dynamics (migration, bottlenecks), kinship, inbreeding, etc. [14,48] have used microsatellites, or short tandem repeats [48]. Where microsatellite markers have already been developed for the species of interest, or in a closely related species that may carry similar markers, they provide a cost-effective, quick conservation management tool [48,49]. However, for those species where appropriate microsatellite markers are not currently available, or cross-species microsatellite amplification is not effective, and a reference genome is also not available, considerable time and resources are required to develop species-specific microsatellite markers. For example, prior to sequencing the Tasmanian devil genome, 11 putatively neutral microsatellite markers were developed to assess genetic diversity in Tasmanian devils [39]. The development of these microsatellites involved the creation and screening of a genomic library, sequencing of positive clones, primer design, and PCR optimization [39]. Several years later, MHC-linked microsatellite markers were developed in a similar manner as a cheaper and faster method of investigating MHC diversity when compared to traditional MHC typing techniques, such as cloning and sequencing particular MHC regions [41]. This traditional microsatellite isolation and the marker development approaches require considerable laboratory expertise, time, and funds [49], that today may be better spent developing more powerful molecular approaches (see Reduced Representation Sequencing section below).

Contrarily, the availability of the Tasmanian devil reference genome enabled 22 additional microsatellite markers to be identified and developed in a much faster, cost-effective manner using bioinformatic methods [50]. More importantly, each of these microsatellites were known to be in non-coding regions across all of the autosomes, providing a greater representation of neutral genome-wide diversity in comparison to the original 11 putatively neutral microsatellites. It has previously been estimated that the development of just 10 microsatellite markers without prior genetic data can cost up to $10000 [51]. The availability of a reference genome mitigates the need for traditional microsatellite isolation procedures, and therefore, significantly reduces costs associated with marker development (< $1000 for primer optimization and testing). Additionally, the commercial development of microsatellite-based PCR kits resulted in further reductions in the time and cost associated with microsatellite marker development and use [50]. To date 33 microsatellite markers have successfully been applied to Tasmanian devil conservation to investigate inbreeding [50], reconstruct the pedigree of offspring born in group housing and on Maria Island [50,52–54], and investigate mate choice within captivity and the wild [55] (Table 1). These microsatellite markers have also successfully been applied to genotype individuals using non-invasive scat samples [56], which are notoriously known for producing low quantities of low-quality DNA [57]. Globally, microsatellite markers continue to be an effective tool in conservation decision making by answering population questions [58–62]. They are particularly valuable when using non-invasive samples that are often unsuitable for more complex genomic methods that require high-quality input DNA, such as reduced representation sequencing and other whole-genome sequencing methods [15]. A reference genome allows for fast, easy, and inexpensive development of such markers, improving their utility in the conservation management space.

Table 1. Examples of Tasmanian devil conservation questions, actions, and outcomes that have been facilitated by the reference genome.

Reference Genome Use	Conservation Questions Addressed	Conservation Actions	Conservation Outcomes
• Microsatellite development • Genome-wide SNP analysis	• Were the founders related? • Does the metapopulation have equal founder representation to ensure the maintenance of gene diversity? • Is inbreeding accumulating in group housing and Maria island insurance populations?	• Resolved relatedness of founders [43] • Resolved parentage in group housing within the metapopulation [50,52,54] • Reconstructed pedigree of island population [53] • Informed translocation recommendations [63]	• Tool for selecting individuals for translocations based on genetic complementation • Improved maintenance of genetic diversity across captive populations • Increased genetic diversity of hybrid individuals at wild release sites
• The characterization of DFTD strains	• How many DFTD strains exist?	• Appropriate management of wild populations [46,64,65]	• Assisted in managing the spread of new DFTD strains
• The characterization of immune genes • Primer design and SNP panel development • Targeted SNP analysis	• Can we develop a vaccine for DFTD? • Can we improve Tasmanian devil immune diversity?	• Immunization development and deployment [66]. Immune gene diversity analysis for informed translocation recommendations [67–75]	• Improved immune responses of devils released to the wild • Improved immunogenetic diversity of released Tasmanian devils and their resultant offspring
• Development of blocking primer for metagenomics diet analysis	• What constitutes the complete diet of Tasmanian devils on Maria Island?	• Investigating the impact of an introduced carnivore to island wildlife	• Mitigation implemented to reduce the impact on highly consumed species
• Alignment of resequenced genomes • SNP Analysis and Annotation • GWAS	• Are devils evolving host-parasite resistance to DFTD?	• Ongoing monitoring to ensure releases do not impact the evolution of potential resistance alleles [76–79]	• Assisted in understanding regions of the genome that are potentially involved in DFTD resistance

3.1.2. Reduced Representation Sequencing

While microsatellite analysis is one of the most common population genetics tools, sometimes more statistical power is needed to address specific conservation management questions, particularly in species with low genetic diversity [43,80,81]. For instance, in the Tasmanian devil, microsatellite analysis was unable to accurately estimate the relatedness of founders sourced for the insurance population between 2006 and 2008 [43]. Single nucleotide polymorphisms (SNPs) enable greater resolution for addressing some common conservation issues such as resolving parentage and population structure, understanding genetic diversity, and identifying regions of the genome, which may be linked to important phenotypes [42]. When compared to a microsatellite approach, only 3–8 biallelic SNPs are required to be as informative as one microsatellite marker [82,83]. Reduced representation sequencing (RRS) is a simple, cost-effective approach for generating genome-wide SNP data and is gaining popularity in the conservation sector [15,42,84]. RRS relies on high-throughput sequencing of fragments generated by restriction enzyme digestion of the genome and can, therefore, easily be applied in any species. There are a variety of RRS methods currently available, including traditional RADseq [85], ddRAD [86], DArTseq [87], and others [42].

Both DArTseq and RADseq have been employed to collect RRS data from over 1,000 Tasmanian devils from the insurance population, Maria island and a number of wild sites [76,77,84,88,89]. RRS methods have shown to be superior in accurately estimating diversity and inferring genome-wide heterozygosity compared with microsatellite analysis and other targeted techniques [89]. Although this approach does not require a reference genome for development and use, coupling RRS data with a reference genome is advantageous in that it: i) improves the reliability of genotype calls [90]; ii) reduces the required coverage for accurate genotyping [91]; iii) provides for a greater number of SNPs [92]; iv) improves downstream population genetic inferences [92]; v) allows for SNP annotation with gene information [93]; and vi) provides the ability to compare results from differing RRS methods which are particularly important when different methods are used across time for endangered species.

Using a reference genome guided approach in the Tasmanian devil enabled 2060 SNPs to be identified [84] much more quickly than a de novo approach. Aligning the RRS data to the reference genome provides the ability to identify genes which may be targets of future analysis, and to separate functional vs. non-functional genome diversity which could have conservation implications [94]. For example, the reference genome was able to identify candidate genes within a genomic region that displayed signatures of selection in RRS data [76], and to identify cancer-resistance candidate genes from phenotype association tests of RRS data [77] (Table 1). A number of non-synonymous SNPs have also been identified within particular genes, which have the potential to impact phenotype. Furthermore, reference alignment allows SNPs from alternative RRS datasets to be compared and combined, such as the DArTseq and RADseq data, which are important for reusing previous investments of limited conservation dollars. Recent work investigating New Zealand threatened bird species also showed the benefits of calling SNPs against conordinal, confamilial, cogeneric, and conspecific reference genomes [95]. This highlights that not every threatened species requires a reference genome, although the quality of the SNP data reduces as you move away from the genus and family level.

3.2. Further Species-Specific Applications

3.2.1. Reference Gene Characterization

A valuable advantage of having access to a reference genome is the ability to characterize particular genes, or gene families, that are relevant to species-specific conservation [23]. Gene characterisation is often undertaken in two main ways: in-depth, manual characterization of a specific set of genes of interest, and automatic, whole-genome annotation. The latter is achieved in two main stages: the computational phase and the annotation phase [96,97]. During the computational phase, initial gene predictions are based on several lines of evidence including transcriptome and protein data from the species of interest and several closely-related, or well-annotated species [96,97]. During the annotation phase, the most representative gene predictions (defined by the annotation pipeline) are synthesized into the final gene annotations [96,97]. The whole-genome annotation of the Tasmanian devil reference genome was achieved using the Ensemble genome annotation pipeline [46,98,99]. This automatic annotation of 18775 protein-coding genes was critical to the development of targeted SNP panels to explore diversity at important immune genes in the Tasmanian devil [69–71] (see SNP Panel section below), and in the identification of genes that may be linked to DFTD [46,76–78,100] (Table 1).

While modern-day tools, such as trainable automated gene prediction algorithms, have increased the feasibility of genome annotation of newly sequenced species within individual research groups, complete genome annotation still requires considerable bioinformatics expertise [96,97]. Manual annotation of a subset of target genes is often required. This is particularly relevant for genes that have experienced duplications and are, therefore, often unable to be automatically annotated [23,96]. In the Tasmanian devil, this was true for a number of gene families, including the Major Histocompatibility Complex (MHC), toll-like receptors (TLR), natural killer (NK) receptors, cathelicidins, behavior, and reproductive genes which were all manually annotated [69,72,75,101,102]. Annotation of these genes was essential in facilitating species-specific downstream research and informing conservation management decisions in the Tasmanian devil, such as genetic variation analyses [69,70,72,75]; selection of individuals for release to the wild [63], individuals response to the immunotherapy [66]; changes of immune function with the onset of puberty [73]; and the influence of age and DFTD on immune function [74] (Table 1). This highlights the potential of a reference genome for exploratory analysis of gene families involved in key biological processes of threatened species such as immunity, reproduction, and behavior.

3.2.2. Targeted SNP Panels

Targeted SNP panels enable diversity at particular genes to be investigated based on current conservation concerns/questions [103]. In the Tasmanian devil, an SNP panel targeting immune, behavioral, and putatively neutral loci was developed and used to genotype over 300 individuals in the insurance population [71]. This involved low-coverage resequencing of a number of individuals (see the Whole-Genome Resequencing section below), alignment of data to the reference genome, identification of target SNPs, primer design, pilot sequencing, and final genotyping. The SNP panel resolved parentage with higher confidence than microsatellite markers and also provided representative measures of genetic diversity at both functional and non-functional loci [71]. Development of another SNP panel, which targeted a range of immune genes, showed considerably low immune diversity in the species [70], which has led to further research into ways of breeding Tasmanian devils to improve genome-wide heterozygosity and functional diversity [67,68]. The Tasmanian devil reference genome was essential for aligning sequencing data and target SNP discovery allowing for management decisions to be based on both genome-wide and functional diversity (Table 1). Although custom SNP panel development can be expensive and is not simple, once developed it provides fast, accurate measures of diversity at particular genes, or genome regions, across a large number of individuals [71,104,105].

3.2.3. Whole-Genome Resequencing

Whole-genome resequencing (WGR) involves sequencing the genome of several individuals to a predetermined level of coverage (usually between 2× and 60×) and aligning this data to an available reference genome (for examples in non-model species, see Fuentes-Pardo and Ruzzante [15]). A major application of whole-genome resequencing (WGR) is the identification of variation throughout the genome, enabling the development of more targeted approaches that can be used to explore diversity at key regions in a larger cohort of individuals [70,71]. The Tasmanian devil targeted SNP panels were created using low-coverage WGR (10–15×) data from 7–12 individuals aligned against the annotated reference genome [70,78]. A major limitation of using this low-coverage resequencing strategy is that genome regions with lower coverage can often contain sequencing errors that may not be distinguished from true SNPs [106]. This led to a number of the SNPs identified in the Tasmanian devil resequencing data not being present in the downstream SNP panel data [70,78]. While the best way to overcome this limitation is to increase the sequencing coverage of individuals, other methods, such as calling SNPs across individuals, can assist in more accurate variant calling in low-coverage WGR datasets [107].

Higher-coverage sequence data enables variants and heterozygosity to be called much more accurately than low-coverage sequence data and hence allows for SNPs to be called more confidently without additional targeted sequencing (e.g., SNP panels) [108]. High-coverage (~45×) WGR of 25 Tasmanian devils has allowed for reliable estimates of genome-wide heterozygosity, which are being used to assess the accuracy of estimates from other techniques including microsatellites, SNP panels and RRS data. The higher cost of high-coverage data causes a trade-off between investigating the whole genome of a relatively small number of individuals versus using a targeted subset of loci across many individuals (as of 2019, WGR routinely costs over $1000 per individual whereas RRS costs less than $100 per individual). This trade-off needs to be acknowledged, is dependent on the conservation research questions, and requires careful consideration prior to the commencement of sequencing [13]. Fortunately, a number of alternative cost-effective WGR approaches are available and may be suitable when high-coverage WGR is not possible. For a review of the different types of WGR and their different applications in conservation [15].

Whilst targeted sequencing approaches are useful for the exploration of genes known to be important to species biology, sometimes genetic mechanisms driving particular phenomena that are vital to species adaptation and survival may not be known or detected in other reduced sequencing techniques like RRS [109]. Whole-genome resequencing (WGR) enables conservation researchers to ask and answer a wide range of questions that are not possible using other approaches. For example, WGR also enables the use of genome-wide association studies to determine the genetic basis of particular phenotypic traits that are important to species conservation [13,15]. In the case of the Tasmanian devil, some individuals have been found to display a resistant phenotype to DFTD, enabling spontaneous tumor regression [110]. Identifying the potential genetic basis of this phenotype is important to understanding which individuals may be more resilient to the disease and provide targets for the development of potential treatments [76–78] (Table 1). Low-coverage WGR of individuals showing tumor regression and those that succumbed to the disease enabled a genome-wide association study to be undertaken, which identified two genomic regions that may be associated with resistance to DFTD including PAX3 and TLL1 loci [78]. A follow up study, Wright et al. [78] resequenced 10 individuals to a higher coverage (20–30×) and was able to identify a larger number of genomic regions that may underlie tumor regression in the Tasmanian devil [100]. This work demonstrates the ability of WGR data, along with an annotated reference genome, in exploring the genetic basis of phenotypic traits that could have important conservation implications [13,15,78,100] (Table 1). It is important to note that often larger numbers of individuals are required to identify genes underlying certain phenotypes, particularly in species with higher genetic diversity and reduced selective pressure on the phenotype of interest [111]. This requires careful consideration of trade-offs between the sequencing approach (targeted vs. RRS vs. WGR), number of samples and sequencing coverage, and will often depend upon some prior knowledge (or preliminary testing), budget, and access to samples. Overall, WGR data is better able to separate out and compare functional versus non-functional diversity than RRS methods, which is valuable in understanding the adaptive potential of species [94].

There are many other advantages of using this high-resolution genomic data,, including i) more robust insights into the evolutionary and demographic histories of a species; ii) more accurate measures of diversity, inbreeding and population structure; and iii) the ability to identify and investigate signatures of selection and adaptive genetic variation [15,16,18]. WGR data in the Tasmanian devil is currently being employed to assess selection and mutation rates within populations and in identifying runs of homozygosity (ROH) throughout the genome (for examples in other species, see Ceballos, et al. [112] and Hodgkinson, et al. [113]). These analyses are useful in the investigation of well-known issues in conservation, including inbreeding depression [112] and adaptation to captivity [114].

Some of the current limitations for using WGR in conservation contexts are the cost, the required computing power and respective expertise, and the availability of reference genomes [13,15]. Costs vary greatly and depend on the number of individuals or loci you wish to use, and the required depth of sequencing [15]. In addition, this approach requires significant expertise and compute power to execute, which limits its applicability to many conservation contexts [15]. Creating partnerships between academic researchers with the required expertise and computing resources and conservation managers is key to overcoming many limitations of using genomics in conservation, and has been successfully implemented in the conservation of the Tasmanian devil [35]. A reference genome is essential for WGR, so the significant lack of published genomes (<1%) for threatened species (or their closely-related counterparts) prevents many conservation managers from taking full advantage of high-resolution genomic data. However, in the dawn of large genomic consortia such as the Earth Biogenome Project, which aims to sequence the genomes of all of the Earth's eukaryotic biodiversity over the next 10 years [20], lack of a reference genome will soon become a thing of the past.

Overall, WGR paired with an annotated reference genome opens up a realm of possibilities for downstream conservation research by developing more cost-effective approaches when data from a large number of individuals is necessary for making informed conservation management decisions. As costs of sequencing continue to decrease, and the availability of reference genomes continue to rise, the use of this high-resolution genomic data in conservation research will likely become the norm [12] and is already being applied to some bird species [95].

4. Reference Genome Quality

An important factor to consider in the creation of reference genomes is the quality of the assembly. Consortia such as the Vertebrate Genome Project and the Earth Biogenome Project have proposed specific standards that reference genomes should meet [20,26] (Table S1). However, it is important to understand whether such high standards are necessary or achievable for conservation management. A number of statistics are used to evaluate the different aspects of genome quality including accuracy (e.g., average read coverage and quality), continuity (e.g., N50, N90, number of contigs/scaffolds, average length of contigs/scaffolds, gap percentage, etc.), and completeness (e.g., BUSCO (Benchmarking sets of Universal Single-Copy Orthologs)/CEGMA (Core Eukaryotic Genes Mapping Approach) scores, number of genes, etc.) (see Wajid and Serpedin [115] for a more exhaustive list). While the ideal reference genome would consist of a completely annotated, gap-free, chromosome-length assembly, even the some of the best model species genomes, such as the human genome, currently do not reach this standard. Furthermore, the ease and ability to reach chosen standards depends on many factors, including genome size, genome structure (e.g., repetitive content), level of heterozygosity, sample availability/quantity, as well as the cost and expertise of the sequencing types and computing resources available [24] (for reviews on reference genome creation including available sequencing types and their associated advantages/disadvantages see Ekblom and Wolf [96], Wajid and Serpedin [115], and Sedlazeck, et al. [116]). It is important to note that the current Tasmanian devil reference genome was sequenced in 2011 by Murchison et al. [46], so it does not meet the minimum standards set by the EBP (Earth Biogenome Project) or VGP (Vertebrate Genomes Project) (Table S1). Despite this, the Tasmanian devil genome has still been able to facilitate an enormous amount of conservation research. A higher-quality genome which is more complete, correct, and contiguous, has a number of advantages such as improved identification and characterization of genes and other genomic regions; more accurate ROH (runs of homozygosity) analysis and structural variant analysis; and higher resolution of chromosomal organization allowing for improved comparative genomic and evolutionary analyses [117].

Naturally, genome quality is also a factor of input DNA quality. High molecular weight DNA, generally greater than 40 kb in length, is required to generate the multiple sequencing types used to construct a high-quality genome [118]. Extracting high molecular weight DNA often requires additional consideration during the sample collection phase, such as flash-freezing tissues in liquid nitrogen, storage at −80 °C or below, and avoiding freeze-thaw. However, for species of high conservation concern, or those that inhabit difficult field locations, this could be challenging. In these scenarios, researchers may utilize museum specimens. However, this can introduce additional problems associated with sample preservation and degraded DNA, which may not be suited to long-read sequencing technologies [119]. As such, the ability to collect, store, and extract high-quality DNA should not be underestimated, as this is an essential first step towards generating high-quality genome. However, it is important to weigh up whether the cost, computing resources, expertise, and time of creating an improved or "Gold standard" assembly is necessary to answer the conservation research questions at hand. For example, Patton et al. [47] showed that the improvement of contiguity of the newly released 2019 Tasmanian devil assembly had minimal impacts on inferred patterns of historical effective population size when compared to the current reference assembly. Hence, in many cases, a simple short-read genome assembly is enough to answer many basic conservation management questions and also enable a number of more in-depth species-specific analyses mentioned in the sections above. Nevertheless, as sequencing technologies and computational infrastructure continue to advance and become more affordable, high-quality reference genomes would become easier to create and would overcome many of the limitations of currently fragmented reference assemblies such as incomplete gene characterization, comparative evolutionary limitations, and increased computational requirements [117]. Despite this, without advances in sequencing chemistry and library preparation to reduce input DNA quality and quantity, the availability of high-quality samples and ensuing high molecular weight DNA may continue to limit the creation of high-quality reference genomes in some species.

5. Conclusions

The Tasmanian devil reference genome has enhanced our capacity to manage this species in the face of an infectious, clonal cancer. By having the reference genome, we have been able to develop a range of genomic tools that have been used to investigate DFTD (e.g., [46]), investigate the interplay between the Tasmanian devils and the disease (e.g., [76–79]), inform development of immunotherapy and vaccine protocols [66], inform the management of the insurance population [38,65], and provide advice on the translocation of Tasmanian devils to wild populations to improve both genome-wide and functional diversity (e.g., [63,89]). Tasmanian devils are not the only species who are threatened globally by disease; other examples include black-footed ferret and distemper [120], bats and white-nose syndrome [121], and frogs and chytrid [122]. Here we have presented a strong case study of the benefits of using reference genomes for the conservation of threatened species. As the threat to global biodiversity increases, the management of threatened species becomes more pronounced. Reference genomes could be used by conservation managers to develop a range of genetic tools such as designing species-specific microsatellite markers for population data and differentiation; developing targeted SNP panels, or aligning and calling RRS data, for higher resolution population information or data on particular genes of interest; and conducting exploratory analyses (e.g., genome-wide association studies) using variant calling of whole-genome resequencing data.

Despite the challenges in obtaining high-quality samples for genome sequencing and expertise for the creation of reference genomes for threatened species, there is value in them. Reduced costs and lower input DNA requirements, as well as improved bioinformatic assembly and annotation pipelines based on non-model non-eutherian species, mean that these technologies are becoming more attainable by conservation programs and should be used more routinely where budgets allow [96]. Reference genomes enable a wealth of genetic/genomic applications and are an important asset in our ongoing fight to preserve global biodiversity. We would recommend that conservation managers who are seeking to use the types of methods we have described herein collaborate with global genome consortia (like the Earth Biogenome Project) or national/local consortia (like the Oz Mammal Genome Initiative) to utilize the full potential of genomic resources and join the genomics revolution. This allows conservation managers to focus on conservation and work with geneticists who can help them make adaptive management decisions in real-time [35].

Although here we have presented a unique case study of a species with significantly low levels of genetic diversity and a large threatening disease process, the techniques described for the Tasmanian devil can be applied more broadly to many species of conservation concern. The applications of what we have described herein for devils is not unique to this species as many of the questions we have answered are posed by those managing other threatened species. These include understanding historical demography and current population structure, minimizing inbreeding, maximizing adaptive potential, and identifying the basis of important phenotypic traits (whether these be related to disease, behavior, or reproduction). Hence, despite differences in threatening processes and current state of vulnerable species, the nature of their small population sizes will result in a number of common conservation concerns that could be informed using genomic data [15,18]. In the midst of the sixth mass extinction event, we advocate the use of reference genomes and associated genetic tools to arm conservation managers with ways to assist the long-term survival of species.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Table S1: Comparison of model and non-model mammalian/marsupial reference genomes to the G10K and EBP minimum reference genome quality standards.

Author Contributions: This project was conceived by K.B. and C.J.H. All authors contributed to writing and reviewing the manuscript. This work was funded by the Australian Research Council Grants to K.B. and C.J.H.

Funding: This research was funded by the Australian Research Council (LP140100508; DP180102465).

Acknowledgments: Thanks to all the staff, students, and affiliates of the Australasian Wildlife Genomics Group who have contributed to discussions over the years in regards to the benefit of genomics data to conservation, in particular, R. Johnson, C. Grueber, B. Wright and Y. Cheng. Thanks also to the staff of the Save the Tasmanian Devil Program and the members of the Zoo and Aquarium Association Australasia, without whom our devil work would not be possible.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ceballos, G.; Ehrlich, P.R.; Barnosky, A.D.; García, A.; Pringle, R.M.; Palmer, T.M. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Sci. Adv.* **2015**, *1*, e1400253, doi:10.1126/sciadv.1400253.
2. IUCN. The IUCN Red List of Threatened Species. Version 2019-1. Available online: http://www.iucnredlist.org (accessed on 2 July 2019).
3. Johnson, C.N.; Isaac, J.L. Body mass and extinction risk in Australian marsupials: The ‘Critical Weight Range’ revisited. *Austral. Ecol.* **2009**, *34*, 35–40, doi:10.1111/j.1442-9993.2008.01878.x.
4. Johnson, C.N.; Isaac, J.L.; Fisher, D.O. Rarity of a top predator triggers continent-wide collapse of mammal prey: Dingoes and marsupials in Australia. *Proc. R. Soc. Biol. Sci. Ser. B* **2006**, *274*, 341–346.
5. Short, J.; Smith, A. Mammal decline and recovery in Australia. *J. Mammal.* **1994**, *75*, 288–297, doi:10.2307/1382547.
6. Mittermeier, R.A. *Megadiversity: Earth’s Biologically Wealthiest Nations*; Agrupacion Sierra Madre: Mexico City, Mexico, 1997.
7. Chapman, A.D. *Numbers of Living Species in Australia and the World*, 2nd ed.; Australian Government, Department of the Environment and Energy: Canberra, Australia, 2009.
8. Department of the Environment and Energy. Recovery Plans. Available online: https://www.environment.gov.au/biodiversity/threatened/recovery-plans (accessed on 8/8/2019).
9. Ballou, J.D.; Lees, C.; Faust, L.J.; Long, S.; Lynch, C.; Bingaman Lackey, L.; Foose, T.J. Demographic and genetic management of captive populations. In *Wild Mammals in Captivity: Principles and Techniques for Zoo Management*, 2nd ed.; The University of Chicago Press: Chicago, IL, USA, 2010; pp. 219–252.
10. Lacy, R.C. Importance of Genetic Variation to the Viability of Mammalian Populations. *J. Mammal.* **1997**, *78*, 320–335, doi:10.2307/1382885.
11. Frankham, R.; Ballou, J.D.; Briscoe, D.A. *Introduction to Conservation Genetics*, 2 ed.; Cambridge University Press: Cambridge, UK, 2010; doi:10.1017/CBO9780511809002.
12. Johnson, W.E.; Koepfli, K. The role of genomics in conservation and reproductive sciences. In *Reproductive Sciences in Animal Conservation*; Springer: Berlin, Germany, 2014; pp. 71–96.
13. Supple, M.A.; Shapiro, B. Conservation of biodiversity in the genomics era. *Genome Biol.* **2018**, *19*, 131, doi:10.1186/s13059-018-1520-3.
14. Allendorf, F.W. Genetics and the conservation of natural populations: Allozymes to genomes. *Mol. Ecol.* **2017**, *26*, 420–430, doi:10.1111/mec.13948.
15. Fuentes-Pardo, A.P.; Ruzzante, D.E. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Mol. Ecol.* **2017**, *26*, 5369–5406, doi:10.1111/mec.14264.
16. Larsen, P.A.; Matocq, M.D. Emerging genomic applications in mammalian ecology, evolution, and conservation. *J. Mammal.* **2019**, *100*, 786–801, doi:10.1093/jmammal/gyy184.

17. McMahon, B.J.; Teeling, E.C.; Höglund, J. How and why should we implement genomics into conservation? *Evol. Appl.* **2014**, *7*, 999–1007.
18. Khan, S.; Nabi, G.; Ullah, M.W.; Yousaf, M.; Manan, S.; Siddique, R.; Hou, H. Overview on the role of advance genomics in conservation biology of endangered species. *Int. J. Genomics* **2016**, *2016*, 1–8, doi:10.1155/2016/3460416.
19. Kitts, P.A.; Church, D.M.; Thibaud-Nissen, F.; Choi, J.; Hem, V.; Sapojnikov, V.; Smith, R.G.; Tatusova, T.; Xiang, C.; Zherikov, A. Assembly: A resource for assembled genomes at NCBI. *Nucleic Acids Res.* **2015**, *44*, D73–D80.
20. Lewin, H.A.; Robinson, G.E.; Kress, W.J.; Baker, W.J.; Coddington, J.; Crandall, K.A.; Durbin, R.; Edwards, S.V.; Forest, F.; Gilbert, M.T.P. Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4325–4333, doi:10.1073/pnas.1720115115.
21. Li, Q.; Xuan, Z.; Li, Y.; Zheng, H.; Bai, Y.; Li, B.; Hu, Y.; Liu, X.; Zhang, Z.; Li, D.; et al. The sequence and de novo assembly of the giant panda genome. *Nature* **2010**, *463*, 311–317, doi:10.1038/nature08696.
22. Groenen, M.A.; Archibald, A.L.; Uenishi, H.; Tuggle, C.K.; Takeuchi, Y.; Rothschild, M.F.; Rogel-Gaillard, C.; Park, C.; Milan, D.; Megens, H.-J. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **2012**, *491*, 393–398, doi:10.1038/nature11622.
23. Johnson, R.N.; O'Meally, D.; Chen, Z.; Etherington, G.J.; Ho, S.Y.W.; Nash, W.J.; Grueber, C.E.; Cheng, Y.; Whittington, C.M.; Dennison, S.; et al. Adaptation and conservation insights from the koala genome. *Nat. Genet.* **2018**, *50*, 1102–1111, doi:10.1038/s41588-018-0153-5.
24. Koepfli, K.-P.; Paten, B.; Genome 10K Community of Scientists; O'Brien, S.J. The Genome 10K Project: A way forward. *Annu. Rev. Anim. Biosci.* **2015**, *3*, 57–111, doi:10.1146/annurev-animal-090414-014900.
25. Genome 10K Community of Scientists. Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J. Hered.* **2009**, *100*, 659–674.
26. Genome 10K Community of Scientists. Vertebrate Genomes Project. Available online: https://vertebrategenomesproject.org (accessed on 16 August 2019).
27. China National GeneBank. B10K. Available online: https://b10k.genomics.cn/ (accessed on 16 August 2019).
28. Teeling, E.C.; Vernes, S.C.; Dávalos, L.M.; Ray, D.A.; Gilbert, M.T.P.; Myers, E.; Bat1K Consortium. Bat biology, genomes, and the Bat1K project: To generate chromosome-level genomes for all living bat species. *Annu. Rev. Anim. Biosci.* **2018**, *6*, 23–46, doi:10.1146/annurev-animal-022516-022811.
29. GIGA Community of Scientists. The Global Invertebrate Genomics Alliance (GIGA): Developing community resources to study diverse invertebrate genomes. *J. Hered.* **2013**, *105*, 1–18.
30. Voolstra, C.R.; Wörheide, G.; Lopez, J.V. Corrigendum to: Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA). *Invertebr. Syst.* **2017**, *31*, 231–231.
31. Potter, S.; Eldridge, M. Oz Mammal Genomics. *Australas. Sci.* 2017, *38*, 19–21.
32. Ralls, K.; Ballou, J.D.; Dudash, M.R.; Eldridge, M.D.; Fenster, C.B.; Lacy, R.C.; Sunnucks, P.; Frankham, R. Call for a paradigm shift in the genetic management of fragmented populations. *Conserv. Lett.* **2018**, *11*, e12412, doi:10.1111/conl.12412.
33. Taylor, H.R.; Dussex, N.; van Heezik, Y. Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Glob. Ecol. Conserv.* **2017**, *10*, 231–242, doi:10.1016/j.gecco.2017.04.001.
34. Holderegger, R.; Balkenhol, N.; Bolliger, J.; Engler, J.O.; Gugerli, F.; Hochkirch, A.; Nowak, C.; Segelbacher, G.; Widmer, A.; Zachos, F.E. Conservation genetics: Linking science with practice. *Mol. Ecol.* **2019**, *28*, 3848–3856, doi:10.1111/mec.15202.
35. Hogg, C.J.; Grueber, C.E.; Pemberton, D.; Fox, S.; Lee, A.V.; Ivy, J.A.; Belov, K. "Devil Tools & Tech": A Synergy of Conservation Research and Management Practice. *Conserv. Lett.* **2017**, *10*, 133–138.
36. Grueber, C.E. Comparative genomics for biodiversity conservation. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 370–375, doi:10.1016/j.csbj.2015.05.003.
37. Lazenby, B.T.; Tobler, M.W.; Brown, W.E.; Hawkins, C.E.; Hocking, G.J.; Hume, F.; Huxtable, S.; Iles, P.; Jones, M.E.; Lawrence, C. Density trends and demographic signals uncover the long-term impact of transmissible cancer in Tasmanian devils. *J. Appl. Ecol.* **2018**, *55*, 1368–1379, doi:10.1111/1365-2664.13088.
38. Hogg, C.J.; Fox, S.; Pemberton, D.; Belov, K. *Saving the Tasmanian Devil*; Hogg, C.J., Fox, S., Pemberton, D., Belov, K., Eds.; CSIRO Publishing: Clayton South, VIC, Australia, 2019.
39. Jones, M.E.; Paetkau, D.; Geffen, E.; Moritz, C. Microsatellites for the Tasmanian devil (Sarcophilus laniarius). *Mol. Ecol. Notes* **2003**, *3*, 277–279, doi:10.1046/j.1471-8286.2003.00425.x.

40. Siddle, H.V.; Marzec, J.; Cheng, Y.; Jones, M.; Belov, K. MHC gene copy number variation in Tasmanian devils: Implications for the spread of a contagious cancer. *Proc. R. Soc. Biol. Sci. Ser. B* **2010**, *277*, 2001–2006, doi:10.1098/rspb.2009.2362.
41. Cheng, Y.; Belov, K. Isolation and characterisation of 11 MHC-linked microsatellite loci in the Tasmanian devil (*Sarcophilus harrisii*). *Conserv. Genet. Resour.* **2012**, *4*, 463–465, doi:10.1007/s12686-011-9575-4.
42. Andrews, K.R.; Good, J.M.; Miller, M.R.; Luikart, G.; Hohenlohe, P.A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **2016**, *17*, 81, doi:10.1038/nrg.2015.28.
43. Hogg, C.J.; Ivy, J.A.; Srb, C.; Hockley, J.; Lees, C.; Hibbard, C.; Jones, M. Influence of genetic provenance and birth origin on productivity of the Tasmanian devil insurance population. *Conserv. Genet.* **2015**, *16*, 1465–1473, doi:10.1007/s10592-015-0754-9.
44. Hogg, C.J.; Wright, B.; Morris, K.M.; Lee, A.V.; Ivy, J.A.; Grueber, C.E.; Belov, K. Founder relationships and conservation management: Empirical kinships reveal the effect on breeding programmes when founders are assumed to be unrelated. *Anim. Conserv.* **2019**, *22*, 348–361, doi:10.1111/acv.12463.
45. Miller, W.; Hayes, V.M.; Ratan, A.; Petersen, D.C.; Wittekindt, N.E.; Miller, J.; Walenz, B.; Knight, J.; Qi, J.; Zhao, F.; et al. Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil). *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12348–12353, doi:10.1073/pnas.1102838108.
46. Murchison, E.P.; Schulz-Trieglaff, O.B.; Ning, Z.; Alexandrov, L.B.; Bauer, M.J.; Fu, B.; Hims, M.; Ding, Z.; Ivakhno, S.; Stewart, C. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* **2012**, *148*, 780–791, doi:10.1016/j.cell.2011.11.065.
47. Patton, A.H.; Margres, M.J.; Stahlke, A.R.; Hendricks, S.; Lewallen, K.; Hamede, R.K.; Ruiz-Aravena, M.; Ryder, O.; McCallum, H.I.; Jones, M.E.; et al. Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in Tasmanian Devils. *Mol. Biol. Evol.* **2019**, msz191, doi:10.1093/molbev/msz191.
48. Selkoe, K.A.; Toonen, R.J. Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. *Ecol. Lett.* **2006**, *9*, 615–629, doi:10.1111/j.1461-0248.2006.00889.x.
49. Abdul-Muneer, P.M. Application of microsatellite markers in conservation genetics and fisheries management: Recent advances in population structure analysis and conservation strategies. *Genet. Res. Int.* **2014**, *2014*, 1–11, doi:10.1155/2014/691759.
50. Gooley, R.; Hogg, C.J.; Belov, K.; Grueber, C.E. No evidence of inbreeding depression in a Tasmanian devil insurance population despite significant variation in inbreeding. *Sci. Rep.* **2017**, *7*, 1830, doi:10.1038/s41598-017-02000-y.
51. Abdelkrim, J.; Robertson, B.C.; Stanton, J.-A.L.; Gemmell, N.J. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques* **2009**, *46*, 185–192, doi:10.2144/000113084.
52. Gooley, R.M.; Hogg, C.J.; Belov, K.; Grueber, C.E. The effects of group versus intensive housing on the retention of genetic diversity in insurance populations. *BMC Zool.* **2018**, *3*, 2, doi:10.1186/s40850-017-0026-x.
53. McLennan, E.A.; Gooley, R.M.; Wise, P.; Belov, K.; Hogg, C.J.; Grueber, C.E. Pedigree reconstruction using molecular data reveals an early warning sign of gene diversity loss in an island population of Tasmanian devils (Sarcophilus harrisii). *Conserv. Genet.* **2018**, *19*, 439–450, doi:10.1007/s10592-017-1017-8.
54. Farquharson, K.A.; Hogg, C.J.; Grueber, C.E. A case for genetic parentage assignment in captive group housing. *Conserv. Genet.* **2019**, *20*, 1–7, doi:10.1007/s10592-019-01198-w.
55. Day, J.; Gooley, R.M.; Hogg, C.J.; Belov, K.; Whittington, C.M.; Grueber, C.E. MHC-associated mate choice under competitive conditions in captive versus wild Tasmanian devils. *Behav. Ecol.* **2019**, *30*, 1196–1204.
56. Grueber, C.E.; Chong, R.; Gooley, R.M.; McLennan, E.A.; Barrs, V.R.; Belov, K.; Hogg, C.J. Genetic analysis of scat samples to inform conservation of Tasmanian devil. *Aust. Zool.* (In press).
57. Taberlet, P.; Waits, L.P.; Luikart, G. Noninvasive genetic sampling: Look before you leap. *Trends Ecol. Evol.* **1999**, *14*, 323–327, doi:10.1016/S0169-5347(99)01637-7.
58. Armstrong, A.J.; Dudgeon, C.L.; Bustamante, C.; Bennett, M.B.; Ovenden, J.R. Development and characterization of 17 polymorphic microsatellite markers for the reef manta ray (Mobula alfredi). *BMC Res. Notes* **2019**, *12*, 233, doi:10.1186/s13104-019-4270-8.
59. Faria, J.; Pita, A.; Rivas, M.; Martins, G.M.; Hawkins, S.J.; Ribeiro, P.; Neto, A.I.; Presa, P. A multiplex microsatellite tool for conservation genetics of the endemic limpet Patella candei in the Macaronesian archipelagos. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **2016**, *26*, 775–781, doi:10.1002/aqc.2651.

60. Shaney, K.J.; Adams, R.; Kurniawan, N.; Hamidy, A.; Smith, E.N.; Castoe, T.A. A suite of potentially amplifiable microsatellite loci for ten reptiles of conservation concern from Africa and Asia. *Conserv. Genet. Resour.* **2016**, *8*, 307–311, doi:10.1007/s12686-016-0534-y.
61. Storfer, A.; Epstein, B.; Jones, M.; Micheletti, S.; Spear, S.F.; Lachish, S.; Fox, S. Landscape genetics of the Tasmanian devil: Implications for spread of an infectious cancer. *Conserv. Genet.* **2017**, *18*, 1287–1297, doi:10.1007/s10592-017-0980-4.
62. Grueber, C.E.; Fox, S.; McLennan, E.A.; Gooley, R.M.; Weiser, E.L.; Pemberton, D.; Hogg, C.J.; Belov, K. Complex problems need detailed solutions: Harnessing multiple data types to inform genetic rescue in the wild. *Evol. Appl.* **2019**, *12*, 280–291.
63. Hogg, C.J.; McLennan, E.A.; Wise, P.; Lee, A.; Pemberton, D.; Fox, S.; Belov, K.; Grueber, C.E. Preserving the integrity of a single source population during multiple translocations. *Biol. Conserv.* (In press)
64. Pye, R.J.; Pemberton, D.; Tovar, C.; Tubio, J.M.; Dun, K.A.; Fox, S.; Darby, J.; Hayes, D.; Knowles, G.W.; Kreiss, A. A second transmissible cancer in Tasmanian devils. *Proc. Natl. Acad. Sci.* **2016**, *113*, 374–379, doi:10.1073/pnas.1519691113.
65. Hogg, C.; Lee, A.; Srb, C.; Hibbard, C. Metapopulation management of an endangered species with limited genetic diversity in the presence of disease: The Tasmanian devil Sarcophilus harrisii. *Int. Zoo Yearb.* **2017**, *51*, 137–153.
66. Pye, R.; Patchett, A.; McLennan, E.; Thomson, R.; Carver, S.; Fox, S.; Pemberton, D.; Kreiss, A.; Baz Morelli, A.; Silva, A. Immunization strategies producing a humoral IgG immune response against devil facial tumor disease in the majority of Tasmanian devils destined for wild release. *Front. Immunol.* **2018**, *9*, 259, doi:10.3389/fimmu.2018.00259.
67. Grueber, C.; Peel, E.; Wright, B.; Hogg, C.; Belov, K. A Tasmanian devil breeding program to support wild recovery. *Reprod. Fertil. Dev.* **2019**, *31*, 1296–1304, doi:10.1071/Rd18152.
68. McLennan, E.A.; Grueber, C.E.; Wise, P.; Belov, K.; Hogg, C.J. Mixing genetic lineages sucessfully boosts diversity of an endangered carnivore. *Anim. Conserv.* (under review)
69. Morris, K.M.; Cheng, Y.; Warren, W.; Papenfuss, A.T.; Belov, K. Identification and analysis of divergent immune gene families within the Tasmanian devil genome. *BMC Genomics* **2015**, *16*, 1017, doi:10.1186/s12864-015-2206-9.
70. Morris, K.M.; Wright, B.; Grueber, C.E.; Hogg, C.; Belov, K. Lack of genetic diversity across diverse immune genes in an endangered mammal, the Tasmanian devil (Sarcophilus harrisii). *Mol. Ecol.* **2015**, *24*, 3860–3872, doi:10.1111/mec.13291.
71. Wright, B.; Morris, K.; Grueber, C.E.; Willet, C.E.; Gooley, R.; Hogg, C.J.; O'Meally, D.; Hamede, R.; Jones, M.; Wade, C. Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population. *BMC Genomics* **2015**, *16*, 791, doi:10.1186/s12864-015-2020-4.
72. Cheng, Y.; Belov, K. Characterisation of non-classical MHC class I genes in the Tasmanian devil (Sarcophilus harrisii). *Immunogenetics* **2014**, *66*, 727–735, doi:10.1007/s00251-014-0804-3.
73. Cheng, Y.; Heasman, K.; Peck, S.; Peel, E.; Gooley, R.M.; Papenfuss, A.T.; Hogg, C.J.; Belov, K. Significant decline in anticancer immune capacity during puberty in the Tasmanian devil. *Sci. Rep.* **2017**, *7*, 44716, doi:10.1038/srep44716.
74. Cheng, Y.; Makara, M.; Peel, E.; Fox, S.; Papenfuss, A.T.; Belov, K. Tasmanian devils with contagious cancer exhibit a constricted T-cell repertoire diversity. *Commun. Biol.* **2019**, *2*, 99, doi:10.1038/s42003-019-0342-5.
75. Cui, J.; Cheng, Y.; Belov, K. Diversity in the Toll-like receptor genes of the Tasmanian devil (Sarcophilus harrisii). *Immunogenetics* **2015**, *67*, 195–201, doi:10.1007/s00251-014-0823-0.
76. Epstein, B.; Jones, M.; Hamede, R.; Hendricks, S.; McCallum, H.; Murchison, E.P.; Schönfeld, B.; Wiench, C.; Hohenlohe, P.; Storfer, A. Rapid evolutionary response to a transmissible cancer in Tasmanian devils. *Nat. Commun.* **2016**, *7*, 12684, doi:10.1038/ncomms12684.
77. Margres, M.J.; Jones, M.E.; Epstein, B.; Kerlin, D.H.; Comte, S.; Fox, S.; Fraik, A.K.; Hendricks, S.A.; Huxtable, S.; Lachish, S. Large-effect loci affect survival in Tasmanian devils (Sarcophilus harrisii) infected with a transmissible cancer. *Mol. Ecol.* **2018**, *27*, 4189–4199, doi:10.1111/mec.14853.
78. Wright, B.; Willet, C.E.; Hamede, R.; Jones, M.; Belov, K.; Wade, C.M. Variants in the host genome may inhibit tumour growth in devil facial tumours: Evidence from genome-wide association. *Sci. Rep.* **2017**, *7*, 423.

79. Hohenlohe, P.A.; McCallum, H.I.; Jones, M.E.; Lawrance, M.F.; Hamede, R.K.; Storfer, A. Conserving adaptive potential: Lessons from Tasmanian devils and their transmissible cancer. *Conserv. Genet.* **2019**, *20*, 81–87, doi:10.1007/s10592-019-01157-5.
80. Fernández, M.E.; Goszczynski, D.E.; Lirón, J.P.; Villegas-Castagnasso, E.E.; Carino, M.H.; Ripoli, M.V.; Rogberg-Muñoz, A.; Posik, D.M.; Peral-García, P.; Giovambattista, G. Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. *Genet. Mol. Biol.* **2013**, *36*, 185–191, doi:10.1590/S1415-47572013000200008.
81. Tokarska, M.; Marshall, T.; Kowalczyk, R.; Wójcik, J.; Pertoldi, C.; Kristensen, T.; Loeschcke, V.; Gregersen, V.; Bendixen, C. Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: The case of European bison. *Heredity* **2009**, *103*, 326, doi:10.1038/hdy.2009.73.
82. Rosenberg, N.A.; Li, L.M.; Ward, R.; Pritchard, J.K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **2003**, *73*, 1402–1422, doi:10.1086/380416.
83. Schopen, G.; Bovenhuis, H.; Visker, M.; Van Arendonk, J. Comparison of information content for microsatellites and SNPs in poultry and cattle. *Anim. Genet.* **2008**, *39*, 451–453, doi:10.1111/j.1365-2052.2008.01736.x.
84. Wright, B.; Farquharson, K.A.; McLennan, E.A.; Belov, K.; Hogg, C.J.; Grueber, C.E. From reference genomes to population genomics: Comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. *BMC Genomics* **2019**, *20*, 453, doi:10.1186/s12864-019-5806-y.
85. Davey, J.W.; Blaxter, M.L. RADSeq: Next-generation population genetics. *Briefings Funct. Genomics* **2010**, *9*, 416–423, doi:10.1093/bfgp/elq031.
86. Peterson, B.K.; Weber, J.N.; Kay, E.H.; Fisher, H.S.; Hoekstra, H.E. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS ONE* **2012**, *7*, e37135, doi:10.1371/journal.pone.0037135.
87. Von Mark, V.C.; Kilian, A.; Dierig, D.A. Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species. *PLoS ONE* **2013**, *8*, e64062.
88. Hendricks, S.; Epstein, B.; Schönfeld, B.; Wiench, C.; Hamede, R.; Jones, M.; Storfer, A.; Hohenlohe, P. Conservation implications of limited genetic diversity and population structure in Tasmanian devils (Sarcophilus harrisii). *Conserv. Genet.* **2017**, *18*, 977–982, doi:10.1007/s10592-017-0939-5.
89. McLennan, E.A.; Wright, B.R.; Belov, K.; Hogg, C.J.; Grueber, C.E. Too much of a good thing? Finding the most informative genetic data set to answer conservation questions. *Mol. Ecol. Resour.* **2019**, *19*, 659–671.
90. Torkamaneh, D.; Laroche, J.; Belzile, F. Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. *PLoS ONE* **2016**, *11*, e0161333, doi:10.1371/journal.pone.0161333.
91. Davey, J.W.; Hohenlohe, P.A.; Etter, P.D.; Boone, J.Q.; Catchen, J.M.; Blaxter, M.L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **2011**, *12*, 499, doi:10.1038/nrg3012.
92. Shafer, A.B.; Peart, C.R.; Tusso, S.; Maayan, I.; Brelsford, A.; Wheat, C.W.; Wolf, J.B. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods Ecol. Evol.* **2017**, *8*, 907–917.
93. Gurgul, A.; Miksza-Cybulska, A.; Szmatoła, T.; Jasielczuk, I.; Piestrzyńska-Kajtoch, A.; Fornal, A.; Semik-Gurgul, E.; Bugno-Poniewierska, M. Genotyping-by-sequencing performance in selected livestock species. *Genomics* **2019**, *111*, 186–195, doi:10.1016/j.ygeno.2018.02.002.
94. Hoelzel, A.R.; Bruford, M.W.; Fleischer, R.C. *Conservation of Adaptive Potential and Functional Diversity*; Springer: Berlin, Germany, 2019.
95. Galla, S.J.; Forsdick, N.J.; Brown, L.; Hoeppner, M.; Knapp, M.; Maloney, R.F.; Moraga, R.; Santure, A.W.; Steeves, T.E. Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management. *Genes* **2019**, *10*, 9, doi:10.3390/genes10010009.
96. Ekblom, R.; Wolf, J.B. A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.* **2014**, *7*, 1026–1042, doi:10.1111/eva.12178.
97. Yandell, M.; Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **2012**, *13*, 329, doi:10.1038/nrg3174.

98. Curwen, V.; Eyras, E.; Andrews, T.D.; Clarke, L.; Mongin, E.; Searle, S.M.; Clamp, M. The Ensembl automatic gene annotation system. *Genome Res.* **2004**, *14*, 942–950, doi:10.1101/gr.1858004.
99. Potter, S.C.; Clarke, L.; Curwen, V.; Keenan, S.; Mongin, E.; Searle, S.M.; Stabenau, A.; Storey, R.; Clamp, M. The Ensembl analysis pipeline. *Genome Res.* **2004**, *14*, 934–941, doi:10.1101/gr.1859804.
100. Margres, M.J.; Ruiz-Aravena, M.; Hamede, R.; Jones, M.E.; Lawrance, M.F.; Hendricks, S.A.; Patton, A.; Davis, B.W.; Ostrander, E.A.; McCallum, H. The genomic basis of tumor regression in Tasmanian devils (Sarcophilus harrisii). *Genome Biol. Evol.* **2018**, *10*, 3012–3025, doi:10.1093/gbe/evy229.
101. Peel, E.; Cheng, Y.; Djordjevic, J.; Fox, S.; Sorrell, T.; Belov, K. Cathelicidins in the Tasmanian devil (Sarcophilus harrisii). *Sci. Rep.* **2016**, *6*, 35019, doi:10.1038/srep35019.
102. van der Kraan, L.E.; Wong, E.S.; Lo, N.; Ujvari, B.; Belov, K. Identification of natural killer cell receptor genes in the genome of the marsupial Tasmanian devil (Sarcophilus harrisii). *Immunogenetics* **2013**, *65*, 25–35, doi:10.1007/s00251-012-0643-z.
103. van Tienderen, P.H.; de Haan, A.A.; van der Linden, C.G.; Vosman, B. Biodiversity assessment using markers for ecologically important traits. *Trends Ecol. Evol.* **2002**, *17*, 577–582.
104. Russell, T.; Cullingham, C.; Kommadath, A.; Stothard, P.; Herbst, A.; Coltman, D. Development of a novel mule deer genomic assembly and species-diagnostic SNP panel for assessing introgression in mule deer, white-tailed deer, and their interspecific hybrids. *G3 Genes Genomes Genet.* **2019**, *9*, 911–919, doi:10.1534/g3.118.200838.
105. Zhao, H.; Fuller, A.; Thongda, W.; Mohammed, H.; Abernathy, J.; Beck, B.; Peatman, E. SNP panel development for genetic management of wild and domesticated white bass (Morone chrysops). *Anim. Genet.* **2019**, *50*, 92–96, doi:10.1111/age.12747.
106. Li, R.; Li, Y.; Fang, X.; Yang, H.; Wang, J.; Kristiansen, K.; Wang, J. SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **2009**, *19*, 1124–1132, doi:10.1101/gr.088013.108.
107. Cheng, A.Y.; Teo, Y.-Y.; Ong, R.T.-H. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics* **2014**, *30*, 1707–1713, doi:10.1093/bioinformatics/btu067.
108. Kishikawa, T.; Momozawa, Y.; Ozeki, T.; Mushiroda, T.; Inohara, H.; Kamatani, Y.; Kubo, M.; Okada, Y. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci. Rep.* **2019**, *9*, 1784, doi:10.1038/s41598-018-38346-0.
109. Hoban, S.; Kelley, J.L.; Lotterhos, K.E.; Antolin, M.F.; Bradburd, G.; Lowry, D.B.; Poss, M.L.; Reed, L.K.; Storfer, A.; Whitlock, M.C. Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *Am. Nat.* **2016**, *188*, 379–397, doi:10.1086/688018.
110. Pye, R.; Hamede, R.; Siddle, H.V.; Caldwell, A.; Knowles, G.W.; Swift, K.; Kreiss, A.; Jones, M.E.; Lyons, A.B.; Woods, G.M. Demonstration of immune responses against devil facial tumour disease in wild Tasmanian devils. *Biol. Lett.* **2016**, *12*, 20160553, doi:10.1098/rsbl.2016.0553.
111. Hong, E.P.; Park, J.W. Sample size and statistical power calculation in genetic association studies. *Genomics Inform.* **2012**, *10*, 117, doi:10.5808/GI.2012.10.2.117.
112. Ceballos, F.C.; Hazelhurst, S.; Ramsay, M. Assessing runs of homozygosity: A comparison of SNP array and whole genome sequence low coverage data. *BMC Genomics* **2018**, *19*, 106, doi:10.1186/s12864-018-4489-0.
113. Hodgkinson, A.; Casals, F.; Idaghdour, Y.; Grenier, J.-C.; Hernandez, R.D.; Awadalla, P. Selective constraint, background selection, and mutation accumulation variability within and between human populations. *BMC Genomics* **2013**, *14*, 495, doi:10.1186/1471-2164-14-495.
114. Willoughby, J.R.; Ivy, J.A.; Lacy, R.C.; Doyle, J.M.; DeWoody, J.A. Inbreeding and selection shape genomic diversity in captive populations: Implications for the conservation of endangered species. *PloS ONE* **2017**, *12*, e0175996, doi:10.1371/journal.pone.0175996.
115. Wajid, B.; Serpedin, E. Do it yourself guide to genome assembly. *Brief. Funct. Genomics* **2014**, *15*, 1–9.
116. Sedlazeck, F.J.; Lee, H.; Darby, C.A.; Schatz, M.C. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **2018**, *19*, 329–346, doi:10.1038/s41576-018-0003-4.
117. Lee, H.; Gurtowski, J.; Yoo, S.; Nattestad, M.; Marcus, S.; Goodwin, S.; McCombie, W.R.; Schatz, M. Third-generation sequencing and the future of genomics. *BioRxiv* **2016**, 048603.
118. Rhoads, A.; Au, K.F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinform.* **2015**, *13*, 278–289, doi:10.1016/j.gpb.2015.08.002.

119. McDonough, M.M.; Parker, L.D.; Rotzel McInerney, N.; Campana, M.G.; Maldonado, J.E. Performance of commonly requested destructive museum samples for mammalian genomic studies. *J. Mammal.* **2018**, *99*, 789–802, doi:10.1093/jmammal/gyy080.
120. Thorne, E.T.; Williams, E.S. Disease and endangered species: The black-footed ferret as a recent example. *Conserv. Biol.* **1988**, *2*, 66–74.
121. Blehert, D.S.; Hicks, A.C.; Behr, M.; Meteyer, C.U.; Berlowski-Zier, B.M.; Buckles, E.L.; Coleman, J.T.; Darling, S.R.; Gargas, A.; Niver, R. Bat white-nose syndrome: An emerging fungal pathogen? *Science* **2009**, *323*, 227–227, doi:10.1126/science.1163874.
122. Berger, L.; Speare, R.; Hyatt, A. Chytrid fungi and amphibian declines: Overview, implications and future directions. In *Declines and disappearances of Australian frogs*; Campbell, A. Ed; Environment Australia: Canberra, Australia 1999, 23–33.

A1.2 TEN SIMPLE RULES FOR GETTING STARTED WITH COMMAND-LINE BIOINFORMATICS

The PDF version of the editorial titled “Ten simple rules for getting started with command-line bioinformatics” published in *PLOS Computational Biology* (2020; 17(2), e1008645), which comprises Chapter 2 of this thesis, is presented on the following pages.

EDITORIAL

Ten simple rules for getting started with command-line bioinformatics

Parice A. Brandies, **Carolyn J. Hogg***

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

* carolyn.hogg@sydney.edu.au

OPEN ACCESS

Citation: Brandies PA, Hogg CJ (2021) Ten simple rules for getting started with command-line bioinformatics. PLoS Comput Biol 17(2): e1008645. https://doi.org/10.1371/journal.pcbi.1008645

Editor: Scott Markel, Dassault Systemes BIOVIA, UNITED STATES

Published: February 18, 2021

Funding: Computing infrastructure support was provided through ch81 of the Australian National Computational Merit Scheme. Cloud compute was provided through in-kind support from Amazon Web Services, Intel, RONIN and Microsoft Azure. The Australasian Wildlife Genomics Group is supported by funding from the Australian Research Council (CJH) and the University of Sydney (CJH, PB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Sequencing technologies are becoming more advanced and affordable than ever before. In response, growing international consortia such as the Earth BioGenomes Project (EBP) [1], the Genome 10K project (G10K) [2,3], the Global Invertebrate Genomics Alliance (GIGA) [4,5], the Insect 5K project (i5K) [6,7], the 10,000 plants project (10KP) [8], and many others have big plans to sequence all life on earth. These consortia aim to utilise genomic data to uncover the biological secrets of our planet's biodiversity and apply this knowledge to real-world matters, such as improving our understanding of species' evolution, assisting with conservation of threatened species, and identifying new targets for medical, agricultural, or industrial purposes [1]. All of these goals rely on someone to analyse and make sense of the tremendous amounts of biological data, making bioinformaticians more sought-after than ever. Many researchers with a background in biology and genetics are stepping up to the challenge of big data analysis, but it can be a little daunting to start down the path of bioinformatics, particularly using the command line, without a strong background in computing and/or computer science. A recent "Ten simple rules" article highlighted the importance of bioinformatics research support [9]. Here we provide 10 simple rules for anyone interested in taking the leap into the realm of bioinformatics using the command line. We have put together these 10 simple rules for those starting on their bioinformatics journey, whether you be a student, an experienced biologist or geneticist, or anyone else who may be interested in this emerging field. The rules are presented in chronological order, together encompassing a simple 10-step process for getting started with command-line bioinformatics (Fig 1). This is by no means an exhaustive introduction to bioinformatics, but rather a simple guide to the key components to get you started on your way to unlocking the true potential of biological big data.

Rule 1: Get familiar with computer terminology

The first step in your command-line bioinformatics journey can be overwhelming due to the wealth of new terminology. This is where you need to channel your inner computer geek and learn the new language of computer terminology. In fact, this very paper is riddled with it, so our first rule addresses this tricky obstacle. Having a basic understanding of computing and associated terminology can be really useful in determining how to run your bioinformatics pipelines effectively. It can also help you troubleshoot many errors along the way. Understanding the terminology allows you to talk with your institutional information technology (IT) departments and communicate your computational needs to answer your biological questions. This will allow you to be able to source the resources you will need. A number of basic definitions of the main terms that you will likely come across as you enter the world of bioinformatics is presented in Box 1.

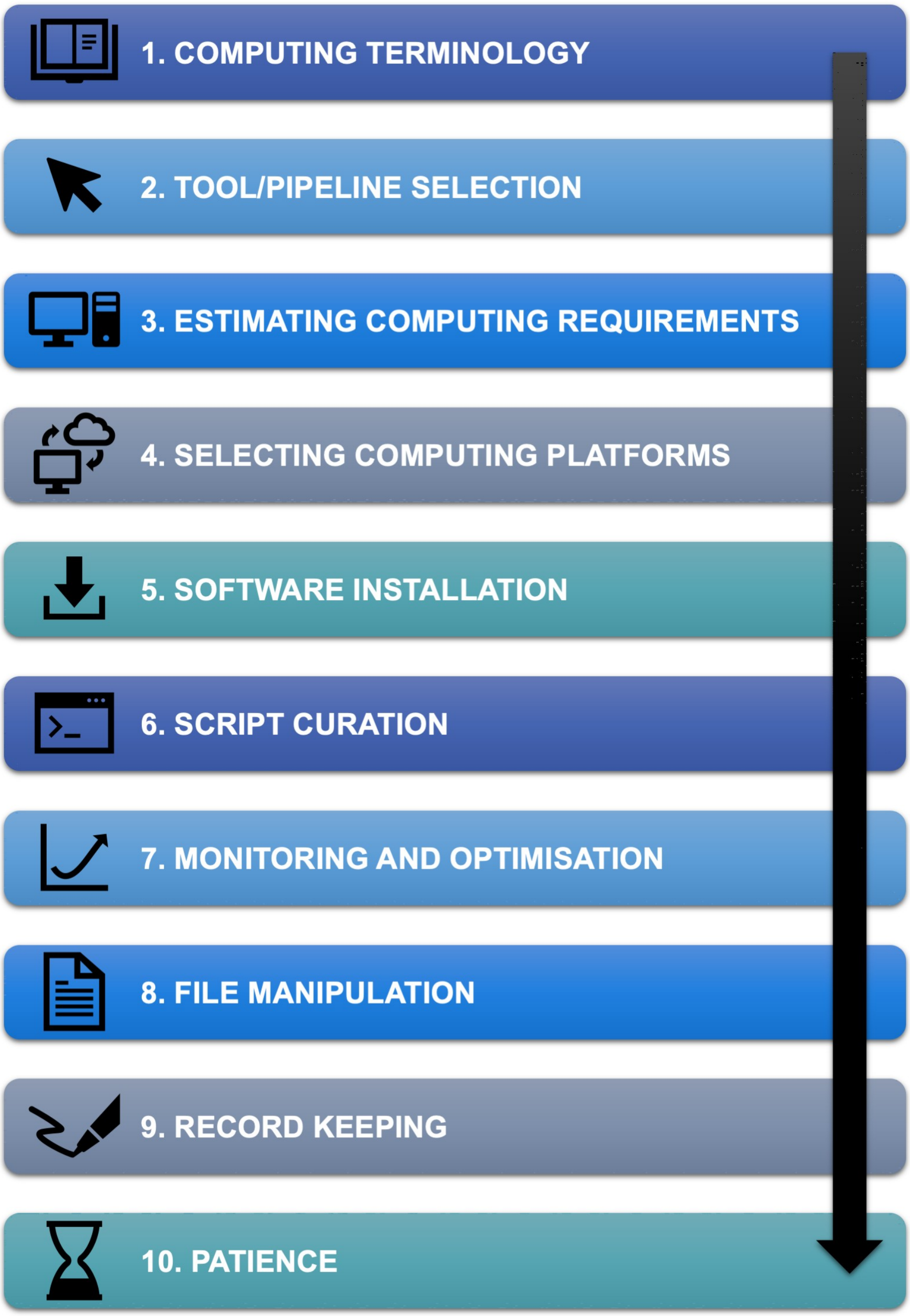

1. COMPUTING TERMINOLOGY
2. TOOL/PIPELINE SELECTION
3. ESTIMATING COMPUTING REQUIREMENTS
4. SELECTING COMPUTING PLATFORMS
5. SOFTWARE INSTALLATION
6. SCRIPT CURATION
7. MONITORING AND OPTIMISATION
8. FILE MANIPULATION
9. RECORD KEEPING
10. PATIENCE

Fig 1. Our 10-step process for getting started with command-line bioinformatics. Each step corresponds to each of our 10 simple rules presented below.

https://doi.org/10.1371/journal.pcbi.1008645.g001

Box 1. Some simple definitions of common computer terms

Algorithm: The set of rules or calculations that are performed by a computer program. Certain algorithms may be more suitable for particular datasets and may have differences in performance (e.g., in speed or accuracy).

Central processing unit (CPU): The chip that performs the actual computation on a compute node or VM.

Compute node: An individual computer that contains a number of CPUs and associated RAM.

Core: Part of a CPU. Single-core processors contain 1 core per CPU, meaning CPUs and cores are often interchangeable terms.

CPU time: The time CPUs have spent actually processing data (often CPU time ~ = Walltime * Number of CPUs).

Dependency: Software that is required by another tool or pipeline for successful execution.

Executable: The file that contains a tool/program. Some software has a single executable, while others have multiple executables for different commands/steps.

High performance computer (HPC): A collection of connected compute nodes.

Operating system (OS): The base software that supports a computer's basic functions. Some of the most common linux-based operating systems include those of the Debian distribution (Ubuntu) and those of the RedHat distribution (Fedora and CentOS).

Pipeline: A pipeline is a workflow consisting of a variety of steps (commands) and/or tools that process a given set of inputs to create the desired output files.

Programming languages: Specific syntax and rules for instructing a computer to perform specific tasks. Common programming language used in bioinformatics include Bash, Python, Perl, R, C, and C++.

Random access memory (RAM): Temporarily stores all the information the CPUs require (can be accessed by all of the CPUs on the associated node or VM).

Scheduler: Manages jobs (scripts) running on shared HPC environments. Some common schedulers include SLURM, PBS, Torque, and SGE.

Script: A file which contains code to be executed in a single programming language.

Thread: Number of computations that a program can perform concurrently—depends on the number of cores (usually 1 core = 1 thread).

Tool: A software program that performs an analysis on an input dataset to extract meaningful outputs/information—Tool, software, and program are often used interchangeably but refer to the core components of bioinformatics pipelines.

VM: Virtual machine—Similar to a compute node as it behaves as a single computer and contains a desired number of CPUs and associated RAM (usually associated with cloud computing).

Walltime: The time a program takes to run in our clock-on-the-wall time.

Rule 2: Know your data and needs to determine which tool or pipeline to use

This can often be one of the most difficult steps as there are usually many different tools and pipelines to choose from for each particular bioinformatic analysis. While you may think about creating your own tool to perform a particular task, more often than not, there is already a preexisting tool that will suit your needs, or perhaps only need minor tweaking to achieve the required result. Having a clear understanding of your data and the types of questions you are wanting to ask will go a long way to assisting in your tool or pipeline selection. Selecting the most suitable pipeline or tool will be dependent on a number of factors including:

Your target species and quality of data. Some bioinformatic pipelines/software may work better for a particular species based on their unique features (e.g., genome size, repeat complexity, ploidy, etc.) or based on the quality of data (e.g., scaffold length, short reads versus long reads, etc.). Reading other published papers on similar species will assist with being able to define this.

Your available computing resources and time restrictions. Certain software may be based of different algorithms which can result in significant reductions or increases of computational resources and walltime. Some shared HPC infrastructure may have walltime limitations in place, or the amount of RAM or cores may be a limiting factor when using personal computing resources. Make enquiries with your institutional IT department regarding limits on personal computing or HPC infrastructure before you start.

Which tools are readily available. Many bioinformatic pipelines and tools are freely available for researchers, though some require purchasing of a license. Additionally, some tools/pipelines may already be available on your desired computing infrastructure or through your local institution. There are a number of "standard" bioinformatic command line tools that have broad applicability across a variety of genomic contexts and are therefore likely already installed on shared infrastructure. Such examples include tabix, FastQC, samtools, vcftools/bcftools, bedtools, GATK, BWA, PLINK, and BUSCO. Furthermore, collaborators or other researchers may have already tested and optimised a particular pipeline on a certain infrastructure and have therefore already overcome the first hurdle for you.

Talking with colleagues who are working on similar projects and reading through the literature is often the best way to decide on which software to use for a particular analysis. There are many publications that benchmark different tools and compare the advantages and disadvantages of similar pipelines. There are also many online web forums (e.g., BioStars [10]) that may also assist with your decision-making process. Be sure to search through the different web forums to see whether another researcher has also asked the same or similar question as you (this is often the case). If you cannot find a solution, ensure any questions you post are clear and detailed, with examples of code or errors provided to have the best chance of helpful replies and answers. Beginning with a pipeline that has previously been tested and optimised on a particular platform is helpful in getting a head start, though do not be scared to try out a new or different pipeline if it seems better suited to your data or desired outcome.

Rule 3: Estimate your computing requirements

Once you have selected your desired tool or pipeline, the next crucial step involves estimating the desired computing requirements for your chosen analysis. Estimating your requirements will not only allow you to determine which platforms may be most suitable to run your pipeline (e.g., cloud versus HPC; see Rule 4) but will also reduce time spent on troubleshooting basic resource errors (e.g., running out of RAM or storage space). Furthermore, this step is almost always necessary prior to running any tool or pipeline on any given compute infrastructure. For instance, on shared HPC environments, your job script will need to include your requested computational resources (cores, RAM, walltime), and you will need to make sure you have enough disk space available for your account. Similarly, for cloud computing, you will need to decide what size machine/s (cores and RAM) and how much attached storage you need for your analysis. Estimating incorrectly can be frustrating as you will waste time in queues on shared HPC infrastructure, only to have your analysis terminated prematurely, or waste money in the cloud specifying more resources than you actually need. Many bioinformatics tools can be run on a single core by default, but this can result in much greater walltimes [11] (which are often restricted on shared HPC infrastructure). Increasing the number of cores can greatly reduce your walltime, though there is often a balance between this and other important factors such as RAM usage, cost, queueing time, etc. [11].

It can be a little tricky estimating computing requirements for a pipeline you have never run before, or on a species that the pipeline has never been tested with before. Never fear though as there are a number of places you can seek out information on computing requirements. First and foremost, read the documentation for the pipeline/tool you are running. Some tool documentation will provide an example of the compute resources required or provide suggestions. Additionally, many programs will provide a test dataset to ensure the pipeline is working correctly before employing your own datasets. These test datasets are a great start for estimating minimal computational requirements and to obtain some general benchmarks when using different parameters or computing resources. If the tool documentation does not provide a guide of computing requirements or an example dataset, you may wish to use a smaller subset of your own data for initial testing. The literature may also provide a guide for general computing requirements that have been used for a particular tool or pipeline for a similar species or sample size. There are many publications where common bioinformatics pipelines are compared with one another to assess performance and results across a variety of organisms (e.g., [12–15]). These can be found with a simple citation search. Finally, another great resource for estimating your computing requirements is from other researchers. Talking to others in your field who may work with similar data or utilising online forums such as BioStars [10] will assist in understanding the resources required.

In general, 32 cores and 128 GB of RAM is usually sufficient for most common bioinformatics pipelines to run within a reasonable timeframe. With that being said, some programs might require much less than this, while others may have much higher memory requirements or enable greater parallelisation.

Rule 4: Explore different computing options

After estimating your computing requirements for your chosen pipeline, you will then need to determine where such resources are available and which infrastructure will best suit your needs. Some tools may easily run on a personal computer, though many of the large bioinformatics pipelines (particularly when working on organisms with large genomes like mammals and plants) require computational resources that will well exceed a standard PC. Many institutions have a local HPC or access to national/international HPC infrastructure. However, the unprecedented generation of sequencing data has started to push these shared infrastructures to their limits. These resources are not always well suited to the requirements of bioinformatic pipelines such as their high I/O demands and "bursty" nature (see Rule 7) [16]. This is why cloud computing is becoming increasingly popular for bioinformaticians [16–20].

Cloud computing provides a number of key advantages over traditional shared HPC resources including:

- The ability to tailor your computing resources for each bioinformatic tool or pipeline you wish to use;
- Complete control over your computing environment (i.e., operating system, software installation, file system structure, etc.);
- Absence of a queuing system resulting in faster time to research;
- Unlimited scalability and ease of reproducibility.

Utilising cloud resources also prevents the need for researchers to purchase and maintain their own physical computer hardware (which can be time consuming, costly, and nowhere near as scalable [21]). However, commercial cloud computing does come at a cost and can be a bit of a steep learning curve. Fortunately, services like RONIN (https://ronin.cloud) have simplified the use of cloud computing for researchers and allow for simple budgeting and cost monitoring to ensure research can be conducted in a simple, cost-effective manner. Researchers at academic institutions may also have access to other free cloud compute services such as Galaxy (https://usegalaxy.org/), ecocloud (https://ecocloud.org.au/), nectar (https://nectar.org.au/cloudpage/), and CyVerse (https://www.cyverse.org).

Overall, deciding where to run your analysis will be dependent on your data/species, what platforms are most easily accessible to you, your prior experience, your timeline, and your budget. Exploring different compute options will allow you to choose which infrastructure best suits your needs and enable you to adapt to the fast-evolving world of bioinformatics.

Rule 5: Understand the basics of software installation

When wanting to utilise a personal resource for your bioinformatic pipelines, such as a cloud VM or a personal computer, you will need to get familiar with the various installation methods for your required tools. While software installation is sometimes provided as a service for some shared HPC platforms, understanding the basics of software installation is useful in helping you troubleshoot any installation-based errors and identify which software you can likely install locally yourself (i.e., without requiring root user privileges). There are numerous ways software can be installed, but we have provided 4 main methods that should cover most bioinformatics software (Box 2).

Once you have your software installed, it is good practice to try and run the program with the help command-line option (i.e., -h/—help/-help), or with no parameters, to ensure it has been installed correctly. If the help option displays some information about running the program and the different command-line options, it is usually a good sign that your software was installed successfully and is ready to go. If your tool does not seem to be working, you may need to ensure the executable for your tool (and sometimes its required dependencies) is available in your path. But what exactly is your path and why is it important? Well, whenever we call upon a particular input file or output directory within a command, we often use an absolute or relative path to show the program where that file or directory is sitting within the file system hierarchy. We can also call upon tools or executables the same way, though it is not efficient to provide a path to a tool every time we need to use it. The path environmental variable overcomes this issue by providing a list of directories that contain tools/executables you may wish to execute.

By default, the path variable is always set to include some standard directories that include a variety of system command-line utilities. So, to ensure a new program can be called upon anywhere without specifying the path to the program, you can either move or copy the tool/

Box 2. Common software installation methods for bioinformatics tools

Package managers

APT (Advanced Package Tool) (https://www.debian.org/doc/manuals/apt-guide/index.en.html) is a package manager that is often already installed by default on many Debian distributions and enables very simple installation of available tools. APT works with a variety of core libraries to automate the download, configuration, and installation of software packages and their dependencies. A number of common bioinformatics tools are available through APT including NCBI blast+, samtools, hmmer, vcftools, bcftools, bedtools among others. If working on a RedHat operating system, the package manager YUM (Yellowdog Updater, Modified) (https://access.redhat.com/solutions/9934) is the equivalent of APT.

Conda

Conda (https://docs.conda.io/en/latest/) is also a package management tool, though it sits somewhere between package managers like APT and containers (see below) due to its ability to also manage environments (i.e., collections of software). This feature makes conda extremely useful, particularly for bioinformatics software where different pipelines may utilise the same tools but require different versions of a particular tool. Conda allows you to easily install and run pipelines in their own separate environments so they do not interfere with one another and also enables you to easily update software when new versions are made available. Bioconda [22] is a channel for conda which specialises in bioinformatics software and includes a myriad of the most commonly used bioinformatic tools. Furthermore, conda also enables the installation and management of popular programming languages such as python or R, along with their respective libraries and packages. It is a great resource for bioinformaticians of all levels and is particularly helpful as a stepping-stone before stepping down a container lane.

Containers

Containers package up software and all dependencies, as well as all of the base system tools and system libraries into a separate environment so that they can be reliably run on different computing platforms. Containers are similar to conda environments, but they differ in the sense that containers include absolutely everything they need within the container itself (even including the base operating system). It is sometimes easier to think about containers as installing a whole separate machine that just utilises the same computing resources and hardware as the local machine it is installed on. The main advantage of a container over a conda environment is the ease of reproducibility due to the ability to pull a specific container each time you want to run, or re-run, a certain pipeline or use a particular tool, no matter what computing platform you are using. Reproducibility can be achieved with conda environments too, but this often requires exporting and keeping track of saved environments.

There are 2 main options when wanting to use a container: Docker [23] or Singularity [24]. Docker is the most standard container service available with thousands of containers available from DockerHub (https://hub.docker.com) or from other container registries such as quay.io (https://quay.io). Bioinformatics software that is available via bioconda also has a respective docker container on quay.io through the BioContainers architecture [25]. This means many common bioinformatics software and pipelines are already available in a containerised environment. Otherwise, some software developers make their own containers available, e.g., Trinity (for RNA-seq assembly) (see https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-in-Docker) or BUSCO v4 (for assessing assembly completeness) (see https://busco.ezlab.org/busco_userguide.html#docker-image). There are also thousands of other public docker containers across a range of online container registries that may have the software you are looking for, or there is always the option to create your own Docker container for reproducible pipelines. Obviously, Docker can be used to download and employ Docker containers, but Singularity is another program that can also be used to download and employ Docker containers (particularly on HPC environments). Both have advantages and disadvantages, so it is usually down to user preference as to which to choose. If you are new to containers, we suggest starting with Singularity. Not only will this allow you to easily be able to scale up your containerised pipelines to HPC environments but also makes reading and writing files to and from the container from the local machine a bit more straightforward.

Manual installation

If none of the above methods are available for your chosen software, you may need to install it manually. This process is usually explained step-by-step in the software documentation but typically involves a number of steps including: (1) Downloading a tar package (or zip file) of the source code (or cloning a Git repository) from GitHub (https://github.com) (or another website); (2) Unpacking the source code to extract its contents; (3) Configuring the software to check your environment and ensure all of the required dependencies are available; (4) Building the finished software from the source code; and (5) Installing the software, i.e., copying the software executables, libraries, and documentation to the required locations. This process is what package managers and containers do automatically for you. There are a number of standard dependencies that are usually required for manual installation (e.g., the build-essential package, the dh-autoreconf package, and the libarchive-dev package) so it is often handy to install these using APT before attempting to manually install any other software. You will be notified of any other required dependencies you may be missing during the installation process.

executable to a directory that is already listed in your path variable, or add a new directory to the path variable that contains the program. New directories can be added to your path either temporarily (by simply exporting the path variable with the added directory included) or permanently (by editing your.bash_profile). Another thing to be aware of is that the order of directories in your path is important because if the same program (or executable with the same name) is found in 2 different directories, the one that is found first in your path will be used. Always keep this in mind when adding new directories to your path to determine where they should sit in the list of paths. [The sheer number of times we mentioned the word “path” in this rule alone should emphasise how important paths really are—though we promise there are no more mentions of it for the rest of this article].

Rule 6: Carefully curate and test your scripts

In other words, always double-check (or triple-check) your scripts and perform test runs at each step along the way. Before you run your pipeline, it is important to first read through the software documentation to ensure you understand the different inputs, outputs, and analysis options. Ensure that the documentation is for the correct version of the software as particular command-line options may change version to version. Many bioinformatics programs have extensive documentation online, either through their GitHub or another website. The basic documentation for most tools can be accessed using the command-line help options (which is also a great way to determine whether your required tool is available and installed correctly—see Rule 5). Sometimes more detailed information can be found in a README file in the source code directory. Most documentation should provide some example commands on how to run the program with basic or default options which should assist you in curating a successful script.

Once you have your final script, it is essential to give it a quick test to determine if there are any immediate errors that will prevent your script from running successfully. From simple spelling mistakes or syntax errors which result in files or directories not being found or commands being confused with invalid options, to not being able to locate the desired software or the software being configured incorrectly with problematic dependencies. These are the "facepalm" errors that any bioinformatician is aware of as we have all been there, time and time again. The good news is that these errors are often quite simple to fix. Yet it is better to catch them early rather than waiting in queues only for your script to error as soon as it starts, or leaving your script to run in the cloud only to come back and realise the machine has been sitting there idle the whole time due to a minor scripting error. Testing your scripts in the cloud is usually as simple as running the script or command and watching to see whether any errors are immediately thrown on-screen, but to test scripts in a shared HPC environment, you may need to utilise an interactive queue. Interactive queues allow you to run commands directly from the command line with a small subset of HPC resources. These resources are usually not enough to run an entire pipeline but are quite useful for testing and debugging purposes. Obviously, your script may still run into errors later on in your pipeline, but testing your script before you submit it properly should alert you to any preliminary errors that would prevent the pipeline from starting successfully and prevent any precious time being wasted in queues or precious dollars being wasted on idle cloud compute.

Rule 7: Monitor and optimise your pipelines

Once you have your script running, it is important to monitor your pipelines to determine whether it is effectively utilising the computational resources you have allocated to it. Understanding what resources your pipeline utilises can help you scale up or down your compute so that you are not wasting resources or hitting resource limits that may slow down your pipeline. On shared HPC infrastructure, you will usually be able to see a summary of the computational resources used from either the job log files or scheduler-specific commands. Metrics such as maximum RAM and CPU usage as well as CPU time and walltime are useful in adjusting future scripts so that they request the optimum amount of resources needed. This enables the pipeline to run efficiently without any unnecessary queue time. Storage space of output files should also be monitored periodically to ensure you are not exceeding your allocated quota.

More specific monitoring is possible when running pipelines in the cloud as you have full control over all computing resources. Simple programs like htop (https://hisham.hm/htop/) can be used for fast real-time monitoring of basic metrics like CPU and RAM usage, while more in-depth programs like Netdata (https://www.netdata.cloud) can assist with tracking a large variety of metrics both in real-time and across an entire pipeline using hundreds of preconfigured interactive graphs. Many bioinformatic pipelines are “bursty” in nature, meaning different steps in a single pipeline may have vastly different computing requirements. Some steps/tools may have high memory requirements but only utilise a small number of cores, while others may multithread quite well across a large number of cores but require minimal memory. Knowing the required computing resources for each step may help you break up your pipeline and run each stage on a different machine type for greater cost efficiency. Monitoring disk space requirements throughout a pipeline is also important as many bioinformatics tools require large amounts of temporary storage that are often cleaned upon completion of the pipeline. Attached storage can be quite costly in the cloud, so ensuring you only request what is necessary will also reduce pipeline costs.

Overall, monitoring of bioinformatics pipelines is key to improving pipeline efficiency, optimising computing resources, reducing wasted queue time, and reducing cloud costs.

Rule 8: Get familiar with basic bash commands

As a bioinformatician, your main role is to make sense of biological datasets, and this often means manipulating, sorting, and filtering input and output files to and from various bioinformatic tools and pipelines. For example, you may want to extract information for a certain sample or a certain gene of interest. Or in a file containing a table of data, you may want to sort an output file by a particular column or select rows that contain a particular value. You may want to replace a certain ID with a respective name from a list or perform a calculation on values within a column. Fortunately, many of the input and output files used in bioinformatics are regular text files, so these tasks can easily be achieved. One might think about using common spreadsheet applications such as Microsoft excel to perform these tasks; however, while this may suffice for small files, excel is not too fond of the sometimes millions of rows of data that are characteristic of a number of common bioinformatic files. This is where some standard unix shell command-line utilities come into play, namely the grep, AWK, and sed utilities.

Global regular expression print (grep) is a command-line utility which searches a text file for a regular expression (i.e., a pattern of text) and returns lines containing the matched expression (Table 1). This tool is useful when wanting to filter or subset a file based on the presence of a particular word or pattern of text (e.g., a sample name or genomic location, etc.). AWK is much more extensive command-line utility which enables more specific file manipulation of column-based files (Table 1). For example, AWK can return lines where a column contains a particular value or regular expression; in addition, it can output only particular columns, perform calculations on values within the columns, and work with multiple files at once. The extensive abilities of AWK are too grand to cover here but just know that this clever little tool will likely hold a special place in any bioinformatician’s heart. Lastly, stream editor (sed) has a basic “find and replace” usage allowing you to transform defined patterns in your text. In its most basic form, sed can replace a word with another given word (Table 1) but can also perform more useful functions like removing everything before or after a certain pattern or adding text at certain places in a file.

Table 1. Basic usage examples of the grep, awk, and sed commands.

Command	Example	Description
grep	grep "chr5" file	Print all lines that contain the string "chr5" in the named file
awk	awk ’$1 == 5 {print $2, $3}’ file	For rows in the named file where the value in column 1 is equal to 5, print columns 2 and 3
sed	sed ’s/sample1/ID7037/g’ file	Replace all occurrences of "sample1" with "ID7037" in the named file and print the result

https://doi.org/10.1371/journal.pcbi.1008645.t001

Of course, grep, AWK, and sed all have their limitations, and more extensive file manipulation may be better suited to a python or perl script (and there is already a great "Ten simple rules" article for biologists wanting to learn how to program [26]); but for simple processing, filtering, and manipulation of bioinformatics files, look no further than these 3 useful command-line utilities.

Rule 9: Write it down!

A previous "Ten simple rules" article has highlighted the importance of keeping a laboratory notebook for computational biologists [27], and another covered some best practices around the documentation of scientific software [28]. Many components from these articles apply to our rule of writing it down and keeping helpful notes when getting started with command-line bioinformatics. The number of pipelines or analyses that can be run on a single set of biological data can sometimes be quite extensive and usually coincides with a lot of trial and error of different parameters, computing resources, and/or tools. Even those with a great memory will often look back at results at the time of publication and ponder "why did we use that tool?", or "what parameters did we end up deciding on for that analysis?". Keeping detailed notes can be a real lifesaver. Not only is it important to keep track of your different script files, and the required computing resources for each script, but also the accompanied notes about why you chose a particular tool and any troubleshooting you had to do to run the pipeline successfully. An easy-to-access document of all of your favourite commands and nifty pieces of code that may come in handy time and time again is also a must! Getting familiar with helpful code text editors like Visual Studio Code (https://code.visualstudio.com), or Atom (https://atom.io), as well as investing some time into learning helpful mark-up languages like Markdown will assist with keeping detailed, organised, and well-formatted scripts and documentation for the pipelines you are using. Exactly how you decide to keep your notes is completely up to you, but just ensure to keep everything well-organised, up-to-date, and backed up. Also, publishing your scripts as markdown files in supplementary material ensures the utility (and citability) of your work.

Rule 10: Patience is key

The number 1 key (that we've saved until last) to being a successful bioinformatician is patience. A large proportion of your time will be spent troubleshooting software installation, computing errors, pipeline errors, scripting errors, or weird results. Some problems are simple to solve, while others may take quite some time. You will likely feel that with every step forward, there is just another hurdle to cross. Yet if you are patient and push through every error that is thrown your way, the euphoria of conquering a bioinformatics pipeline and turning a big lump of numeric data or As, Ts, Cs, and Gs into something biologically meaningful is well worth it. Also, as many past "Ten simple rules" articles in this field have addressed, do not be afraid to raise your hand and ask for help when you get stuck. Most of the time, someone before you has been in the exact same situation and encountered the same error or tackled a similar problem. Google will become your best friend and first port of call when things are not going as planned. And on the rare occasion where endless googling leads you nowhere, talk with your peers and reach out to the bioinformatic community; people are often more than happy to share their knowledge and put their problem-solving skills to the test.

Conclusion

In the new era of whole genome sequencing, bioinformaticians are now more sought-after than ever before. Stepping into the world of command-line bioinformatics can be a steep learning curve but is a challenge well worth undertaking. We hope these 10 simple rules will give any aspiring bioinformatician a head start on their journey to unlocking the meaningful implications hidden within the depths of their biological datasets.

Acknowledgments

Our journey into bioinformatics and this article would not have been possible without the support of the teams at the Sydney Informatics Hub, Australian BioCommons, RONIN, Amazon Web Services (AWS), Pawsey Supercomputing Centre, National Computational Infrastructure, and Microsoft Azure. Special thanks to Rosie Sadsad, Nathan Albrighton, Johan Gustafsson, Jeff Christiansen, Kathy Belov, and the team members of the Australasian Wildlife Genomics Group.

References

1. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A. 2018; 115(17):4325–33. Epub 2018/04/25. https://doi.org/10.1073/pnas.1720115115 PMID: 29686065; PubMed Central PMCID: PMC5924910.
2. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. J Hered. 2009; 100(6):659–74. https://doi.org/10.1093/jhered/esp086 PMID: 19892720
3. Koepfli K-P, Paten B, Genome 10K Community of Scientists, O'Brien SJ. The Genome 10K Project: a way forward. Annu Rev Anim Biosci. 2015; 3(1):57–111. Epub 2015/02/18. https://doi.org/10.1146/annurev-animal-090414-014900 PMID: 25689317; PubMed Central PMCID: PMC5837290.
4. GIGA Community of Scientists. The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes. J Hered. 2013; 105(1):1–18.
5. Voolstra CR, Wörheide G, Lopez JV. Corrigendum to: Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA). Invertebr Syst. 2017; 31(2):231–.
6. Consortium iK. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 2013; 104(5):595–600. https://doi.org/10.1093/jhered/est050 PMID: 23940263
7. Levine R. i5k: the 5,000 insect genome project. Am Entomol. 2011; 57(2):110–3.
8. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, et al. 10KP: A phylodiverse genome sequencing plan. Gigascience. 2018; 7(3):giy013. https://doi.org/10.1093/gigascience/giy013 PMID: 29618049
9. Kumuthini J, Chimenti M, Nahnsen S, Peltzer A, Meraba R, McFadyen R, et al. Ten simple rules for providing effective bioinformatics research support. PLoS Comput Biol. 2020; 13(3):e1007531. https://doi.org/10.1371/journal.pcbi.1007531 PMID: 32214318
10. Parnell LD, Lindenbaum P, Shameer K, Dall'Olio GM, Swan DC, Jensen LJ, et al. BioStar: an online question & answer resource for the bioinformatics community. PLoS Comput Biol. 2011; 7(10): e1002216. https://doi.org/10.1371/journal.pcbi.1002216 PMID: 22046109
11. Kawalia A, Motameny S, Wonczak S, Thiele H, Nieroda L, Jabbari K, et al. Leveraging the power of high performance computing for next generation sequencing data analysis: tricks and twists from a high throughput exome workflow. PLoS ONE. 2015; 10(5):e0126321. https://doi.org/10.1371/journal.pone.0126321 PMID: 25942438
12. Zhang C, Zhang B, Lin L-L, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics. 2017; 18(1):583. https://doi.org/10.1186/s12864-017-4002-1 PMID: 28784092
13. Cornish A, Guda C. A comparison of variant calling pipelines using genome in a bottle as a reference. Biomed Res Int. 2015;2015.
14. Khan AR, Pervez MT, Babar ME, Naveed N, Shoaib M. A comprehensive study of de novo genome assemblers: current challenges and future prospective. Evol Bioinform. 2018; 14:1176934318758650. https://doi.org/10.1177/1176934318758650 PMID: 29511353
15. Schilbert HM, Rempel A, Pucker B. Comparison of read mapping and variant calling tools for the analysis of plant NGS data. Plants. 2020; 9(4):439. https://doi.org/10.3390/plants9040439 PMID: 32252268

16. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform. 2013; 46(5):774–81. https://doi.org/10.1016/j.jbi.2013.07.001 PMID: 23872175

17. Kwon T, Yoo WG, Lee W-J, Kim W, Kim D-W. Next-generation sequencing data analysis on cloud computing. Genes Genom. 2015; 37(6):489–501.

18. Shanker A. Genome research in the cloud. OMICS J Integr Biol. 2012; 16(7–8):422–8. https://doi.org/10.1089/omi.2012.0001 PMID: 22734722

19. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010; 11(5):207. https://doi.org/10.1186/gb-2010-11-5-207 PMID: 20441614

20. Zhao S, Watrous K, Zhang C, Zhang B. Cloud computing for next-generation sequencing data analysis. In: Jaydip Sen, editor. Cloud Computing-Architecture and Applications. Rijeka: InTech; 2017. p. 29–51.

21. Fox A. Cloud Computing—What's in It for Me as a Scientist? Science. 2011; 331(6016):406–7. https://doi.org/10.1126/science.1198981 PMID: 21273473

22. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018; 15(7):475–6. https://doi.org/10.1038/s41592-018-0046-7 PMID: 29967506

23. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014; 2014(239):2.

24. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS ONE. 2017; 12(5):e0177459. https://doi.org/10.1371/journal.pone.0177459 PMID: 28494014

25. da Veiga Leprevost F, Grüning BA, Alves Aflitos S, Röst HL, Uszkoreit J, Barsnes H, et al. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics. 2017; 33(16):2580–2. https://doi.org/10.1093/bioinformatics/btx192 PMID: 28379341

26. Carey MA, Papin JA. Ten simple rules for biologists learning to program. PLoS Comput Biol. 2018; 14(1):e1005871. https://doi.org/10.1371/journal.pcbi.1005871 PMID: 29300745

27. Schnell S. Ten simple rules for a computational biologist's laboratory notebook. PLoS Comput Biol. 2015; 11(9):e1004385. https://doi.org/10.1371/journal.pcbi.1004385 PMID: 26356732

28. Lee BD. Ten simple rules for documenting scientific software. PLoS Comput Biol. 2018; 14(12): e1006561. https://doi.org/10.1371/journal.pcbi.1006561 PMID: 30571677

A1.3 CHARACTERISATION OF REPRODUCTIVE GENE DIVERSITY IN THE ENDANGERED TASMANIAN DEVIL

The PDF version of the article titled “Characterisation of reproductive gene diversity in the endangered Tasmanian devil” published in *Molecular Ecology Resources* (2020; 00, 1-12), which comprises Chapter 3 of this thesis, is presented on the following pages.

Received: 25 August 2020 | Revised: 25 October 2020 | Accepted: 5 November 2020

DOI: 10.1111/1755-0998.13295

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Characterization of reproductive gene diversity in the endangered Tasmanian devil

Parice A. Brandies[1] | Belinda R. Wright[1] | Carolyn J. Hogg[1] | Catherine E. Grueber[1,2] | Katherine Belov[1]

[1]School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, NSW, Australia

[2]San Diego Zoo Global, San Diego, CA, USA

Correspondence
Katherine Belov, School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia.
Email: kathy.belov@sydney.edu.au

Funding information
Australian Research Council, Grant/Award Number: LP140100508 and DP170101253

Abstract

Interindividual variation at genes known to play a role in reproduction may impact reproductive fitness. The Tasmanian devil is an endangered Australian marsupial with low genetic diversity. Recent work has shown concerning declines in productivity in both wild and captive populations over time. Understanding whether functional diversity exists at reproductive genes in the Tasmanian devil is a key first step in identifying genes that may influence productivity. We characterized single nucleotide polymorphisms (SNPs) at 214 genes involved in reproduction in 37 Tasmanian devils. Twenty genes contained nonsynonymous substitutions, with genes involved in embryogenesis, fertilization and hormonal regulation of reproduction displaying greater numbers of nonsynonymous SNPs than synonymous SNPs. Two genes, *ADAMTS9* and *NANOG*, showed putative signatures of balancing selection indicating that natural selection is maintaining diversity at these genes despite the species exhibiting low overall levels of genetic diversity. We will use this information in future to examine the interplay between reproductive gene variation and reproductive fitness in Tasmanian devil populations.

KEYWORDS

conservation, genetic variation, reproduction, *Sarcophilus harrisii*, Tasmanian devil

1 | INTRODUCTION

Globally the number of species threatened with extinction is increasing as a result of human-induced activities including habitat fragmentation, invasive predators and pollution. Genetic diversity at functional gene families can have long-term consequences on species adaptation and survival in a changing world (Holderegger et al., 2006; Mimura et al., 2017). Understanding the causes and consequences of interindividual variation sits at the core of evolution and ecology, yet despite decades of molecular research, the genetic basis of phenotypic variation (i.e., genetic polymorphism) remains poorly quantified for the vast majority of species and traits (Forsman & Wennersten, 2016; Mimura et al., 2017). However, recent advances in sequencing technology have better enabled researchers to investigate interindividual variation at gene families and determine how this variation is linked to important phenotypic traits. For example, genetic diversity at immune genes, particularly genes of the major histocompatibility complex (MHC), have been associated with a range of key biological phenomena such as disease susceptibility and mate choice (Brandies et al., 2018; Sommer, 2005). These phenomena have significant implications on fitness and as a result interindividual variation at MHC loci has been extensively studied across a number of threatened species (Ujvari & Belov, 2011). Studies of MHC and other immune genes have demonstrated how characterizing genetic variation is crucial to predicting which genes may contribute to variable phenotypes, and the resultant implications for species conservation. However, little is currently known about diversity at other important gene families in threatened species.

Variation at reproductive genes may contribute to key productivity traits that impact the survival of threatened species. Relationships between gene variants and reproductive phenotypes have been extensively studied across a range of model organisms, from *Drosophila* to humans. For example, polymorphisms in male reproductive genes have been associated with variation in sperm competitive ability in *Drosophila* (Fiumera et al., 2005) and a range of gene mutations have been linked to infertility in humans (Layman, 2002). Associations between variants of key reproductive genes (e.g., those involved in the production or binding of reproductive hormones) and reproductive traits have also been reported in livestock species where high productivity is important (Kirkpatrick, 2002). Examining diversity at genes known to be involved in reproduction is a fundamental first step in determining which loci have the potential to underlie important reproductive traits. However, little is currently known about the variation at reproductive genes in wildlife species, particularly in threatened species that exhibit low levels of genetic diversity overall.

The Tasmanian devil (*Sarcophilus harrisii*) is one such threatened species that is suffering from a range of threatening processes, in addition to having low genome-wide diversity. Devils are the largest extant carnivorous marsupial and are native to the island state of Tasmania, Australia (Owen & Pemberton, 2005). Populations have declined by up to 80% across this species' range due to a contagious cancer, known as devil facial tumour disease (DFTD). Historical population declines and contemporary habitat fragmentation have resulted in the erosion of genetic diversity (Jones et al., 2004; Miller et al., 2011), particularly at immune gene loci that are highly polymorphic in other species (Cheng et al., 2012; Morris et al., 2015). Tasmanian devils exhibit a number of interesting life-history strategies such the ability of females to undergo up to three oestrous cycles per breeding season (Keeley et al., 2012), the production of up to 30 embryos, of which only four can be supported by the four teats (Guiler, 1970; Hughes, 1982), precocial breeding (Lachish et al., 2009; Russell et al., 2019) and multiple paternity litters (Russell et al., 2019). Despite these unique reproductive traits, Tasmanian devils have shown concerning declines in productivity in both captivity (Farquharson et al., 2017) and the wild (Farquharson et al., 2018). So, an understanding of whether diversity exists at reproductive genes is a fundamental step in identifying genes that may be associated with differential reproductive phenotypes, and hence may influence reproductive fitness. Armed with this basic knowledge, conservation managers can then use this information in their management decisions pertaining to captive breeding and translocations.

Here, we aimed to identify and characterize reproductive genes, and then examine single nucleotide polymorphism (SNP) diversity at these genes using 37 resequenced Tasmanian devil genomes. We explore signatures of selection to identify polymorphic genes with adaptive potential (i.e., genes where specific alleles may result in differential phenotypes that are beneficial under particular circumstances). The results from this study provide a resource for future research to examine the association between reproductive diversity and productivity in the Tasmanian devil.

2 | MATERIALS AND METHODS

2.1 | Gene identification and characterization

In total, 250 genes that have previously been associated with reproduction in mammalian species were selected based on literature searches using the search terms "reproduction" and "gene," as well as mining the human gene database GeneCards (www.genecards.org, Stelzer et al., 2016) using the keyword "reproduction." The identified genes are involved in a variety of reproductive stages including: the hormonal regulation of reproduction, sexual/reproductive development, gametogenesis, fertilization and embryogenesis. Predicted complete and partial gene sequences from NCBI's or Ensembl's automatic annotation process were identified in the Tasmanian devil genome reference assembly on NCBI (Devil_ref v7.0 [GCA_000189315.1], Murchison et al., 2012).

Gene predictions in the Tasmanian devil genome were checked using a number of methods including: (a) confirming gene synteny against model organisms (human and mouse) and the current highest-quality marsupial genome (koala) using NCBI's genome viewer (NCBI Resource Coordinators, 2017); (b) mapping the predicted coding sequences (CDS) back to the reference genome using SPLIGN (Kapustin et al., 2008) to ensure all exons were correctly identified and confirm that CDS were complete and did not contain any premature stop codons or frameshift mutations; and (c) performing a BLASTP (Altschul et al., 1990) search on the predicted translated sequences against the UniProt (Consortium, 2018) database to confirm identity and protein lengths. For genes with multiple isoforms, the first-named isoform (Variant X1) was investigated (usually the longest). All genes were utilized in downstream analyses.

For partial gene predictions, any missing exons were identified by comparison to well-annotated model organism orthologues using the NCBI genome viewer (NCBI Resource Coordinators, 2017) and TBLASTN (Altschul et al., 1990) searches. Where exons were unable to be fully resolved (i.e., due to gaps in the reference sequence, genome fragmentation, etc.) partial sequences were utilized in downstream analyses. For any genes not automatically annotated in the reference genome by NCBI or Ensembl, the predicted location of these genes was identified through gene synteny and TBLASTN searches with model organisms (human and mouse), and gene prediction was performed using FGENESH+ (Solovyev, 2004) with koala orthologues as an input. If an orthologous sequence was not available in koala, human or mouse orthologues were used as an input instead.

2.2 | Sample collection and genome resequencing

Two existing data sets of resequenced genomes were used to explore reproductive gene diversity in the Tasmanian devil. The first data set comprised 25 individuals (including 12 wild-born founders [Figure S1] and nine parent–offspring trios [Figure S2]) that were sequenced to a high coverage of ~45× (SRA accessions: SRX6096677–SRX6096696, Wright et al., 2020). The second data set included

12 wild individuals from a separate wild population (Figure S1) sequenced to a low coverage of 10–15× (SRA accessions: ERS682204–ERS682210; ERS1202857–ERS1202861, Wright et al., 2015, 2017). This low-coverage data set was only included following the preliminary SNP identification to minimize the risk of this data set introducing false SNPs. We refer to the 12 low-coverage genomes as “12L” to differentiate it from the data set encompassing the 25 high-coverage resequenced genomes (“25H”).

2.3 | Preliminary SNP identification

To identify an initial high-confidence target SNP set, whole-genome alignment and SNP calling was performed on the 25H data set following the methods given by Wright et al. (2020). Briefly, reads were aligned to the Tasmanian devil reference genome assembly version 7.0 (GenBank: GCA_000189315.1, Murchison et al., 2012) using BWA version 0.7.15 (Li & Durbin, 2009). PCR duplicates were removed with PICARDTOOLS version 1.119 (http://broadinstitute.github.io/picard/) and indel realignment was performed with GATK version 3.6 (McKenna et al., 2010). SNPs were called using SAMTOOLS version 1.6 (Li et al., 2009) with minimum base and mapping quality of 30 and a coefficient for downgrading mapping quality for reads containing excessive mismatches of 50. ANNOVAR version 20180416 (Yang & Wang, 2015) gene-based annotation was used to annotate all variants from each of the 25H resequenced genomes aligned to the reference genome using the corresponding genome annotation file from NCBI (O'Leary et al., 2015). Any genes not included in the NCBI annotation were checked for SNPs manually in GENEIOUS (Kearse et al., 2012). SNPs associated with the reproductive genes in the 25H Tasmanian devils were identified by filtering the ANNOVAR output, and the total number of each type of SNP (synonymous, nonsynonymous, splicing, UTR5, UTR3, intronic, upstream, downstream) was calculated for each gene. Reproductive genes containing nonsynonymous SNPs were targeted for further analysis. The 12L data set was not included in the initial SNP identification procedure in order to minimize the risk of false positive SNPs, which may have resulted in inaccurate target gene identification, because SNPs from low-coverage data sets cannot be called as confidently as from higher-coverage data.

2.4 | Nonsynonymous SNP confirmation and analysis

Reproductive genes containing nonsynonymous SNPs were investigated further in both the original 25H resequenced genomes as well as the 12L resequenced genomes. Variants within the target reproductive genes of the 12L resequenced genomes were called together with the 25H resequenced genomes using the same parameters, as above. This method was chosen as multisample callers result in the best accuracy when lower coverage samples are called simultaneously with a larger number of higher coverage individuals (Cheng et al., 2014). Individual sample VCF files were then subset from the multisample VCF file and filtered to exclude variants below a filtered depth threshold using BCFTOOLS version 1.3.1 (Li et al., 2009). We chose a minimum filtered read depth of 10 for the 25H resequenced genomes and a minimum filtered read depth of five for the 12L resequenced genomes to increase confidence in the variant calls while preventing excessive data loss. The remaining variants were then merged into a multisample VCF file and converted to transposed PLINK format (Purcell et al., 2007) using VCFTOOLS version 0.1.14 (Danecek et al., 2011). PLINK version 1.90 was used to calculate minor allele frequencies (MAFs) and determine genotypes for all variants present within the coding regions of the target reproductive genes. Any variants with an MAF below 0.05 that were called in only one individual and had a low allelic depth (below 10), were removed in GENEIOUS. Any positions that were called as variants relative to the reference, but which were monomorphic across the 37 resequenced genomes (i.e., MAF = 0), were also filtered out using GATK and BCFTOOLS. The final variant call files were used to create consensus sequences for each individual using GATK. IUPAC ambiguity codes were used to represent heterozygous positions in the individual consensus sequences, and any positions below the specified filtered read depth (as above), or with a missing genotype, were masked. Extraction of CDS for the target genes was performed using BEDTOOLS version 2.25 (Quinlan & Hall, 2010) with a custom bed file containing the target gene regions and exon positions. Alignments of the CDS were mapped to the reference in GENEIOUS to confirm all synonymous and nonsynonymous SNPs. Missing data/genotyping rate (by locus and individual), MAFs, heterozygosity and deviations from Hardy–Weinberg equilibrium were calculated for the identified nonsynonymous SNPs in PLINK version 1.90 (Purcell et al., 2007). These analyses were performed on all samples and again with the nine known offspring removed to ensure the measures were not influenced by relatedness.

2.5 | Population diversity analysis

CDS alignments of genes confirmed to contain SNPs were converted to PHASE format using SEQPHASE (Flot, 2010). PHASE version 2.1 (Stephens & Donnelly, 2003; Stephens et al., 2001) was used to construct haplotypes using the original model with default iteration parameters and output probability thresholds (-*p* and -*q*) set to 0. This was performed to ensure any missing SNPs were imputed (based on the distributions of known haplotypes and allele frequencies across the entire data set, see Stephens & Donnelly, 2003; Stephens et al., 2001) prior to performing the population diversity analysis. The -x flag was used to run the algorithm five times (with random seeds for each run) for each gene and the run with the highest goodness-of-fit statistic was selected for the output. SEQPHASE was used to convert the PHASE output files to FASTA format and CERVUS 3.0.7 (Kalinowski et al., 2007) was used to test whether the phased haplotypes were consistent across the nine trios present in the data set. DNASP version 6 (Rozas et al., 2017) was used to infer the number of haplotypes (*h*), haplotype diversity (*hd*) and nucleotide diversity per site (π) for each gene. Deviations from the neutral model of molecular evolution were tested using Tajima's *D* (Tajima, 1989) in DNASP and codon-based *Z*-tests of selection were performed in MEGA7 (Kumar et al., 2016) using the Nei–Gojobori method (Nei & Gojobori, 1986) with variance estimated from 500 bootstraps. These statistics were repeated with the nine known offspring excluded to ensure any significant findings were not influenced by relatedness.

3 | RESULTS

3.1 | Gene characterization

Of 250 genes examined, 214 had predicted (complete or partial) CDS (Table S1). These 214 predicted genes were confirmed through analysis of gene synteny, CDS and BLASTP searches and were investigated in the subsequent SNP analysis. The remaining 36 genes were not automatically annotated by NCBI or Ensembl and could not be identified in the Tasmanian devil genome (Table S2).

3.2 | SNP identification and analysis

Using our 25H resequenced genomes, we identified over 5,000 putative SNPs associated with the 214 reproductive genes investigated (Figure 1) with an average of 28 putative SNPs per gene (range 0–549) (Table S3). Approximately 90% of these SNPs were intronic (Table S3). Forty-nine genes (23% of all genes investigated) were predicted to contain exonic SNPs, with 34 of these genes predicted to contain at least one nonsynonymous SNP (Table S3). Genes involved in embryogenesis, fertilization and hormonal regulation of reproduction displayed greater numbers of nonsynonymous SNPs than synonymous SNPs (Figure 1).

Confirmation of putative nonsynonymous SNPs was performed by analysing data from the 12L and 25H resequenced genomes together, along with additional filtering (see Methods). After filtering, 33 nonsynonymous SNPs across 20 of the genes remained (Table S4). These 20 genes represented molecular processes across a range of reproductive roles in females, males or both sexes (Table 1). For these nonsynonymous SNPs, the genotyping rate (percentage

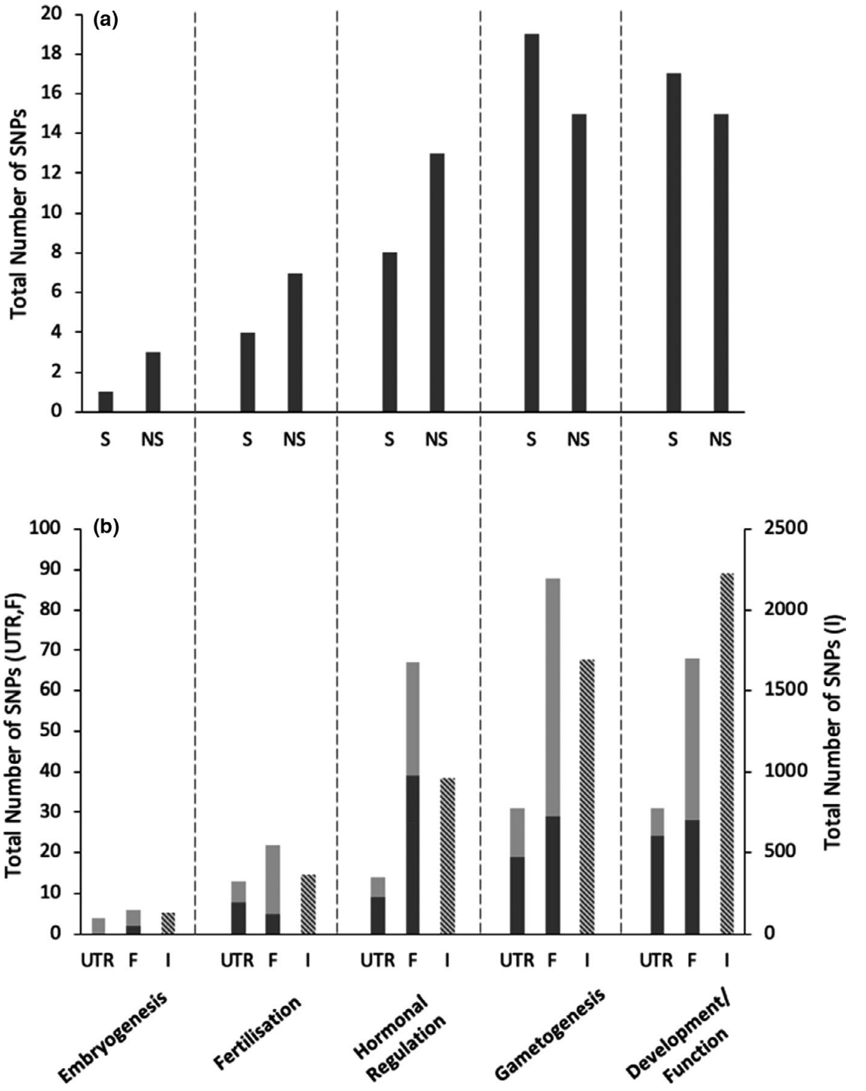


FIGURE 1 Total number of SNPs identified in genes known or predicted to be involved in a variety of reproductive functions including embryogenesis (*N* = 13 genes), fertilization (*N* = 26), hormonal regulation of reproduction (*N* = 43), gametogenesis (*N* = 74), and general reproductive development and function (*N* = 58). (a) Exonic SNPs including synonymous (S) and nonsynonymous (NS) SNPs. (b) Other major SNP types including untranslated regions (UTR), flanking regions (F) and intronic regions (I). Stripes indicate intronic SNPs are plotted on the secondary axis. Light shading indicates SNPs that are 5′ (upstream), and dark shading indicates SNPs that are 3′ (downstream). See Table S3 for more information.

TABLE 1 Reproductive roles of genes found to contain nonsynonymous SNPs

Gene	Role in reproduction	Sex affected	Reference
ADAMTS9	Important in uterine remodelling of implantation, placentation and parturition	Female	Russell et al. (2015)
ADAMTS10	Important for adhesion between the sperm and egg zona pellucida	Male	Dun et al., 2012
ADAMTSL1	Involved in embryonic gonadogenesis	Female	Carré et al. (2011)
AIRE	Mutations result in autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) which can lead to infertility	Both	Aaltonen et al. (1997)
BMP5	Predicted to play a role in ovarian folliculogenesis	Female	Pierre et al. (2005)
CHD7	Mutations result in CHARGE syndrome (pubertal failure and infertility)	Both	Kim et al. (2008)
CLU	Increased expression results in reduced sperm quality and infertility	Male	Zalata et al. (2012)
CYP19A1	A key enzyme in oestrogen biosynthesis and influences female fertility	Female	Simpson et al. (1994); Altmäe et al. (2009)
DIAPH2	Important for normal ovarian development and function	Female	Bione et al. (1998)
DZIP1	Regulator of hedgehog signalling and may participate in spermatogenesis via its interaction with *DAZ*	Male	Moore et al. (2004); Sekimizu et al. (2004)
IRS4	Null mutations can lead to defects in reproduction	Both	Fantin et al. (2000)
KIT	Plays a key role in germ cell development, spermatogenesis and oogenesis	Both	Rossi, 2013; Russell et al. (2015)
LEP	Deficiencies can lead to hypogonadotrophic hypogonadism and infertility	Both	Chehab et al. (1996)
NANOG	Transcription regulator important for embryonic stem cell pluripotency	Both	Pan and Thomson (2007)
PIP	Functions in seminal fluid, important for fertilization	Male	Hassan et al. (2009)
PRDM14	Required for the proper initiation and coordination of the primordial germ cell specific gene expression programme and promotes pluripotency	Both	Hohenauer and Moore (2012)
PTCH1	Mediates hedgehog signalling in developing and adult marsupial gonads	Both	O'Hara et al. (2011)
PTCH2	Mediates hedgehog signalling in developing and adult marsupial gonads	Both	O'Hara et al. (2011)
PTGFRN	Inhibitor of the Prostaglandin F2 Receptor which has multiple roles in reproduction (e.g., progesterone synthesis and ovulation)	Female	Craig (1975)
SPACA6	Involved in sperm–oocyte fusion; gene knockouts result in failed fusion	Male	Lorenzetti et al. (2014)

of individuals successfully genotyped at each SNP) was 82% (77% when excluding the nine known offspring) (Table S5). All nonsynonymous SNPs conformed to Hardy–Weinberg expectations (Table S5).

Haplotypes at 18 of the 20 genes were consistent with the known trio information. *DIAPH2* showed inconsistencies in five sire–dam–offspring trios (offspring haplotypes were not observed in the parents), possibly due to sequence complexity or particular motifs in this gene region resulting in sequencing difficulty (Nakamura et al., 2011). This gene was excluded from further analysis due to the high error rate (25% of SNPs were inconsistent across the nine trios). *PIP* showed two occurrences of trio phasing inconsistency but was included in subsequent analysis due to the low error rate (2.2% of SNPs were inconsistent across the nine trios). This resulted in 19 final genes (following exclusion of *DIAPH2*) that were included in subsequent population diversity analysis.

The total number of SNPs (both synonymous and nonsynonymous) in the coding regions of each of the 19 final genes across the 37 resequenced genomes ranged from one to 10; the number of haplotypes per gene ranged from two to four (Table 2). Mean haplotype diversity was 0.36 (*SD* 0.20) and mean nucleotide diversity was $4.3 \times 10^{--4}$ (*SD* 5.4×10^{-4}) (Table 2).

ADAMTS9 and *NANOG* showed statistically significant deviation from neutrality at the sequence level with positive Tajima's *D* values suggesting population decline or balancing selection (Table 2). *ADAMTS9* also showed evidence of purifying selection at the codon level with a statistically significant negative *Z*-test ($p < .01$; Table 2).

TABLE 2 Diversity statistics and neutrality tests performed on the target reproductive genes

Gene	*n*	CDS length (bp)	SNPs (ns:s)	*h*	*hd*	*π*	Tajima's *D*	Z-test
ADAMTS9	74	5,919	9 (1:8)	4	0.666	7.32	3.52***	−2.63**
ADAMTS10	74	3,342	1 (1:0)	2	0.104	0.31	−0.60	0.98
ADAMTSL1	74	5,298	1 (1:0)	2	0.294	0.55	0.53	1.00
AIRE	74	1,590	10 (4:6)	4	0.451	21.71	1.83	−1.73
BMP5	74	1,368	1 (1:0)	2	0.053	0.39	−0.90	1.04
CHD7	74	9,093	3 (3:0)	3	0.586	1.15	1.34	1.27
CLU	74	1,178	2 (2:0)	3	0.445	4.06	0.27	1.29
CYP19A1	74	1,512	1 (1:0)	2	0.053	0.35	−0.90	1.07
DZIP1	74	2,433	3 (1:2)	3	0.283	3.08	0.41	−1.26
IRS4	74	2,751	1 (1:0)	2	0.053	0.19	−0.90	1.06
KIT	74	2,901	3 (2:1)	4	0.545	3.76	1.47	−0.67
LEP	74	504	1 (1:0)	2	0.217	4.30	0.07	1.04
NANOG	74	936	2 (2:0)	2	0.494	10.55	2.30*	1.01
PIP	74	534	3 (3:0)	4	0.588	12.54	0.17	0.16
PRDM14	74	1,662	2 (1:1)	2	0.217	2.61	0.09	−0.69
PTCH1	74	3,891	1 (1:0)	2	0.344	0.88	0.82	1.01
PTCH2	74	4,524	3 (3:0)	3	0.527	2.34	1.37	1.50
PTGFRN	74	2,892	2 (2:0)	3	0.416	1.54	0.14	1.40
SPACA6	74	1,122	1 (1:0)	2	0.462	4.12	1.53	1.02

Note: n, number of sequences (two allele sequences per individual); *h*, number of inferred haplotypes; *hd*, haplotype diversity; *π*, nucleotide diversity ($\times 10^4$); ns:s, nonsynonymous:synonymous.

*$p < .05$. Did not remain significant after Holm–Bonferroni multiple test correction.

**$p < .01$. Did not remain significant after Holm–Bonferroni multiple test correction.

***$p < .001$. Remained significant after Holm–Bonferroni multiple test correction.

There were no qualitative changes to the results when the nine known offspring were excluded from the analyses (Table S6).

4 | DISCUSSION

As wildlife populations continue to decline globally, understanding the genetic basis of interindividual variation is crucial for determining which genes may govern important phenotypes and contribute to species' long-term survival and fitness. Here we show how genomic data can be used to explore functional genetic diversity in an endangered species. This study identified a surprising amount of putatively functional variation at reproductive genes in an otherwise genetically depauperate species. Tasmanian devils have shown concerning declines in productivity over time in both captivity (Farquharson et al., 2017) and the wild (Farquharson et al., 2018). It is predicted that genetic variation may play a role in such changes (Farquharson et al., 2017; Gooley et al., 2020), although until now there was limited knowledge of whether diversity even exists at their reproductive genes. We characterized genetic variation at 214 reproductive genes in 37 Tasmanian devils and identified 5,933 putative SNPs. Signatures of selection were examined at a subset of 19 target genes that contained nonsynonymous variation, and hence may have functional consequences for reproduction. To the best of our knowledge, this is the first study to examine within-species reproductive gene diversity to this extent in a threatened species.

Tasmanian devils exhibit very low levels of genetic diversity overall (Cheng et al., 2012; Jones et al., 2004; Miller et al., 2011; Morris et al., 2015). Most (77%) of the reproductive genes we examined had monomorphic coding regions in our sample set of 37 resequenced genomes: a low level of diversity that is comparable to that seen in a previous study which examined genetic diversity at 167 immune genes in 10 Tasmanian devils (seven of which were included in the current study) (Morris et al., 2015). However, within those reproductive genes that showed nonsynonymous variation, we found surprisingly high diversity relative to a similar subset of immune genes that also contained nonsynonymous SNPs (Morris et al., 2015). For example, despite a much larger sample size of up to 196 individuals across multiple captive and wild populations (with the majority of individuals presumed to be unrelated), Morris et al. (2015) found a maximum of three SNPs per gene across nine polymorphic immune genes, compared with a maximum of 10 SNPs per reproductive gene here (across the final 19 polymorphic reproductive genes). Mean haplotype diversity was also higher in the current study. Differences in sample origin may contribute to the observed increased levels of diversity herein; however, the finding of higher genetic diversity at reproductive genes compared with immune genes is unexpected given the smaller sample size and presence of related individuals within the current study. We note that Morris et al. (2015) used amplicon sequencing to confirm SNP diversity in the subset of target genes, which resulted in fewer SNPs than predicted by genome resequencing data. Although we did not employ gene-targeted sequencing methods in this study, we believe that the SNPs identified are likely to reflect real diversity, not sequencing artefacts, due to the number of resequenced genomes (particularly those with high coverage, around 45×) and the strict variant calling and filtering parameters employed.

Thirty-six reproductive genes (14% of all genes investigated) present in model species could not be characterized in the Tasmanian devil genome by the methods applied here. For example, there were no TBLASTN hits for a number of genes including *DPPA3/STELLA*, *SEMG1*, *SEMG2*, *TNP2* and *PRM2*, which are either too divergent from known orthologues to be identified by this method, or do not exist in marsupials (Johnson et al., 2018). Additionally, members of the *NLRP* (nucleotide-binding oligomerization domain, leucine-rich repeat and pyrin domain-containing proteins) gene family have shown extensive duplication and diversification in mammalian lineages (Tian et al., 2009) and were unable to be identified in the Tasmanian devil genome. Fragmentation and gaps in the current reference genome precluded characterizing a number of genes such as *KLK3* and ZPBP (Table S2). Genes located on the Y chromosome (e.g., *ATRY*, *DAZ1*, *USP9Y* and *DDX3Y*) could not be identified due to the unavailability of Y-chromosome data in the female reference genome. Sequencing the Y chromosome will be important in the future to focus on male reproduction, as a number of important male reproductive genes are found on the Y chromosome (Murtagh et al., 2010; Toder et al., 2000).

Twenty genes were found to contain nonsynonymous SNPs in the current study (with *DIAPH2* later excluded due to phasing inconsistencies). Since nonsynonymous mutations result in amino acid changes, genes that contain nonsynonymous SNPs may influence phenotype (Shastry, 2009). Although other SNPs, such as synonymous polymorphisms or variants outside the coding sequence, may contribute to phenotype via processes such as mRNA stability (Chamary & Hurst, 2005), these are expected to have a weaker effect on gene function compared with mutations that alter the protein sequence (Tomoko, 1995). The genes found to contain nonsynonymous SNPs in the current study are involved in a variety of reproductive functions in both males and females, and influence fertility-associated phenotypes in humans and other species (see Table 1 for more information). For example, mutations in the *CHD7* gene cause idiopathic hypogonadotropic hypogonadism and Kallmann syndrome in humans, resulting in impaired sexual development in both males and females (Kim et al., 2008). Mutations in the *AIRE* gene cause autoimmune polyendocrinopathy, candidiasis and ectodermal dystrophy (APECED) (Aaltonen et al., 1997), which has also been linked to infertility in both men and women (Perheentupa, 2006). ADAMTS proteases influence a range of reproductive processes in humans and mice (Russell et al., 2015), three of which (*ADAMTS9*, *ADAMTS10* and *ADAMTSL1*) displayed nonsynonymous variation in the current study.

The majority of the individuals in our sample set are known to have successfully reproduced based on breeding records in captive facilities (Figure S2), so most of the nonsynonymous SNPs identified in the current study are unlikely to cause the extreme infertile phenotypes that have been reported in humans and mice. However, these variants may result in more subtle phenotypic effects such as reduced fertilization success or reduced offspring survival. We note that a number of nonsynonymous homozygoyte genotypes were not observed in our data set (Table S5). They may encode more severe phenotypes which could be associated with pregnancy loss or infertility and may exist in a larger sample set or could potentially be lethal and hence never appear in homozygous form. Further research is required to explore the functional consequences of the identified nonsynonymous variants herein. Interestingly, we found that genes involved in embryogenesis, fertilization and hormonal regulation of reproduction displayed greater numbers of nonsynonymous SNPs than synonymous SNPs. This suggests that functional diversity may be important at genes involved in such processes. Tasmanian devils exhibit a number of unique reproductive characteristics including undergoing up to three oestrous cycles within their annual breeding season (Keeley et al., 2012); producing a greater number of embryos (up to 30) than can be supported by their four teats (Guiler, 1970; Hughes, 1982); and multiple paternity litters (Russell et al., 2019) even though mate-guarding is a behavioural reproductive strategy (Hamilton et al., 2019). We hypothesize that these unique reproductive traits may drive functional diversity across genes involved in particular reproductive processes through adaptive evolution. For example, multiple mating by females is known to drive sperm competition, which may result in selective pressures on genes involved in fertilization (Dapper & Wade, 2016; Fiumera et al., 2005). Similarly, fitness advantages associated with the timing or number of oestrous cycles, or the number of viable embryos, could potentially drive natural selection at genes involved in the hormonal regulation of reproduction or embryogenesis respectively. To explore these ideas further we investigated signatures of selection at the reproductive genes containing nonsynonymous SNPs.

Of the 19 final reproductive genes (following exclusion of *DIAPH2* due to phasing inconsistencies), two genes (*ADAMTS9* and *NANOG*) showed statistically significant signatures of selection, suggesting their variants may be linked to important phenotypic traits. After correcting for multiple testing using the Holm–Bonferroni method (Holm, 1979), the Tajima's *D* for *ADAMTS9* remained statistically significant, indicating that this gene may be under balancing selection at the sequence level within the population. Demographic factors such as population bottlenecks can contribute to the value of Tajima's *D* (Tajima, 1989), although demographic factors are likely to affect loci across the whole genome. Because similar patterns of selection were not observed across all of the target loci, we hypothesize that *ADAMTS9* may be a candidate for long-term balancing selection. Balancing selection actively maintains multiple alleles in a population, suggesting that the associated phenotypes may be advantageous under certain circumstances (e.g., Gos et al., 2012). *ADAMTS9* is a pleiotropic gene that belongs to a large, diversified family of *ADAMTS* genes and has been implicated in several crucial female reproductive processes, namely: ovulation, implantation, placentation and parturition (Russell et al., 2015). *ADAMTS9* is also a novel tumour suppressor (Du et al., 2013) and has undergone strong selection for increased longevity in a number of small-bodied mammal lineages (Lambert & Portfors, 2017). This is particularly interesting in our context, as Tasmanian devils have a short lifespan (maximum 5 years in the wild) in comparison to other mammals of their size and show unusually high vulnerability to tumours (Griner, 1979). However, our data cannot disentangle whether potential selection on the *ADAMTS9* gene in Tasmanian devils may be attributed to that gene's role in reproduction and/or its role in tumour suppression and longevity. The attributes of this gene, such as its role in a number of key processes, make it a plausible candidate for adaptation and warrants further investigation.

The *NANOG* gene also showed a putative pattern of balancing selection in the Tasmanian devil, although this result did not remain statistically significant after correcting for multiple testing. *NANOG* is a key transcription factor involved in embryonic stem cell pluripotency (Pan & Thomson, 2007). We identified a multinucleotide nonsynonymous polymorphism within the CDS of *NANOG*. It is currently unknown whether these variants are associated with differential phenotypes. Investigations into whether the identified nonsynonymous SNP is correlated with embryonic survival traits and may influence reproductive success within the Tasmanian devil are required.

Although reproductive genes and their variants have been well studied in model and livestock species (see Hunt et al., 2018), there are few data on reproductive variants in threatened species, many of which typically show low overall levels of genome-wide diversity. As a result, it is difficult to ascertain whether Tasmanian devil reproductive gene diversity is higher or lower than expected compared to other threatened species. Furthermore, our study focused on a relatively small sample set from a limited number of locations in Tasmania and may not have captured the true extent of genetic diversity across the species' range. As whole genome sequencing technology becomes cheaper with time, sampling Tasmanian devils across their range would improve our understanding of their reproductive gene diversity. The full benefit of understanding reproductive gene diversity in Tasmanian devils can be realized by studying the relationship between genetic variation and reproductive phenotypes. For this threatened species, this is possible as the Tasmanian devil insurance population is Australia's largest captive breeding programme (Hogg et al., 2019) with a large number of individuals across multiple generations with DNA samples and extensive reproductive records. This resource will allow us to investigate diversity across a range of candidate genes to determine whether variation in reproductive genes influences reproductive fitness. For example, the *SPACA6* gene has been implicated in fertilization ability of male mice (Lorenzetti et al., 2014) and was found to contain a nonsynonymous SNP among the sampled Tasmanian devils in the current study. By sequencing the *SPACA6* gene across hundreds of male Tasmanian devils using specific PCR primers, or a targeted capture approach, we could statistically determine whether this variant is correlated with an individual's siring ability. Candidate gene approaches have several advantages over whole-genome approaches, namely the higher inherent statistical power and reduced sequencing costs. However, it is possible that other genes or genomic regions that influence reproductive phenotypes may be missed and so a genome-wide association study (GWAS) may be more informative (for a review of candidate gene vs. GWAS approaches see Suh & Vijg, 2005). A combination of these approaches will probably be the best way forward to understanding the interplay between reproductive genotype and phenotype. The rise of whole genome sequencing and global consortia developing reference genomes for wildlife means that our understanding of functional gene diversity in a range of threatened species can only improve with time, particularly in those species where range reduction and population contraction has led them to be genetically depauperate. The approach used in this study demonstrates how these growing genomic resources can be utilized to explore functional diversity in threatened species and how this information can assist with their conservation management.

5 | CONCLUSION

Our study has bioinformatically characterized diversity at 219 reproductive genes in 37 Tasmanian devils. We have identified and examined diversity at 19 polymorphic genes containing nonsynonymous SNPs that may have functional consequences on reproduction. The results from this study provide the foundation for future research to explore whether any of these genes are associated with variable reproductive phenotypes and hence may be involved in the generational productivity declines that have been observed in the Tasmanian devil insurance population (Farquharson et al., 2017; Hogg et al., 2015). If specific genotypes are found to influence productivity, preserving the functional variation described herein may be key to minimizing these declines and facilitating the success of conservation breeding programmes. Beyond assisting with conservation decisions for the Tasmanian devil, the candidate gene approach described here may also be applied to reproductive management in other threatened species conservation programmes.

ACKNOWLEDGEMENTS

This work was funded by the Australian Research Council (LP140100508; DP170101253) and supported by the Save the Tasmanian Devil Program, the Zoo and Aquarium Association Australasia and San Diego Zoo Global. We thank Professor Graham Wallis for insightful discussions on methodology and general themes of the paper. Dr Camilla Whittington provided helpful suggestions of genes important to reproduction. We thank Drew Lee for creating the supplemental map overlay. Computing resources were provided by the Sydney Informatics Hub and the University of Sydney's high-performance computing cluster Artemis.

AUTHOR CONTRIBUTIONS

K.B, C.E.G and C.J.H contributed to the design of the study and sourced funding. P.B compiled the list of target genes and performed bioinformatic gene characterisation and SNP prediction. B.W performed alignments of resequencing data to the reference genome and assisted P.B with the analysis of SNP genotypes. P.B wrote the manuscript with contributions and editing from B.W, C.E.G, C.J.H and K.B. All authors read and approved the final manuscript.

DATA AVAILABILITY

The Tasmanian Devil reference genome and associated data are available from NCBI:

- Devil_ref version 7.0 GenBank: GCA_000189315.1 (Murchison et al., 2012)

Raw genome resequencing reads are available from the referenced NCBI SRA accessions:

- 25 high-coverage: SRX6096677–SRX6096696 (Wright et al., 2020).
- 12 low-coverage: ERS682204–ERS682210; ERS1202857–ERS1202861 (Wright et al., 2015, 2017).

The multisample VCF file for the reproductive genes of interest has been deposited in Dryad

- Dryad https://doi.org/10.5061/dryad.t1g1jwt10

ORCID

Parice A. Brandies https://orcid.org/0000-0003-1702-2938
Belinda R. Wright https://orcid.org/0000-0002-3317-8185
Carolyn J. Hogg https://orcid.org/0000-0002-6328-398X
Catherine E. Grueber https://orcid.org/0000-0002-8179-1822
Katherine Belov https://orcid.org/0000-0002-9762-5554

REFERENCES

Aaltonen, J., Björses, P., Perheentupa, J., Horelli–Kuitunen, N., Palotie, A., Peltonen, L., Lee, Y. S., Francis, F., Henning, S., Thiel, C., Leharach, H., & Yaspo, M. L. (1997). An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nature Genetics*, *17*(4), 399–403. https://doi.org/10.1038/ng1297-399

Altmäe, S., Haller, K., Peters, M., Saare, M., Hovatta, O., Stavreus-Evers, A., Velthut, A., Karro, H., Metspalu, A., & Salumets, A. (2009). Aromatase gene (CYP19A1) variants, female infertility and ovarian stimulation outcome: A preliminary report. *Reproductive Biomedicine Online*, *18*(5), 651–657. https://doi.org/10.1016/s1472-6483(10)60009-0

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bione, S., Sala, C., Manzini, C., Arrigo, G., Zuffardi, O., Banfi, S., Borsani, G., Jonveaux, P., Philippe, C., Zuccotti, M., Ballabio, A., & Toniolo, D. (1998). A human homologue of the Drosophila melanogaster diaphanous gene is disrupted in a patient with premature ovarian failure: Evidence for conserved function in oogenesis and implications for human sterility. *The American Journal of Human Genetics*, *62*(3), 533–541. https://doi.org/10.1086/301761

Brandies, P. A., Grueber, C. E., Hogg, C. J., & Belov, K. (2018). MHC genes and mate choice. In J. Choe (Ed.), *Encyclopedia of animal behaviour*, 2nd ed. Elsevier.

Carré, G.-A., Couty, I., Hennequet-Antier, C., & Govoroun, M. S. (2011). Gene expression profiling reveals new potential players of gonad differentiation in the chicken embryo. *PLoS One*, *6*(9), e23959. https://doi.org/10.1371/journal.pone.0023959

Chamary, J., & Hurst, L. D. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology*, *6*(9), R75. https://doi.org/10.1186/gb-2005-6-9-r75

Chehab, F. F., Lim, M. E., & Lu, R. (1996). Correction of the sterility defect in homozygous obese female mice by treatment with the human recombinant leptin. *Nature Genetics*, *12*(3), 318–320. https://doi.org/10.1038/ng0396-318

Cheng, A. Y., Teo, Y.-Y., & Ong, R.-T.-H. (2014). Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, *30*(12), 1707–1713. https://doi.org/10.1093/bioinformatics/btu067

Cheng, Y., Sanderson, C., Jones, M., & Belov, K. (2012). Low MHC class II diversity in the Tasmanian devil (Sarcophilus harrisii). *Immunogenetics*, *64*(7), 525–533. https://doi.org/10.1007/s00251-012-0614-4

Consortium, U (2018). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515.

Craig, G. M. (1975). Prostaglandins in reproductive physiology. *Postgraduate Medical Journal*, *51*(592), 74–84. https://doi.org/10.1136/pgmj.51.592.74

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dapper, A. L., & Wade, M. J. (2016). The evolution of sperm competition genes: The effect of mating system on levels of genetic variation within and between species. *Evolution*, *70*(2), 502–511. https://doi.org/10.1111/evo.12848

Du, W., Wang, S., Zhou, Q., Li, X., Chu, J., Chang, Z., Tao, Q., Ng, E. K. O., Fang, J., Sung, J. J. Y., & Yu, J. (2013). ADAMTS9 is a functional tumor suppressor through inhibiting AKT/mTOR pathway and associated with poor survival in gastric cancer. *Oncogene*, *32*(28), 3319–3328. https://doi.org/10.1038/onc.2012.359

Dun, M. D., Anderson, A. L., Bromfield, E. G., Asquith, K. L., Emmett, B., McLaughlin, E. A., Aitken, R. J., & Nixon, B. (2012). Investigation of the expression and functional significance of the novel mouse sperm protein, a disintegrin and metalloprotease with thrombospondin type 1 motifs number 10 (ADAMTS10). *International Journal of Andrology*, *35*(4), 572–589. https://doi.org/10.1111/j.1365-2605.2011.01235.x

Fantin, V. R., Wang, Q., Lienhard, G. E., & Keller, S. R. (2000). Mice lacking insulin receptor substrate 4 exhibit mild defects in growth, reproduction, and glucose homeostasis. *American Journal of Physiology-Endocrinology and Metabolism*, *278*(1), E127–E133. https://doi.org/10.1152/ajpendo.2000.278.1.E127

Farquharson, K. A., Gooley, R. M., Fox, S., Huxtable, S. J., Belov, K., Pemberton, D., Hogg, C. J., & Grueber, C. E. (2018). Are any populations 'safe'? Unexpected reproductive decline in a population of Tasmanian devils free of devil facial tumour disease. *Wildlife Research*, *45*(1), 31–37. https://doi.org/10.1071/WR16234

Farquharson, K. A., Hogg, C. J., & Grueber, C. E. (2017). Pedigree analysis reveals a generational decline in reproductive success of captive Tasmanian devil (Sarcophilus harrisii): Implications for captive management of threatened species. *Journal of Heredity*, *108*(5), 488–495. https://doi.org/10.1093/jhered/esx030

Fiumera, A. C., Dumont, B. L., & Clark, A. G. (2005). Sperm competitive ability in Drosophila melanogaster associated with variation in male reproductive proteins. *Genetics*, *169*(1), 243–257. https://doi.org/10.1534/genetics.104.032870

Flot, J. F. (2010). SeqPHASE: A web tool for interconverting PHASE input/output files and FASTA sequence alignments. *Molecular Ecology Resources*, *10*(1), 162–166. https://doi.org/10.1111/j.1755-0998.2009.02732.x

Forsman, A., & Wennersten, L. (2016). Inter-individual variation promotes ecological success of populations and species: Evidence from experimental and comparative studies. *Ecography*, *39*(7), 630–648. https://doi.org/10.1111/ecog.01357

Gooley, R. M., Hogg, C. J., Fox, S., Pemberton, D., Belov, K., & Grueber, C. E. (2020). Inbreeding depression in one of the last DFTD-free wild populations of Tasmanian devils. *PeerJ*, *8*, e9220. https://doi.org/10.7717/peerj.9220

Gos, G., Slotte, T., & Wright, S. I. (2012). Signatures of balancing selection are maintained at disease resistance loci following mating system evolution and a population bottleneck in the genus Capsella. *BMC Evolutionary Biology*, *12*(1), 152. https://doi.org/10.1186/1471-2148-12-152

Griner, L. (1979). Neoplasms in Tasmanian devils (Sarcophilus harrisii). *Journal of the National Cancer Institute*, *62*(3), 589–595. https://doi.org/10.1093/jnci/62.3.589

Guiler, E. (1970). Observations on the Tasmanian devil, *Sarcophilus harrisii* (Marsupialia: Dasyuridae) II. Reproduction, breeding and growth of pouch young. *Australian Journal of Zoology*, *18*(1), 63–70.

Hamilton, D. G., Jones, M. E., Cameron, E. Z., McCallum, H., Storfer, A., Hohenlohe, P. A., & Hamede, R. K. (2019). Rate of intersexual interactions affects injury likelihood in Tasmanian devil contact networks. *Behavioral Ecology*, *30*(4), 1087–1095. https://doi.org/10.1093/beheco/arz054

Hassan, M. I., Waheed, A., Yadav, S., Singh, T., & Ahmad, F. (2009). Prolactin inducible protein in cancer, fertility and immunoregulation: Structure, function and its clinical implications. *Cellular and Molecular Life Sciences*, *66*(3), 447–459. https://doi.org/10.1007/s00018-008-8463-x

Hogg, C. J., Fox, S., Pemberton, D., & Belov, K. (2019). In C. J. Hogg, S. Fox, D. Pemberton, & K. Belov (Eds.), *Saving the Tasmanian Devil*. CSIRO Publishing.

Hogg, C. J., Ivy, J. A., Srb, C., Hockley, J., Lees, C., Hibbard, C., & Jones, M. (2015). Influence of genetic provenance and birth origin on productivity of the Tasmanian devil insurance population. *Conservation Genetics*, *16*(6), 1465–1473. https://doi.org/10.1007/s10592-015-0754-9

Hohenauer, T., & Moore, A. W. (2012). The Prdm family: Expanding roles in stem cells and development. *Development*, *139*(13), 2267–2282. https://doi.org/10.1242/dev.070110

Holderegger, R., Kamm, U., & Gugerli, F. (2006). Adaptive vs. neutral genetic diversity: Implications for landscape genetics. *Landscape Ecology*, *21*(6), 797–807.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Hughes, R. (1982). Reproduction in the Tasmanian devil *Sarcophilus harrisii* (Dasyuridae, Marsupialia). *Carnivorous Marsupials*, *1*, 49–63.

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Flicek, P. (2018). Ensembl variation resources. *Database*, *2018*, bay119.

Johnson, R. N., O'Meally, D., Chen, Z., Etherington, G. J., Ho, S. Y. W., Nash, W. J., Grueber, C. E., Cheng, Y., Whittington, C. M., Dennison, S., Peel, E., Haerty, W., O'Neill, R. J., Colgan, D., Russell, T. L., Alquezar-Planas, D. E., Attenbrow, V., Bragg, J. G., Brandies, P. A., … Belov, K. (2018). Adaptation and conservation insights from the koala genome. *Nature Genetics*, *50*(8), 1102–1111. https://doi.org/10.1038/s41588-018-0153-5

Jones, M. E., Paetkau, D., Geffen, E., & Moritz, C. (2004). Genetic diversity and population structure of Tasmanian devils, the largest marsupial carnivore. *Molecular Ecology*, *13*(8), 2197–2209. https://doi.org/10.1111/j.1365-294X.2004.02239.x

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, *16*(5), 1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Kapustin, Y., Souvorov, A., Tatusova, T., & Lipman, D. (2008). Splign: Algorithms for computing spliced alignments with identification of paralogs. *Biology Direct*, *3*(1), 20. https://doi.org/10.1186/1745-6150-3-20

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Keeley, T., O'Brien, J., Fanson, B., Masters, K., & McGreevy, P. (2012). The reproductive cycle of the Tasmanian devil (*Sarcophilus harrisii*) and factors associated with reproductive success in captivity. *General and Comparative Endocrinology*, *176*(2), 182–191. https://doi.org/10.1016/j.ygcen.2012.01.011

Kim, H.-G., Kurth, I., Lan, F., Meliciani, I., Wenzel, W., Eom, S. H., Kang, G. B., Rosenberger, G., Tekin, M., Ozata, M., Bick, D. P., Sherins, R. J., Walker, S. L., Shi, Y., Gusella, J. F., & Layman, L. C. (2008). Mutations in CHD7, encoding a chromatin-remodeling protein, cause idiopathic hypogonadotropic hypogonadism and Kallmann syndrome. *The American Journal of Human Genetics*, *83*(4), 511–519. https://doi.org/10.1016/j.ajhg.2008.09.005

Kirkpatrick, B. (2002). QTL and candidate gene effects on reproduction in livestock: progress and prospects. Paper presented at the Proceedings of the 7th Word Congress on Genetics Applied to Livestock Production, August.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. https://doi.org/10.1093/molbev/msw054

Lachish, S., McCallum, H., & Jones, M. (2009). Demography, disease and the devil: Life-history changes in a disease-affected population of Tasmanian devils (*Sarcophilus harrisii*). *Journal of Animal Ecology*, *78*(2), 427–436.

Lambert, M. J., & Portfors, C. V. (2017). Adaptive sequence convergence of the tumor suppressor ADAMTS9 between small-bodied mammals displaying exceptional longevity. *Aging (Albany NY)*, *9*(2), 573–582. https://doi.org/10.18632/aging.101180

Layman, L. C. (2002). Human gene mutations causing infertility. *Journal of Medical Genetics*, *39*(3), 153–161. https://doi.org/10.1136/jmg.39.3.153

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lorenzetti, D., Poirier, C., Zhao, M., Overbeek, P. A., Harrison, W., & Bishop, C. E. (2014). A transgenic insertion on mouse chromosome 17 inactivates a novel immunoglobulin superfamily gene potentially involved in sperm–egg fusion. *Mammalian Genome*, *25*(3–4), 141–148. https://doi.org/10.1007/s00335-013-9491-x

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Miller, W., Hayes, V. M., Ratan, A., Petersen, D. C., Wittekindt, N. E., Miller, J., Walenz, B., Knight, J., Qi, J., Zhao, F., Wang, Q., Bedoya-Reina, O. C., Katiyar, N., Tomsho, L. P., Kasson, L. M., Hardie, R.-A., Woodbridge, P., Tindall, E. A., Bertelsen, M. F., … Schuster, S. C. (2011). Genetic diversity and population structure of the endangered marsupial *Sarcophilus harrisii* (Tasmanian devil). *Proceedings of the National Academy of Sciences of the United States of America 108*(30), 12348–12353.

Mimura, M., Yahara, T., Faith, D. P., Vázquez-Domínguez, E., Colautti, R. I., Araki, H., Javadi, F., Núñez-Farfán, J., Mori, A. S., Zhou, S., Hollingsworth, P. M., Neaves, L. E., Fukano, Y., Smith, G. F., Sato, Y.-I., Tachida, H., & Hendry, A. P. (2017). Understanding and monitoring the consequences of human impacts on intraspecific variation. *Evolutionary Applications*, *10*(2), 121–139. https://doi.org/10.1111/eva.12436

Moore, F. L., Jaruzelska, J., Dorfman, D. M., & Reijo-Pera, R. A. (2004). Identification of a novel gene, DZIP (DAZ-interacting protein), that encodes a protein that interacts with DAZ (deleted in azoospermia) and is expressed in embryonic stem cells and germ cells. *Genomics*, *83*(5), 834–843. https://doi.org/10.1016/j.ygeno.2003.11.005

Morris, K. M., Wright, B., Grueber, C. E., Hogg, C., & Belov, K. (2015). Lack of genetic diversity across diverse immune genes in an endangered mammal, the Tasmanian devil (*Sarcophilus harrisii*). *Molecular Ecology*, *24*(15), 3860–3872. https://doi.org/10.1111/mec.13291

Murchison, E. P., Schulz-Trieglaff, O. B., Ning, Z., Alexandrov, L. B., Bauer, M. J., Fu, B., Hims, M., Ding, Z., Ivakhno, S., Stewart, C., Ng, B. L., Wong, W., Aken, B., White, S., Alsop, A., Becq, J., Bignell, G. R., Cheetham, R. K., Cheng, W., … Stratton, M. R. (2012). Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell*, *148*(4), 780–791. https://doi.org/10.1016/j.cell.2011.11.065

Murtagh, V. J., Waters, P. D., & Graves, J. A. M. (2010). Compact but Complex-The Marsupial Y Chromosome. In J.E Deakin P.D. Waters & J.A.M. Graves (Eds.), *Marsupial genetics and genomics* (pp. 207–228). Springer.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M. D., Ogasawara, N., & Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, *39*(13), e90. https://doi.org/10.1093/nar/gkr344

NCBI Resource Coordinators (2017). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *45*(Database issue), D12.

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

O'Hara, W. A., Azar, W. J., Behringer, R. R., Renfree, M. B., & Pask, A. J. (2011). Desert hedgehog is a mammal-specific gene expressed during testicular and ovarian development in a marsupial. *BMC Developmental Biology*, *11*(1), 72. https://doi.org/10.1186/1471-213X-11-72

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., & Ako-Adjei, D. (2015). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745.

Owen, D., & Pemberton, D. (2005). *Tasmanian devil: A unique and threatened animal*. Allen & Unwin.

Pan, G., & Thomson, J. A. (2007). Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Research*, *17*(1), 42–49. https://doi.org/10.1038/sj.cr.7310125

Perheentupa, J. (2006). Autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy. *The Journal of Clinical Endocrinology & Metabolism*, *91*(8), 2843–2850. https://doi.org/10.1210/jc.2005-2611

Pierre, A., Pisselet, C., Dupont, J., Bontoux, M., & Monget, P. (2005). Bone morphogenetic protein 5 expression in the rat ovary: Biological effects on granulosa cell proliferation and steroidogenesis. *Biology of Reproduction*, *73*(6), 1102–1108. https://doi.org/10.1095/biolreprod.105.043091

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rossi, P. (2013). Transcriptional control of KIT gene expression during germ cell development. *International Journal of Developmental Biology*, *57*(2–4), 179–184. https://doi.org/10.1387/ijdb.130014pr

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, *34*(12), 3299–3302. https://doi.org/10.1093/molbev/msx248

Russell, D. L., Brown, H. M., & Dunning, K. R. (2015). ADAMTS proteases in fertility. *Matrix Biology*, *44–46*, 54–63. https://doi.org/10.1016/j.matbio.2015.03.007

Russell, T., Lane, A., Clarke, J., Hogg, C., Morris, K., Keeley, T., Madsen, T., & Ujvari, B. (2019). Multiple paternity and precocial breeding in wild Tasmanian devils, *Sarcophilus harrisii* (Marsupialia: Dasyuridae). *Biological Journal of the Linnean Society*, *128*(1), 201–210. https://doi.org/10.1093/biolinnean/blz072

Sekimizu, K., Nishioka, N., Sasaki, H., Takeda, H., Karlstrom, R. O., & Kawakami, A. (2004). The zebrafish iguana locus encodes Dzip1, a novel zinc-finger protein required for proper regulation of Hedgehog signaling. *Development*, *131*(11), 2521–2532. https://doi.org/10.1242/dev.01059

Shastry, B. S. (2009). SNPs: Impact on gene function and phenotype. In A. Komar (Ed.), *Single Nucleotide polymorphisms* (pp. 3–22). Springer.

Simpson, E. R., Mahendroo, M. S., Means, G. D., Kilgore, M. W., Hinshelwood, M. M., Graham-lorence, S., Amarneh, B., Ito, Y., Fisher, C. R., Michael, M. D., Mendelson, C. R., & Bulun, S. E. (1994). Aromatase cytochrome P450, the enzyme responsible for estrogen biosynthesis. *Endocrine Reviews*, *15*(3), 342–355. https://doi.org/10.1210/edrv-15-3-342

Solovyev, V. (2004). Statistical approaches in Eukaryotic gene prediction. In D. Balding, C. Cannings, & M. Bishop (Eds.), *Handbook of statistical genetics* (3rd ed., pp. 1616). John Wiley & Sons.

Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in Zoology*, *2*(1), 16.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, *54(1), 1.30.31*, - 31.30.33. https://doi.org/10.1002/cpbi.5

Stephens, M., & Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, *73*(5), 1162–1169. https://doi.org/10.1086/379378

Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, *68*(4), 978–989. https://doi.org/10.1086/319501

Suh, Y., & Vijg, J. (2005). SNP discovery in associating genetic variation with human disease phenotypes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *573*(1–2), 41–53. https://doi.org/10.1016/j.mrfmmm.2005.01.005

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Tian, X., Pascal, G., & Monget, P. (2009). Evolution and functional divergence of NLRP genes in mammalian reproductive systems. *BMC Evolutionary Biology*, *9*(202), https://doi.org/10.1186/1471-2148-9-202

Toder, R., Wakefield, M., & Graves, J. (2000). The minimal mammalian Y chromosome–the marsupial Y as a model system. *Cytogenetic and Genome Research*, *91*(1–4), 285–292. https://doi.org/10.1159/000056858

Tomoko, O. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution*, *40*(1), 56–63. https://doi.org/10.1007/BF00166595

Ujvari, B., & Belov, K. (2011). Major histocompatibility complex (MHC) markers in conservation biology. *International Journal of Molecular Sciences*, *12*(8), 5168–5186. https://doi.org/10.3390/ijms12085168

Wright, B. R., Farquharson, K. A., McLennan, E. A., Belov, K., Hogg, C. J., & Grueber, C. E. (2020). A demonstration of conservation genomics for threatened species management. *Molecular Ecology Resources*, *00*, 1–16. https://doi.org/10.1111/1755-0998.13211

Wright, B., Morris, K., Grueber, C. E., Willet, C. E., Gooley, R., Hogg, C. J., O'Meally, D., Hamede, R., Jones, M., Wade, C., & Belov, K. (2015). Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population. *BMC Genomics*, *16*(1), 791. https://doi.org/10.1186/s12864-015-2020-4

Wright, B., Willet, C. E., Hamede, R., Jones, M., Belov, K., & Wade, C. M. (2017). Variants in the host genome may inhibit tumour growth in devil facial tumours: Evidence from genome-wide association. *Scientific Reports*, *7*(1), 423. https://doi.org/10.1038/s41598-017-00439-7

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols*, *10*(10), 1556–1566. https://doi.org/10.1038/nprot.2015.105

Zalata, A., El-Samanoudy, A. Z., Shaalan, D., El-Baiomy, Y., Taymour, M., & Mostafa, T. (2012). Seminal clusterin gene expression associated with seminal variables in fertile and infertile men. *The Journal of Urology*, *188*(4), 1260–1264. https://doi.org/10.1016/j.juro.2012.06.012

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Brandies PA, Wright BR, Hogg CJ, Grueber CE, Belov K. Characterization of reproductive gene diversity in the endangered Tasmanian devil. *Mol Ecol Resour.* 2020;00:1–12. https://doi.org/10.1111/1755-0998.13295

A1.4 THE FIRST ANTECHINUS REFERENCE GENOME PROVIDES A RESOURCE FOR INVESTIGATING THE GENETIC BASIS OF SEMELPARITY AND AGE-RELATED NEUROPATHOLOGIES

The PDF version of the article titled “The first Antechinus reference genome provides a resource for investigating the genetic basis of semelparity and age-related neuropathologies” published in *Gigabyte* (2020; 1(7), 1-22), which comprises Chapter 4 of this thesis, is presented on the following pages.

(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
Revised: 20 October 2020
Accepted: 04 November 2020
Published: 05 November 2020

* 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

Subjects Genetics and Genomics, Animal Genetics, Evolutionary Biology

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

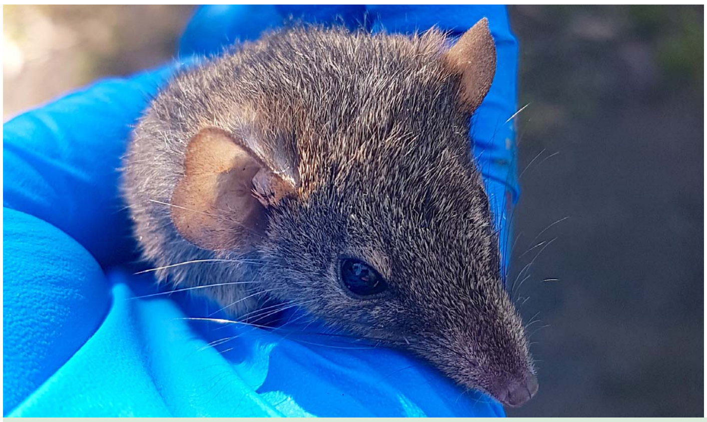


Figure 1. *Antechinus stuartii* individual used for the male reference genome. Image from Carolyn Hogg.

(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 < 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 < 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 (AD-ratio ≤ 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 $\times 10^{-5}$ 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

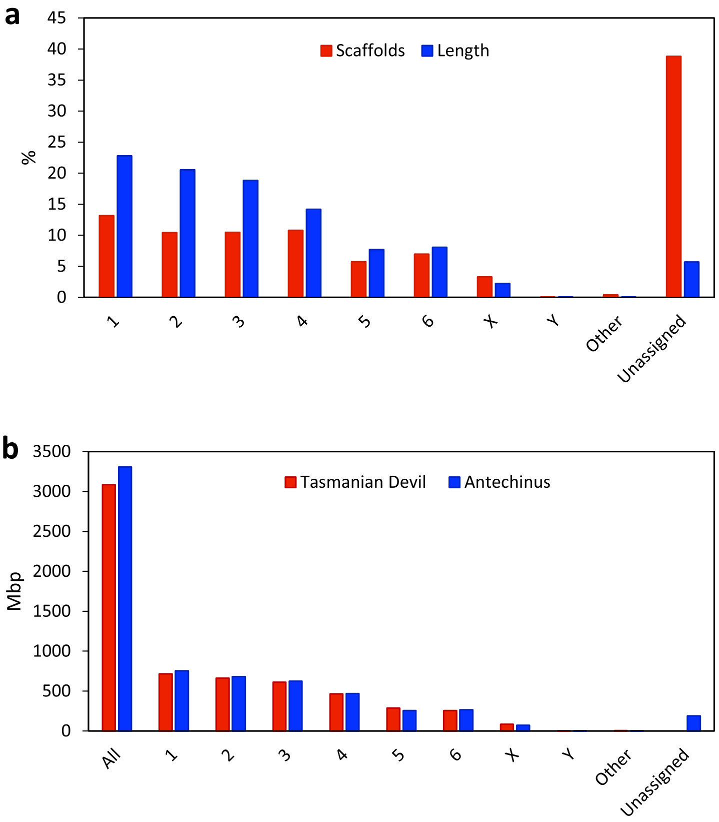


Figure 2. Assignment of antechinus scaffolds to chromosomes by alignment to the Tasmanian devil reference genome. (a) Proportion (%) of scaffolds (blue) and genome length (red) assigned to chromosomes. (b) Comparison of length of sequence assigned to each chromosome from the Tasmanian devil reference 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

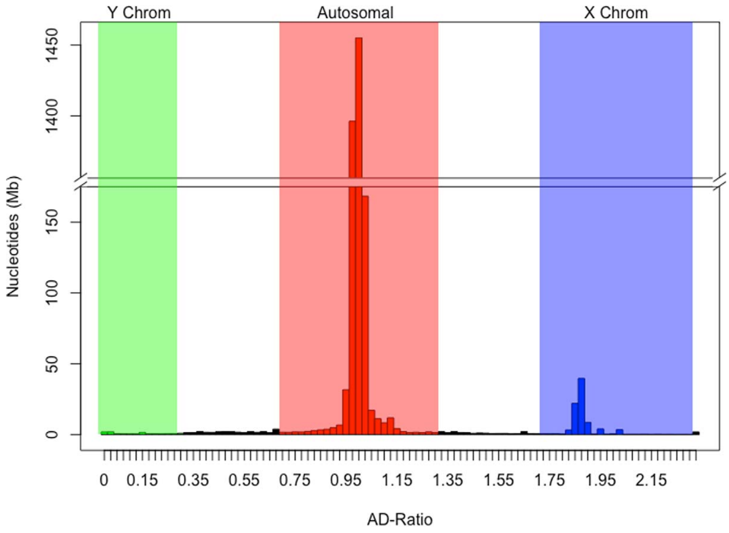


Figure 3. AD-Ratio histogram of antechinus scaffolds. Figure shows the total length of sequence within each 0.025 AD-ratio bin. Scaffolds clustering around an AD-ratio of 0 represent Y-linked sequence (Green), scaffolds clustering around an AD-ratio of 1 represent Autosomal sequence (Red), scaffolds 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

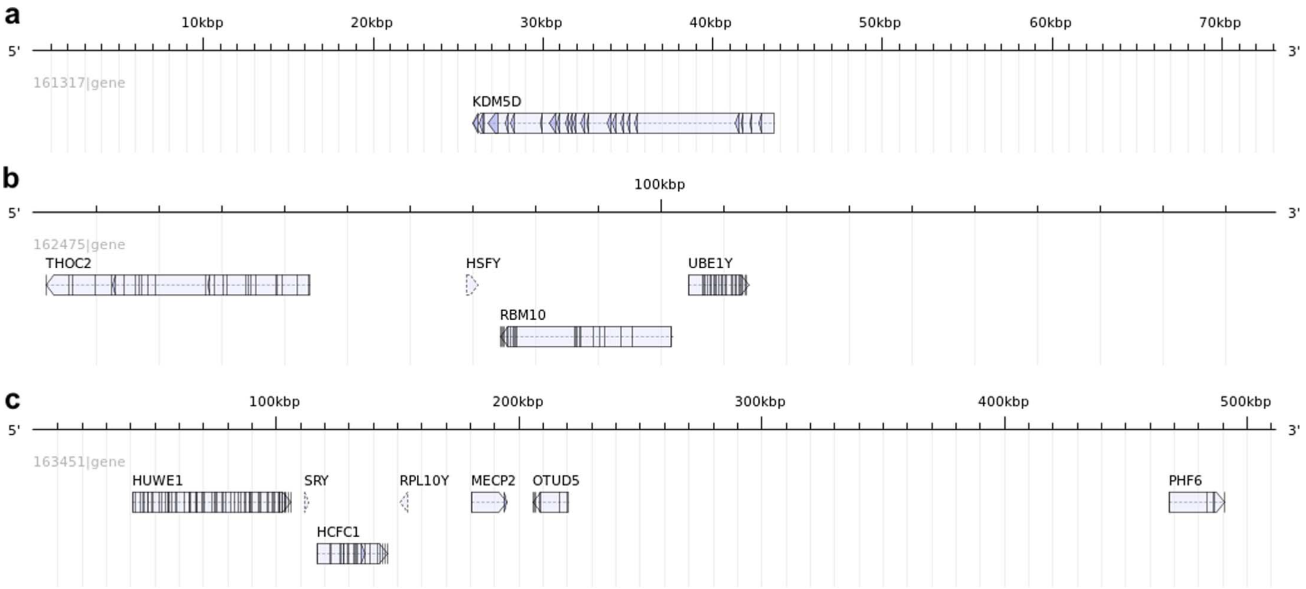


Figure 4. Mapping of known marsupial Y gene homologs on antechinus Y chromosome scaffolds. (a) Scaffold 161317, (b) Scaffold 162475, (c) Scaffold 163451. Figure was created using the AnnotationSketch module from GenomeTools [66].

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

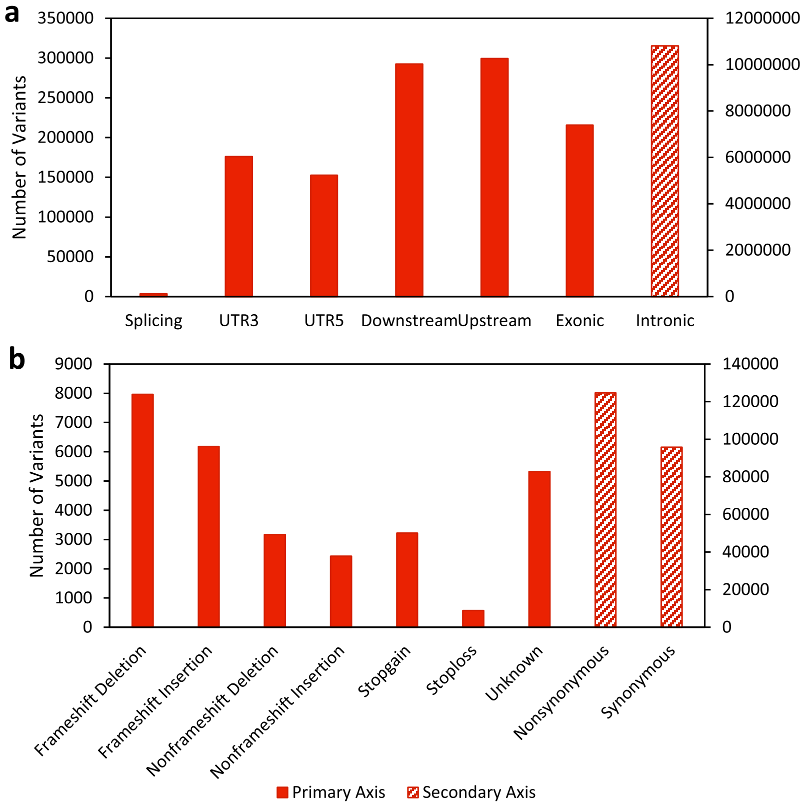


Figure 5. Functional annotation of antechinus variants. (a) Total number of variants annotated to various gene regions including: Splicing (within a splice site of a gene), UTR3 (3′ untranslated region), UTR5 (5′ untranslated region), Downstream (within 1kb downstream of a gene), Upsteam (within 1kb 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

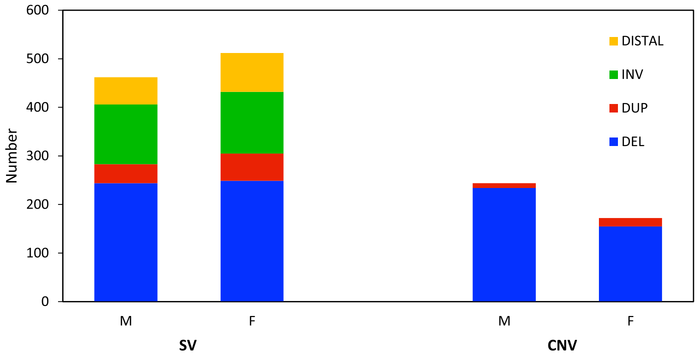


Figure 6. Breakdown of high-quality large structural variants (SVs) and copy number variants (CNVs) in the antechinus. Figure shows both male (M) and female (F) deletions (blue), tandem duplications (red), inversions (green) and distal structural variants (i.e. across two scaffolds, yellow).

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

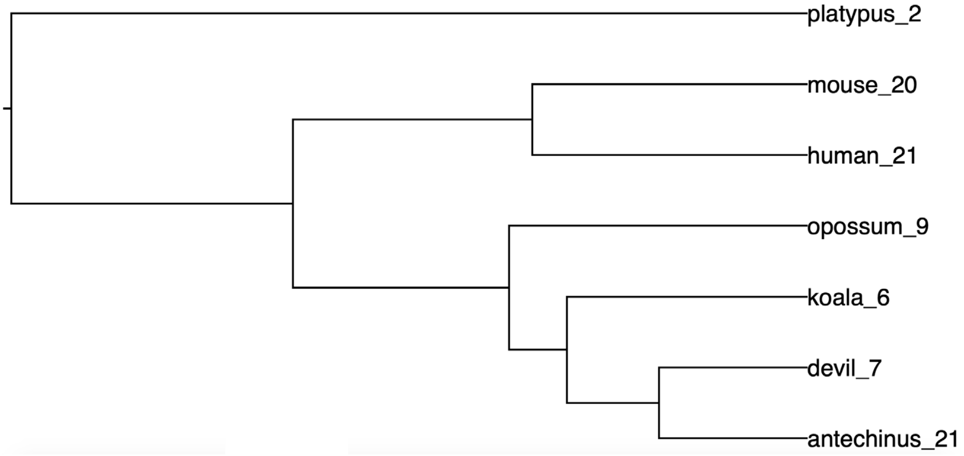
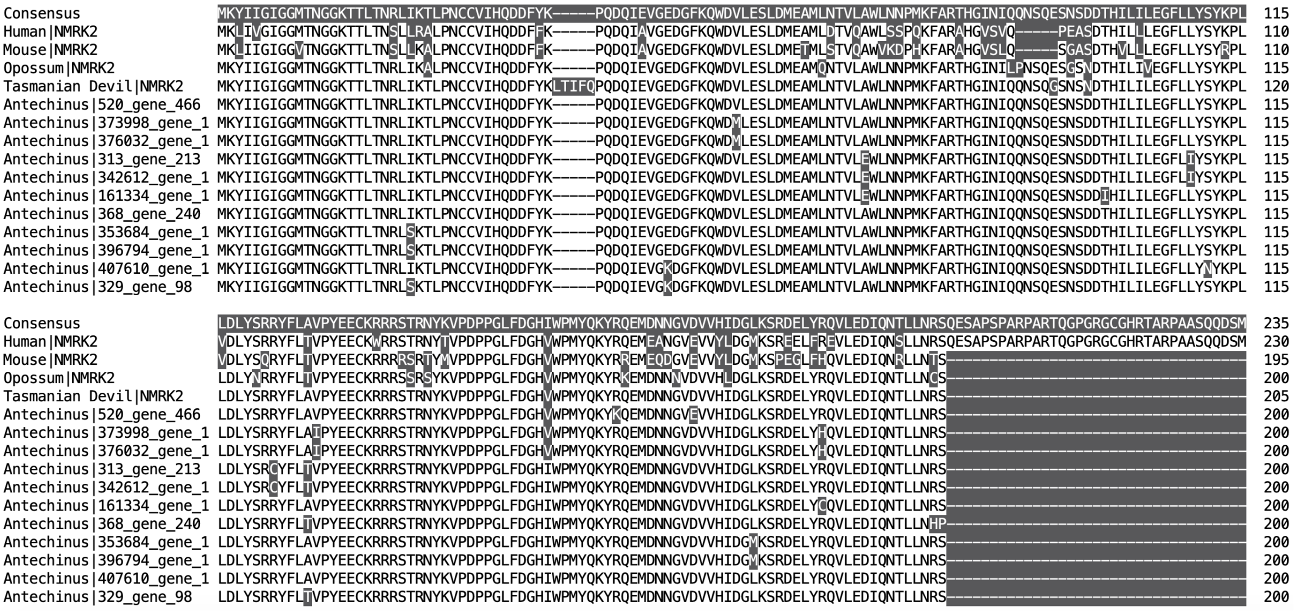


Figure 7. Gene tree showing numbers of Pcdh-γ genes across 7 species.

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.



```
Consensus                MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Human|NMRK2              MKLIVGIGGMTNGGKTTLTNSLLRALPNCCVIHQDDFFK-----PQDQIAVGEDGFKQWDVLESLDMEAMLDTVQAWLSSPQKFARAHGVSVQ-----PEASDTHILLLEGFLLYSYKPL  110
Mouse|NMRK2              MKLIIGIGGVTNGGKTTLTNSLLKALPNCCVIHQDDFFK-----PQDQIAVGEDGFKQWDVLESLDMETMLSTVQAWVKDPHKFARAHGVSLQ-----SGASDTHVLLLEGFLLYSYRPL  110
Opossum|NMRK2            MKYIIGIGGMTNGGKTTLTNRLIKALPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMQNTVLAWLNNPMKFARTHGINILPNSQESGSNDTHILIVEGFLLYSYKPL  115
Tasmanian Devil|NMRK2    MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYKLTIFQPQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQGSNSNDTHILILEGFLLYSYKPL  120
Antechinus|520_gene_466  MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|373998_gene_1 MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDMLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|376032_gene_1 MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDMLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|313_gene_213  MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLEWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLIYSYKPL  115
Antechinus|342612_gene_1 MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLEWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLIYSYKPL  115
Antechinus|161334_gene_1 MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLEWLNNPMKFARTHGINIQQNSQESNSDDIHILILEGFLLYSYKPL  115
Antechinus|368_gene_240  MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|353684_gene_1 MKYIIGIGGMTNGGKTTLTNRLSKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|396794_gene_1 MKYIIGIGGMTNGGKTTLTNRLSKTLPNCCVIHQDDFYK-----PQDQIEVGEDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115
Antechinus|407610_gene_1 MKYIIGIGGMTNGGKTTLTNRLIKTLPNCCVIHQDDFYK-----PQDQIEVGKDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYNYKPL  115
Antechinus|329_gene_98   MKYIIGIGGMTNGGKTTLTNRLSKTLPNCCVIHQDDFYK-----PQDQIEVGKDGFKQWDVLESLDMEAMLNTVLAWLNNPMKFARTHGINIQQNSQESNSDDTHILILEGFLLYSYKPL  115

Consensus                LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRSQESAPSPARPARTQGPGRGCGHRTARPAASQQDSM  235
Human|NMRK2              VDLYSRRYFLTVPYEECKWRRSTRNYTVPDPPGLFDGHVWPMYQKYRQEMEANGVEVVYLDGMKSREELFREVLEDIQNSLLNRSQESAPSPARPARTQGPGRGCGHRTARPAASQQDSM  230
Mouse|NMRK2              VDLYSQRYFLTVPYEECKRRRRSRTYMVPDPPGLFDGHVWPMYQKYRREMEQDGVEVVYLDGMKSPEGLFHQVLEDIQNRLLNTS-----------------------------------  195
Opossum|NMRK2            LDLYNRRYFLTVPYEECKRRRSSRSYKVPDPPGLFDGHVWPMYQKYRKEMDNNNVDVVHLDGLKSRDELYRQVLEDIQNTLLNCS-----------------------------------  200
Tasmanian Devil|NMRK2    LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHVWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  205
Antechinus|520_gene_466  LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHVWPMYQKYKQEMDNNGVEVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|373998_gene_1 LDLYSRRYFLAIPYEECKRRRSTRNYKVPDPPGLFDGHVWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYHQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|376032_gene_1 LDLYSRRYFLAIPYEECKRRRSTRNYKVPDPPGLFDGHVWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYHQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|313_gene_213  LDLYSRCYFLTVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|342612_gene_1 LDLYSRCYFLTVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|161334_gene_1 LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYCQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|368_gene_240  LDLYSRRYFLTVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNHP-----------------------------------  200
Antechinus|353684_gene_1 LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGMKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|396794_gene_1 LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGMKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|407610_gene_1 LDLYSRRYFLAVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
Antechinus|329_gene_98   LDLYSRRYFLTVPYEECKRRRSTRNYKVPDPPGLFDGHIWPMYQKYRQEMDNNGVDVVHIDGLKSRDELYRQVLEDIQNTLLNRS-----------------------------------  200
```

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

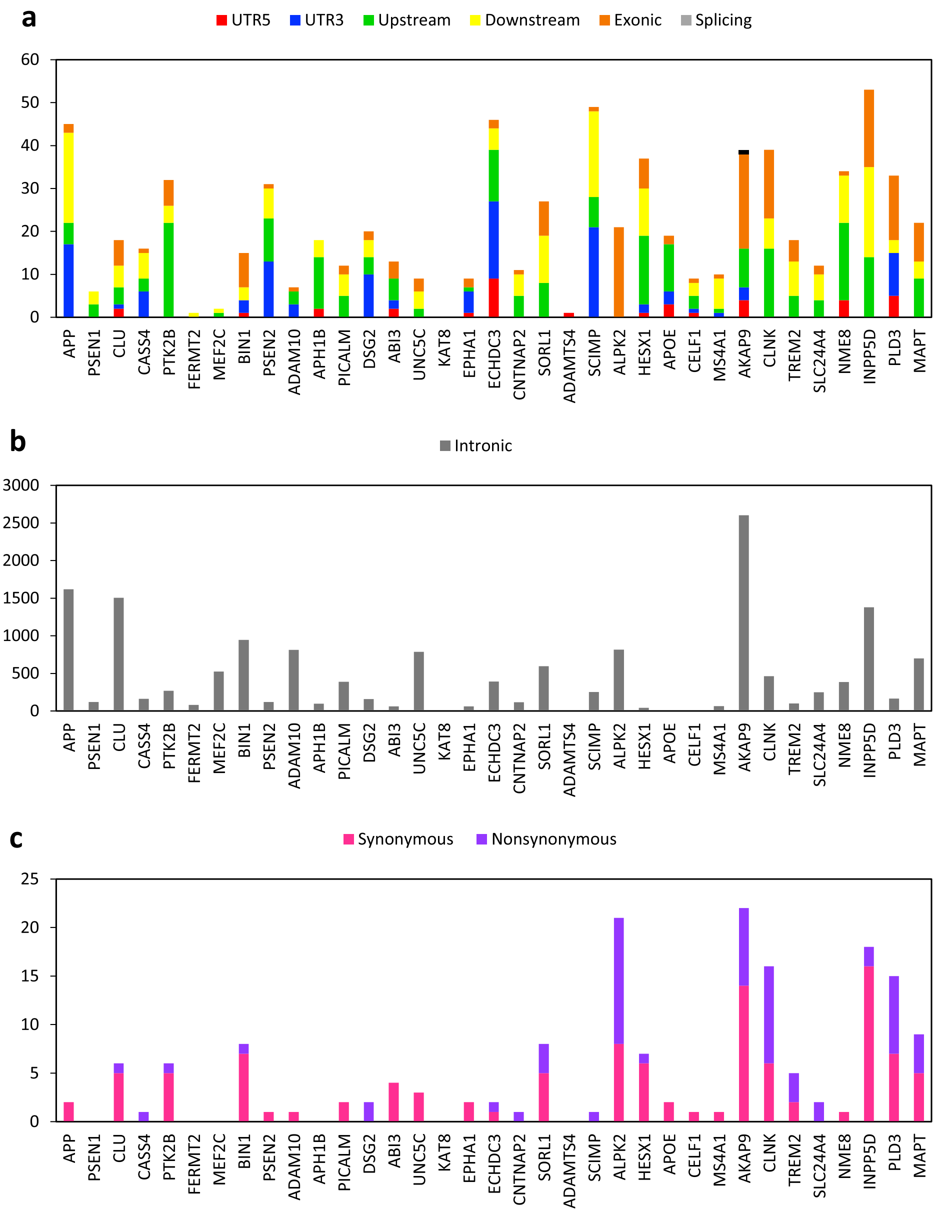


Figure 9. Number of each type of SNV associated with the target Alzheimers-related genes in the antechinus. (a) Numbers of SNVs present in the 5′ UTR, 3′ UTR, 1kb upstream region, 1kb downstream region, exons, and splice sites of each gene. (b) Numbers of intronic SNVs present in each gene. (c) Number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

AVAILABILITY OF SUPPORTING DATA AND MATERIALS

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.

APPENDIX 2: ADAPTATION AND CONSERVATION INSIGHTS FROM THE KOALA GENOME

A2.1 BACKGROUND

The following article details the creation of the first koala reference genome and describes the conservation implications of key downstream findings. At the time of publication, the koala genome was the first genome of an Australian species to utilise third generation (i.e., long read) sequencing and represented the “gold standard” marsupial reference genome. This high-quality reference genome enabled significant insights into the genetic adaptations behind the koala’s unique biology such as expansions within cytochrome P450 genes resulting in the koala’s ability to detoxify eucalyptus leaves and expansions in vomeronasal and taste receptors assisting in the koala’s specialised food choice. Genomic data was also employed to investigate immune genes involved in response to chlamydia vaccine and to identify biogeographical boundaries to gene flow between koala populations. The results from this study were crucial in demonstrating the broad implications of high-quality genomic data in the conservation of a threatened species.

Led by Rebecca N. Johnson and Katherine Belov, the koala genome work comprises a large consortium effort with many experts in different fields bringing their expertise together to better understand the genetics of this iconic but vulnerable Australian species. I contributed to this study by assisting with the annotation of MHC genes and the investigation of candidate genes for chlamydia vaccine response. Please refer to the main article for a complete list of author contributions.

A2.2 MAIN ARTICLE

The article titled “'Adaptation and conservation insights from the koala genome” published in *Nature Genetics* (2018; 50, 1102-1111) is presented on the following pages.

nature genetics

ARTICLES
https://doi.org/10.1038/s41588-018-0153-5

OPEN

Adaptation and conservation insights from the koala genome

Rebecca N. Johnson[1,2,30,31*], Denis O'Meally[2,3,30], Zhiliang Chen[4,30], Graham J. Etherington[5], Simon Y. W. Ho[2], Will J. Nash[5], Catherine E. Grueber[2,6], Yuanyuan Cheng[2,7], Camilla M. Whittington[8], Siobhan Dennison[1], Emma Peel[2], Wilfried Haerty[5], Rachel J. O'Neill[9], Don Colgan[1], Tonia L. Russell[10], David E. Alquezar-Planas[1], Val Attenbrow[1], Jason G. Bragg[11,12], Parice A. Brandies[2], Amanda Yoon-Yee Chong[5,13], Janine E. Deakin[14], Federica Di Palma[5,15], Zachary Duda[9], Mark D. B. Eldridge[1], Kyle M. Ewart[1], Carolyn J. Hogg[2], Greta J. Frankham[1], Arthur Georges[14], Amber K. Gillett[16], Merran Govendir[8], Alex D. Greenwood[17,18], Takashi Hayakawa[19,20], Kristofer M. Helgen[1,21], Matthew Hobbs[1], Clare E. Holleley[22], Thomas N. Heider[9], Elizabeth A. Jones[8], Andrew King[1], Danielle Madden[3], Jennifer A. Marshall Graves[11,14,23], Katrina M. Morris[24], Linda E. Neaves[1,25], Hardip R. Patel[26], Adam Polkinghorne[3], Marilyn B. Renfree[27], Charles Robin[27], Ryan Salinas[4], Kyriakos Tsangaras[28], Paul D. Waters[4], Shafagh A. Waters[4], Belinda Wright[1,2], Marc R. Wilkins[4,10,30], Peter Timms[29,30] and Katherine Belov[2,30,31]

The koala, the only extant species of the marsupial family Phascolarctidae, is classified as 'vulnerable' due to habitat loss and widespread disease. We sequenced the koala genome, producing a complete and contiguous marsupial reference genome, including centromeres. We reveal that the koala's ability to detoxify eucalypt foliage may be due to expansions within a cytochrome P450 gene family, and its ability to smell, taste and moderate ingestion of plant secondary metabolites may be due to expansions in the vomeronasal and taste receptors. We characterized novel lactation proteins that protect young in the pouch and annotated immune genes important for response to chlamydial disease. Historical demography showed a substantial population crash coincident with the decline of Australian megafauna, while contemporary populations had biogeographic boundaries and increased inbreeding in populations affected by historic translocations. We identified genetically diverse populations that require habitat corridors and instituting of translocation programs to aid the koala's survival in the wild.

The koala is an iconic Australian marsupial, instantly recognizable by its round, humanoid face and distinctive body shape. Fossil evidence identifies as many as 15–20 species, following the divergence of koalas (Phascolarctidae) from terrestrial wombats (Vombatidae) 30–40 million years ago[1,2] (Supplementary Fig. 1). The modern koala, *Phascolarctos cinereus*, which first appeared in the fossil record ~350,000 years ago, is the only extant species of the Phascolarctidae. Like other marsupials, koalas give birth to

[1]Australian Museum Research Institute, Australian Museum, Sydney, New South Wales, Australia. [2]School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia. [3]Animal Research Centre, Faculty of Science, Health, Education & Engineering, University of the Sunshine Coast, Maroochydore, Queensland, Australia. [4]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Kensington, New South Wales, Australia. [5]Earlham Institute, Norwich Research Park, Norwich, UK. [6]San Diego Zoo Global, San Diego, CA, USA. [7]UQ Genomics Initiative, University of Queensland, St Lucia, Queensland, Australia. [8]Sydney School of Veterinary Science, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia. [9]Department of Molecular and Cell Biology and Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA. [10]Ramaciotti Centre for Genomics, University of New South Wales, Kensington, New South Wales, Australia. [11]Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia. [12]National Herbarium of New South Wales, Royal Botanic Gardens & Domain Trust, Sydney, New South Wales, Australia. [13]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [14]Institute for Applied Ecology, University of Canberra, Bruce, Australian Capital Territory, Australia. [15]Department of Biological Sciences, University of East Anglia, Norwich, UK. [16]Australia Zoo Wildlife Hospital, Beerwah, Queensland, Australia. [17]Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany. [18]Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany. [19]Department of Wildlife Science (Nagoya Railroad Co., Ltd.), Primate Research Institute, Kyoto University, Inuyama, Japan. [20]Japan Monkey Centre, Inuyama, Japan. [21]School of Biological Sciences, Environment Institute, Centre for Applied Conservation Science, and ARC Centre of Excellence for Australian Biodiversity and Heritage, University of Adelaide, Adelaide, South Australia, Australia. [22]Australian National Wildlife Collection, National Research Collections Australia, CSIRO, Canberra, Australian Capital Territory, Australia. [23]School of Life Sciences, La Trobe University, Bundoora, Victoria, Australia. [24]The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Midlothian, UK. [25]Royal Botanic Garden Edinburgh, Edinburgh, UK. [26]John Curtin School of Medical Research, Australian National University, Acton, Australian Capital Territory, Australia. [27]School of BioSciences, University of Melbourne, Melbourne, Victoria, Australia. [28]Department of Translational Genetics, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. [29]Faculty of Science, Health, Education & Engineering, University of the Sunshine Coast, Maroochydore, Queensland, Australia. [30]These authors contributed equally: Rebecca N. Johnson, Denis O'Meally, Zhiliang Chen, Marc R. Wilkins, Peter Timms, Katherine Belov. [31]These authors jointly supervised this work: Rebecca N. Johnson, Katherine Belov. *e-mail: rebecca.johnson@austmus.gov.au

underdeveloped young. Birth occurs after just 35 d of gestation, with young lacking immune tissues or organs. Their immune system develops while they are in the pouch, meaning survival during early life depends on immunological protection provided by mothers' milk.

A specialist arboreal folivore feeding almost exclusively from *Eucalyptus* spp., the koala has a diet that would be toxic or fatal to most other mammals[3]. Due to the low caloric content of this diet, the koala rests and sleeps up to 22 h a day[4]. A detailed understanding of the mechanisms by which koalas detoxify eucalyptus and protect their young in the pouch has been elusive, as there are no koala research colonies and access to milk and tissue samples is opportunistic. The genome enables unprecedented insights into the unique biology of the koala, without having to harm or disturb an animal of conservation concern.

The genome also enables a holistic, scientifically grounded approach to koala conservation. Australia has the highest mammal extinction record of any country during the Anthropocene[5], and koala numbers have plummeted in northern parts of its range since European settlement of the continent[6], but increased in southern sections of the range, notably in parts of Victoria and South Australia. The uneven response of koala populations throughout its range is one of the most difficult issues in its management[7]. The species was heavily exploited by a pelt trade (1870s to late 1920s), which harvested millions of animals[6,8,9]. Today, the threats are primarily due to loss and fragmentation of habitat, urbanization, climate change and disease. Current estimates put the number of koalas in Australia at only 329,000 (range 144,000–605,000), and a continuing decline is predicted[6]. Koalas present a complex conservation conundrum: in the north, causes of decline include ongoing habitat fragmentation, urbanization and disease. However, decline in the south has followed a different path[10], with widespread, often sequential, translocations (1920–1990s) from a limited founder population, which has resulted in genetically bottlenecked populations that are overabundant to the point of starvation in some areas[11]. There are marked differences in the degree to which threats affect each population, thereby cautioning against one prescription for population recovery.

Adding to the complexity of koala conservation is the impact of disease, specifically koala retrovirus (KoRV) and *Chlamydia*. KoRV is thought to have arrived in Australia via a putative murine vector before cross-species transmission[12,13]. It is now prevalent in northern koalas and appears to be spreading to southern populations[14]. Some strains appear to be more virulent than others and are putatively associated with an increase in neoplastic disease[15]. Similarly, *Chlamydia*, which in some individuals causes severe symptoms yet in others remains asymptomatic, may have crossed the species barrier from introduced hosts such as domestic sheep and cattle following European settlement[16]. A complete koala genome offers insights into the species' genetic susceptibility to these diseases, provides the genomic basis for innovative vaccines, and can underpin new conservation management solutions that incorporate the species' population and genetic structure, such as facilitating gene flow via habitat connectivity or translocations.

Results

Genome landscape. Koalas have 16 chromosomes, differing from the ancestral marsupial $2n = 14$ karyotype by a simple fission of ancestral chromosome 2 giving rise to koala chromosomes 4 and 7[17]. We sequenced the complete genome using 57.3-fold PacBio long-read coverage, generating a 3.42 Gb reference assembly. The primary contigs from the FALCON assembly (representing homozygous regions of the genome) yielded genome version phaCin_unsw_v4.1. This comprised 3.19 Gb, including 1,906 contigs with an N50 of 11.6 Mb and the longest at 40.6 Mb. The heterozygous regions of the genome (representing the alternate contigs from the assembly) totaled 230 Mb, with an N50 of 48.8 kb (Table 1, Supplementary Tables 1–3 and Methods). Approximately 30-fold coverage of Illumina short reads was used to polish the assembly. BioNano optical maps plus additional conserved synteny information for marsupials were used for scaffolding[18] to assemble long-read contigs into 'virtual' chromosome scaffolds ('super-contigs') (Supplementary Tables 4 and 5 and Supplementary Note). The largest super-contig spanned approximately half of koala chromosome 7 (Supplementary Fig. 2).

Our long-read-based sequence presented the opportunity to identify and study centromeres, which are multi-megabase regions that are challenging to construct in eutherian (for example, human and mouse)[19] genome assemblies due to intractable higher order arrays of satellites. Centromeres are smaller in marsupials than in eutherians, and as such are more amenable to analysis[20]. Chromatin immunoprecipitation and sequencing using antibodies to centromeric proteins (CENP-A and CREST)[21] enabled the identification of scaffolds containing putative centromeric regions (Supplementary Fig. 3) and the characterization of known and new repeats, including composite elements within koala centromeric domains (Supplementary Table 6–10) that lack the previously annotated retroelement, kangaroo endogenous retrovirus (KERV), found in some tammar wallaby centromeres[22]. Koala centromeres span a total of 2.6 Mb of the koala haploid genome, equivalent to an average of 300 kb of centromeric material per chromosome. Like those of other species with small centromeres[19,20,23,24], koala centromeres lack higher order satellite arrays (Supplementary Tables 7–10). Among the newly identified repeats, some are similar to composite elements recently described in gibbon centromeres[25], where absence of higher order satellite arrays accompanied the evolution of new composite elements with putative centromere function. The composition of the koala centromere therefore supports mounting evidence that transposable elements represent a major, functional component of small centromeres when higher order satellite arrays are absent[20,24,25].

Interspersed repeats account for approximately 47.5% of the koala genome; 44% of these are transposable elements (Supplementary Table 11). As in other mammalian genomes, short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) are the most numerous elements (35.2% and 28.9% of total number of elements, respectively), with LINEs making up 32.1% of the koala genome. The long-read sequence assembly also enabled full characterization and annotation of repeat-rich long noncoding RNAs, including *RSX*, which mediates X chromosome inactivation in female marsupials[26]. Koala *RSX* represents the first marsupial *RSX* to be fully annotated and to have its structure predicted (Supplementary Fig. 4 and Supplementary Note). As expected, it was expressed in all female tissues, but in no male tissues[27].

The assembled koala genome has very high coverage of coding regions: we recovered 95.1% of 4,104 mammalian benchmarking universal single-copy orthologs (BUSCOs)[28], the highest value for any published marsupial genome (Supplementary Table 5) and comparable with that of the human assembly (GRCh38, which scores 94.1% of orthologs). Analysis of gene family evolution using a maximum-likelihood framework identified 6,124 protein-coding genes in 2,118 gene families with at least two members in koala. Among these, 1,089 have more gene members in koala than in any of the other species (human, mouse, dog, tammar wallaby, Tasmanian devil, gray short-tailed opossum, platypus, chicken; Supplementary Fig. 5).

Having characterized the genome, we undertook detailed analyses of key genes and gene families to gain insights into the genomic basis of the koala's highly specialized biology. Gene families of particular interest were those that encode proteins involved in induced ovulation, those proteins involved in the complex lactation process, those proteins responsible for immunity, and those enzymes that enable the koala to subsist on a toxic diet.

Table 1 | Comparison of assembly quality between koala genome assembly phaCin_unsw_v4.1 and published marsupial and monotreme genomes

Species	Genome size (Gb)	G+C content (%)	No. scaffolds	Scaffold N50 (kb)	Reference
Koala phaCin_unsw_v4.1 (female Bilbo)	3.42	39.0	1,906[a] 5,525[b] (contigs)	11,589 (contig)	This study
Platypus (*Ornithorhynchus anatinus*)	2.3	45.5	200,283	959	Warren et al. 2008[82]
Gray short-tailed opossum (*Monodelphis domestica*)	3.48	37.7	5,223	59,810	Mikkelsen et al. 2007[83]
Tammar wallaby (*Notamacropus eugenii*)	2.7	38.8	277,711	37	Renfree et al. 2011[84]
Tasmanian devil (*Sarcophilus harrisii*)	3.17	36.4	35,974	1,847	Murchison et al. 2012[85]

[a]Homozygous. [b]Heterozygous.

Ability to tolerate a highly toxic diet. The koala's diet of eucalyptus leaves contains high levels of plant secondary metabolites[29], phenolic compounds[30] and terpenes (for example, ref. [31]) that would be lethal to most other mammals[32]. Koalas thus experience little competition for food resources. *Eucalyptus grandis* shows substantial expansion in terpene synthase genes relative to other plant genomes[33]. Eucalypt toxicity is therefore likely to have exerted selection pressure on the koala's ability to metabolize such xenobiotics, so we searched for genes encoding enzymes with a detoxification function and investigated sequence evolution at these loci.

Cytochrome P450 monooxygenase (*CYP*) genes represent a multi-gene superfamily of heme-thiolate enzymes that play a role in detoxification through phase 1 oxidative metabolism of a range of compounds including xenobiotics[34]. These genes have been identified throughout the tree of life, including in plants, animals, fungi, bacteria and viruses[35]. In the koala genome we found two lineage-specific monophyletic expansions of the cytochrome P450 family 2 subfamily C (*CYP2C*s, 31 members in koala) (Fig. 1a). The functional importance of these *CYP2C* genes was further demonstrated through analysis of expression in 15 koala transcriptomes from two koalas, showing particularly high expression in the liver, consistent with a role in detoxification (Supplementary Fig. 6).

Comparing *CYP2C* gene context in mouse versus koala identified conserved flanking markers strongly suggestive of tandem duplication (Fig. 1b). Further sequence-level analysis of the *CYP* expansions indicated that most conserved regions are under strong purifying selection (Fig. 1c). However, there is evidence that individual *CYP* codons have experienced episodic diversifying selection while purifying selection shapes the rest of the gene (Fig. 1c–h, Supplementary Note and Supplementary Tables 12 and 13). Adaptive expansion of *CYP2C* and maintenance of duplicates appear to have worked in concert, resulting in higher enzyme levels for detoxification while the interplay between purifying and diversifying selection resulted in neofunctionalization within the *CYP*s. Such adaptations enable koalas to detoxify their highly specialized diet rich in plant secondary metabolites.

The characterization of koala *CYP2Cs* has significant therapeutic potential. The high expression levels of *CYP2C* genes in the liver helps to explain why meloxicam, a nonsteroidal anti-inflammatory drug (NSAID) known to be metabolized by the protein product of *CYP2C* in humans[36,37] and frequently used for pain relief in veterinary care, is so rapidly metabolized in the koala and a handful of other eucalypt-eating marsupials (common brushtail possum and eastern ringtail possum) compared with eutherian species[37,38]. It is expected that other NSAIDs are also rapidly metabolized in koalas and have little efficacy at suggested doses[39]. Anti-chlamydia antibiotics such as chloramphenicol are degraded rapidly by koalas; treatment with a single dose applicable to humans is insufficient in koalas, which require a daily dose for up to 30 to 45 d. This discovery of *CYP2C* gene expression levels will inform new research into the pharmacokinetics of medicines in koalas.

Taste, smell and food choice. Like many specialist folivores, koalas are notoriously selective feeders, making food choices both to target nutrients and to avoid plant secondary metabolites[40]. Koalas have been observed to sniff leaves before tasting them[41], and their acute discrimination has been correlated with the complexity and concentration of plant secondary metabolites[42]. This suggests an important role for olfaction and vomerolfaction, as well as taste. While most herbivores circumvent plant chemical defenses by detoxifying one or a few compounds[43], the complexity of eucalyptus plant secondary metabolites, in combination with the terpene expansion in eucalypts, led us to hypothesize that the koala requires enhanced capabilities both in specialist detection and in plant secondary metabolite detoxification. We therefore investigated the genomic basis of the koala's taste and smell senses, finding multiple gene family expansions that could enhance its ability to make food choices.

We report an expansion of one lineage of vomeronasal receptor type 1 (*V1R*) genes associated with the detection of nonvolatile odorants (Supplementary Note). There are six such genes in koala, compared with only one in the Tasmanian devil and gray short-tailed opossum, and none found in tammar wallaby, human, mouse, dog, platypus or chicken. The expansion of one lineage of *V1R* genes is consistent with the koala's ability to discriminate among diverse plant secondary metabolites.

Surprisingly, given the degree of its dietary specialization, the olfactory receptor genes ($n = 1,169$) characterized in koala had a gene repertoire that was slightly smaller than that of gray short-tailed opossum (1,431 genes), tammar wallaby (1,660 genes) and Tasmanian devil (1,279 genes) (Supplementary Note). This may be understood in the context of relaxed selection on olfactory receptors among dietary specialists[44].

We also report genomic evidence of expansions within the taste receptor families that would enable the koala to optimize ingestion of leaves with a higher moisture and nutrient content in concert with the concentration of toxic plant secondary metabolites in their food plants. The koala's ability to 'taste water' is potentially enhanced by an apparent functional duplication of the aquaporin 5 gene[45–47] (Supplementary Table 14 and Supplementary Note).

The *TAS2R* family has a role in 'bitter' taste, enabling recognition of structural toxins such as terpenes, phenols and glycosides. These are found in various levels in eucalypts as plant secondary metabolites[3,30,31,48]. In marsupials, the *TAS2R* family includes the orthologous repertoires from eutherians, as well as three specific

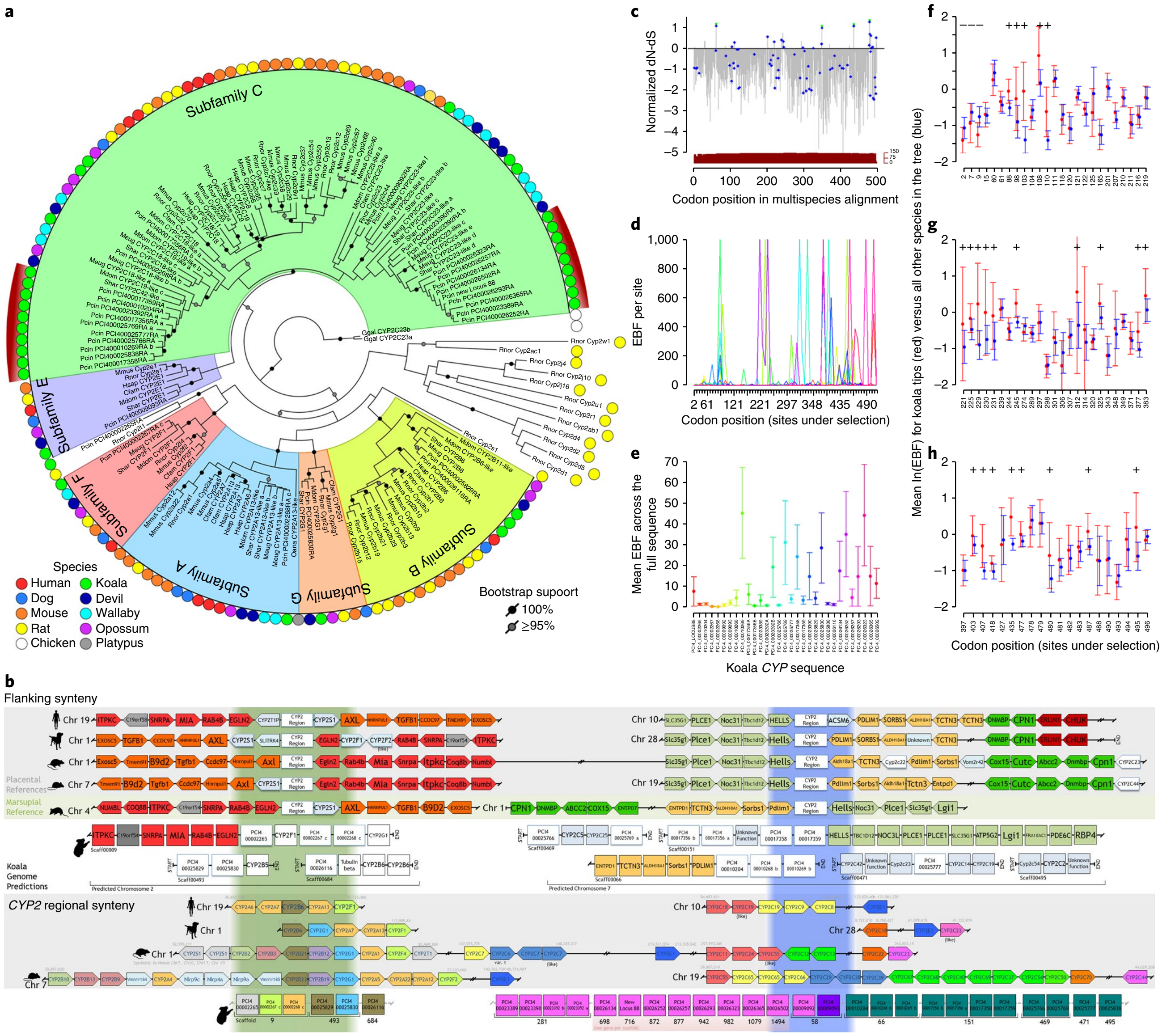


Fig. 1 | Analysis of cytochrome P450 family 2 subfamily C gene family. a, Phylogenetic tree of *CYP2* gene family in koala (Pcin; 31 *CYP2* members), compared with marsupials: tammar wallaby (Meug), Tasmanian devil (Shar), gray short-tailed opossum (Mdom); eutherian mammals: human (Hsap), rat (Rnor), mouse (Mmus), dog (Cfam); monotreme: platypus (Oana); and outgroup chicken (Ggal). Two independent monophyletic expansions are seen in koala in the *CYP2C* subfamily (highlighted by red arcs). **b**, *CYP* synteny map showing expansion of *CYP2C* genes in koala as compared to eutherians (human, dog, rat, mouse) and another marsupial (opossum). **c–h**, Selection analysis of *CYP* gene expansion. **c**, Normalized *dN*-*dS* (SLAC (single-likelihood ancestor counting) method) across the alignment of 152 *CYP* sequences (only sites with data in koala and at least one other species; red bars show sequence depth). Points indicate statistically significant (threshold $\alpha = 0.1$) evidence for codons under selection. Four sites show positive selection across entire tree (SLAC; green points); 70 sites show episodic selection (MEME (mixed effects model of evolution); blue diamonds). **d**, Comparison of episodic selection on particular codons across koala *CYP* genes ($n = 31$); *x* axis shows codons with evidence of statistically significant selection anywhere on the tree (identified in **c**). **e**, Comparison of mean episodic selection among koala *CYP* genes ($n = 70$). Points indicate mean empirical Bayes factor (EBF) for sites under selection for each sequence; error bars, 95% confidence interval. **f–h**, Mean EBF (natural log transformed, EBF values of 0 excluded) for koala tree tips ($n = 31$; red) relative to all others ($n = 121$ in 9 species (see Methods), blue). Points show mean, error bars ± 95% confidence interval, evaluated as 1.96 × s.e.m. (using sequence depth as sample size; red bars in **c**). Codon positions on *x* axis refer to multispecies alignment from **c**. Symbols above each point indicate that the mean value for koala site falls outside the 95% confidence interval for all other species (above, "+"; below, "–"; two-tailed test at $\alpha = 0.05$). Raw statistics shown (unadjusted for multiple comparisons).

expansions in the last common ancestor shared by all marsupials[49,50] (Fig. 2). Large koala-specific duplications in four marsupial orthologous groups have produced a large koala *TAS2R* repertoire of 24 genes (Fig. 2). The koala has more *TAS2Rs* than any other Australian marsupial, and among the most of all mammal species[49,50], including paralogs of human and mouse receptors whose agonists are toxic glycosides (Supplementary Table 15 and Supplementary Note). The *TAS1R* gene families, responsible for sweet taste and umami amino

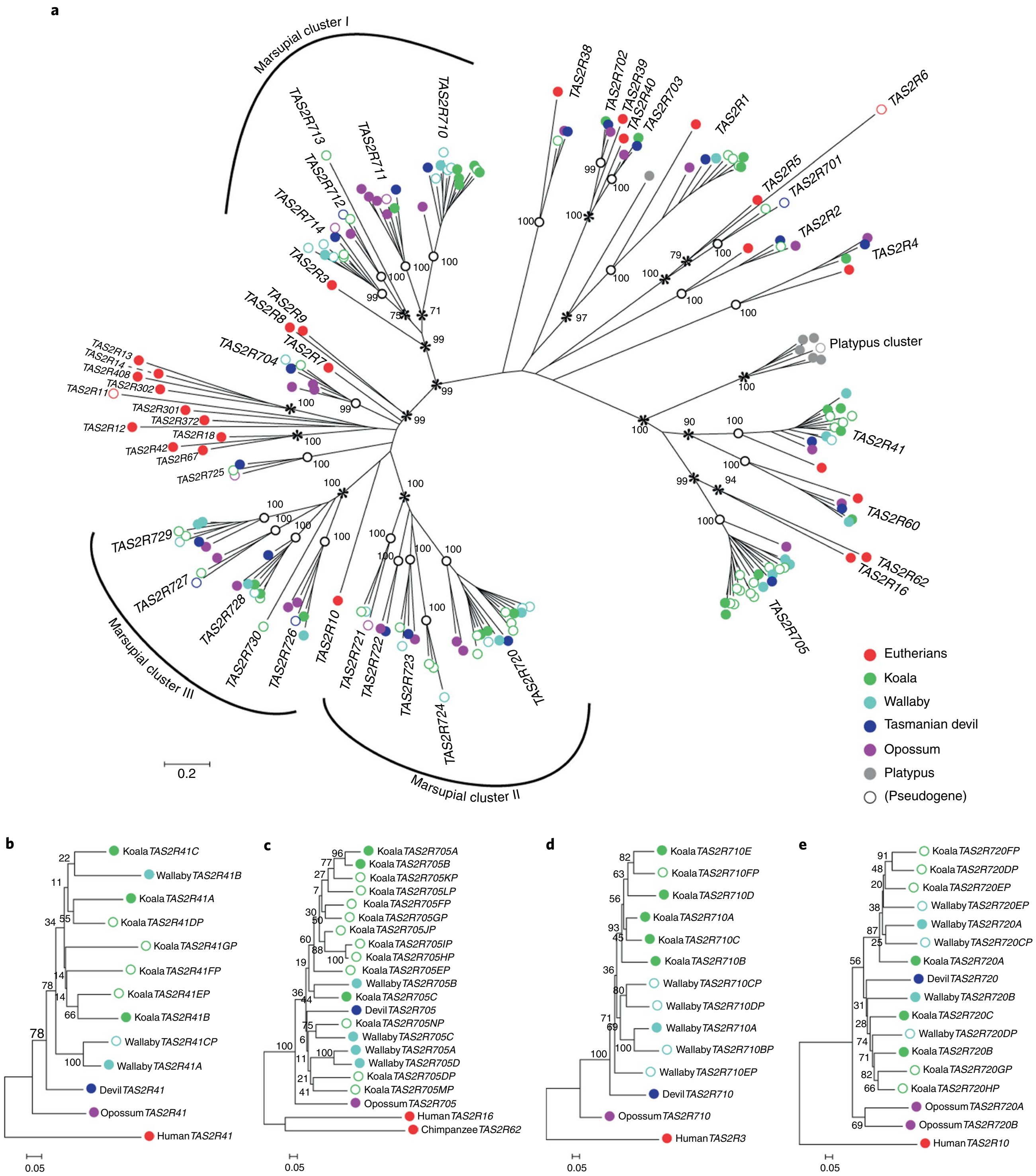


Fig. 2 | Taste receptor analysis in koalas and other mammals identifies three marsupial-specific expansions and further koala-specific duplications. *TAS2R* genes are responsible for bitter taste perception. a, Maximum-likelihood tree of *TAS2R*s (including pseudogenes) in the four marsupials, where the sequences contained 250 amino acids. 28 representative *TAS2R*s of orthologous gene groups (OGGs) in eutherians (red circles) and 7 platypus *TAS2R*s (gray circles) were also used. There were 27 distinct marsupial OGGs (supported by ≥99% bootstrap values), where the nodes of OGG clades are indicated by white open circles. Bootstrap values of ≥70% in the nodes connecting OGG clades are indicated by asterisks. There are three marsupial-specific clusters (I, II and III) where massive expansion events occurred in the common ancestor of marsupials after their split from eutherian ancestors. b–e, Reconstructed maximum-likelihood trees of *TAS2R* orthologs in which there are more than two duplicates of koala *TAS2R*s: b, *TAS2R41*; c, *TAS2R705*; d, *TAS2R710*; and e, *TAS2R720*. Genomic structures of the umami and sweet taste receptor *TAS1R*s were also analyzed and found to be functional in koala (see Supplementary Note).

acid perception, have previously been reported as pseudogenized in eutherians with highly specialized diets, such as the giant panda[51]. In the koala, however, we found that all *TAS1R* genes are putatively functional (Supplementary Fig. 7).

Genomics of an induced ovulator. Koala reproduction is of particular interest because the koala is an induced ovulator[52], with key genes controlling female ovulation (*LHB*, *FSHB*, *ERR1*, *ERR2*), as well as prostaglandin synthesis genes important in parturition and ejaculation (*PTGS1*, *PTGS2*, *PTGS3*) (Supplementary Note). We identified genes putatively involved in the induction of ovulation in the female by male seminal plasma (*NGF*), and in coagulation of seminal fluid (*ODC1*, *SAT1*, *SAT2*, *SMOX*, *SRM*, *SMS*) (Supplementary Note), which may function to prevent sperm leakage from the female reproductive tract in this arboreal species.

Genomic characterization of koala milk. A koala young is about the size of a kidney bean and weighs <0.5 g. It crawls into the mother's posteriorly opening pouch and attaches to a teat, where it remains for 6–7 months. It continues to suck after it has left the pouch until about a year old.

Analysis of the genome, in conjunction with a mammary transcriptome and a milk proteome, enabled us to characterize the main components of koala milk (Supplementary Fig. 8, Supplementary Table 16, Supplementary Note and ref. [53]). The high-quality assembly of the genome allowed both the identification of marsupial-specific genes and determination of their evolutionary origins based on their genomic locations. For instance, we found that there are four Late Lactation Protein (*LLP*) genes tightly linked to both trichosurin and β-lactoglobulin (Supplementary Fig. 8), potentially allowing marsupials to fine-tune milk protein composition across the stages of lactation to meet the changing needs of their young. Additionally, the koala marsupial milk 1 (*MM1*) gene, a novel marsupial gene, is located close to the gene encoding very early lactation protein (*VELP*), an ortholog of *Glycam1* (or *PP3*) that encodes a eutherian antimicrobial protein[53] (Supplementary Fig. 8). In eutherians, this region contains an array of short glycoproteins that have antimicrobial properties and are found in secretions such as milk, tears and sweat. We propose that *MM1* has an antimicrobial role in marsupial milk, along with three other short novel genes located in the same region. We also detected expansions in another antimicrobial gene family, the cathelicidins.

Koala immunome and disease. At the time of European settlement, koalas were widespread in eastern mainland Australia, from north Queensland to the southeastern corner of South Australia. Today they are mainly confined to the east coast and are listed as 'vulnerable' under Australia's *Environment Protection and Biodiversity Conservation Act 1999*[54]. There is strong evidence to suggest that some fragmented populations of koalas are already facing extinction, particularly in formerly densely populated koala territories in southeast Queensland and northern New South Wales. A major challenge for the conservation of these declining koala populations is the high prevalence of disease, especially that caused by the obligate intracellular bacterial pathogen *Chlamydia pecorum*, which is found across the geographic range, with the exception of some offshore islands[55]. A main challenge for managing these populations has been the lack of knowledge about the koala immune response to disease. Recent modeling suggests the best way to stabilize heavily affected koala populations is to target disease[56].

The long-read-based genome enabled the de novo assembly of complex, highly duplicated immune gene families and comprehensive annotation of immune gene clusters[53,57,58]. These include the major histocompatibility complex (*MHC*)[59], as well as T cell receptors (*TCR*), immunoglobulin (*IG*) (Supplementary Fig. 9, Supplementary Tables 17 and 18, and Supplementary Note), natural killer cell (NK) receptor[58] and defensin[60] gene clusters. Together these findings provide a starting point for new disease research and allow us to interrogate the immune response to the most significant pathogen of the koala, *C. pecorum*.

Of the more than 1,000 koalas arriving annually at wildlife hospitals in Queensland and New South Wales, 40% have late-stage chlamydial disease and cannot be rehabilitated. Annotation of koala immune genes enabled us to study variation within candidate genes known to play a role in resistance and susceptibility to chlamydia infection in other species (Supplementary Tables 18–20). Preliminary case/control association tests for five koalas involved in a chlamydia vaccination trial showed that the MHCII *DMA* and *DMB* genes, as well as the *CD8-a* gene, may be involved in differential immune responses to chlamydia vaccine (Supplementary Table 21 and Supplementary Note). We also conducted differential expression analysis of RNA sequencing (RNA-seq) data from conjunctival tissue collected from koalas at necropsy, both with and without signs of ocular chlamydiosis, showing that in diseased animals, 1,508 of the 26,558 annotated genes (5.7%) were twofold upregulated, while 685 (2.6%) were downregulated by greater than twofold when compared with healthy animals (Supplementary Fig. 9 and Supplementary Note). In diseased animals, upregulated genes were associated with Gene Ontology (GO) terms for a range of immunological processes, including signatures of leukocyte infiltration (Supplementary Fig. 9). Immune responses in the affected conjunctivas were directed at T_H1 rather than T_H2 responses. Proinflammatory mediators such as *CCL20*, *IL1α*, *IL1β*, *IL6* and *SSA1* were also upregulated. As in human trachoma, this cascade of proinflammatory products may help to clear the infection but may also lead to tissue damage in the host[61]. Furthermore, resolution of human trachoma infection is thought to require a IFN-γ driven T_H1 response[62], and in diseased koalas we found that IFN-γ was upregulated 4.7-fold in the conjunctival tissue. These annotated koala immune genes will now help us to define features of protective versus pathogenic immunological responses to the disease and may be invaluable for effective vaccine design.

Koala genomes are undergoing genomic invasion by koala retrovirus (KoRV)[63], which is spreading from the north of the country to the south. Both endogenous (germline transmission) and exogenous (infectious 'horizontal' transmission) forms are extant[64]. Our results provide a comprehensive view of KoRV insertions in the koala genome. We found a total of 73 insertions in the phaCin_unsw_4.1 assembly (Supplementary Table 22). It is likely that most of these 73 loci are endogenous, consistent with our observation of integration breakpoint sequences that are shared with one or both of the other koala genomes reported (Supplementary Tables 23 and 24).

We investigated the sites of KoRV insertion to define their proximity to protein-coding genes and explore possible disruptions. This analysis identified insertions into 24 protein-coding genes (Supplementary Table 25). However, none is likely to disrupt protein-coding capacity, since 22 insertions are in introns and the other two are in 3′ untranslated regions. Transcription proceeding from the proviral long terminal repeat (LTR) could possibly affect the transcription of the host genes.

Understanding the genetics of host resistance to chlamydia and the etiology of the retrovirus will help inform the development of vaccines against both diseases, as well as translocation strategies.

Genome-informed conservation. Broad-scale population management of koalas is critical to conservation efforts. This is challenging because distribution models are not easily generalized across bioregions, and further complicated by the unique regional conservation

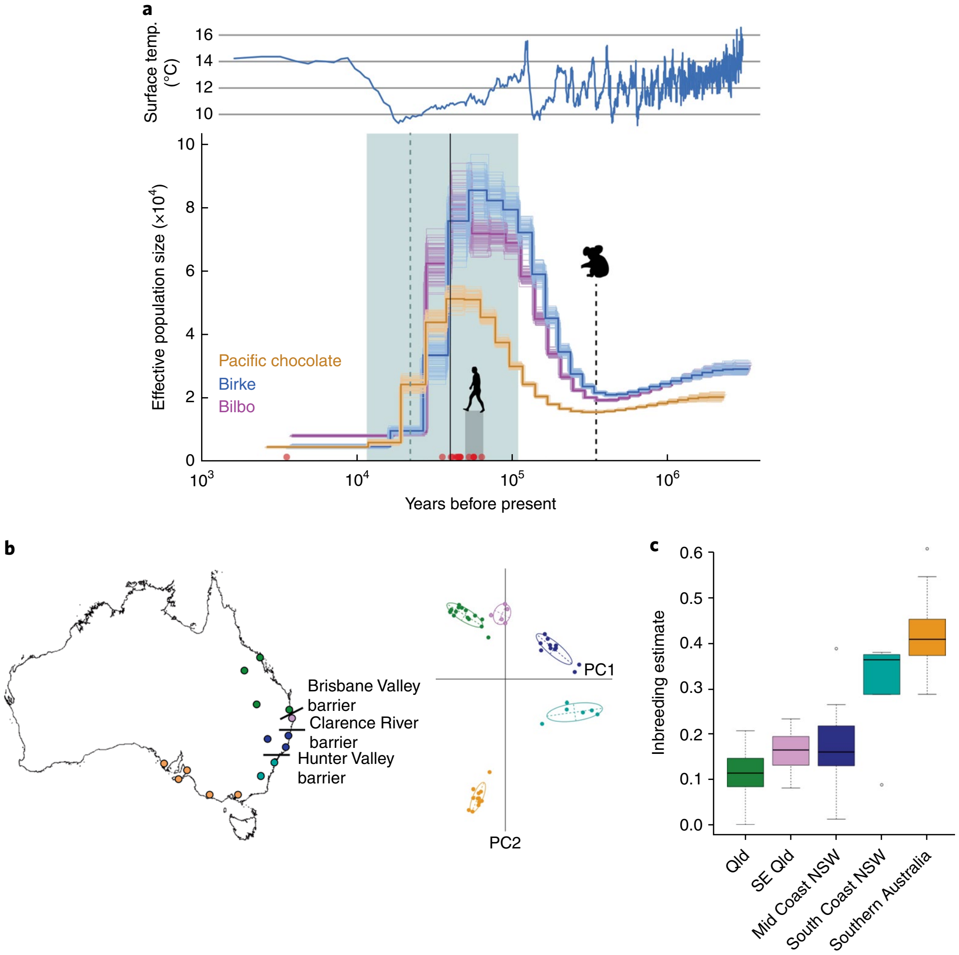


Fig. 3 | Analysis koala populations using genome-mapped markers. a, Top: plot of surface temperature (temp.) over past 3 million years based on a five-point running mean of $\delta^{18}O$ data[76]. Bottom: population demographic history inferred from diploid sequences of three koalas (females 'Pacific Chocolate' and 'Bilbo', male 'Birke') using the pairwise sequential Markovian coalescent (PSMC) method. Koala silhouette indicates earliest fossil record of modern koala[2]. Gray shading, human arrival in Australia[77] (see ref. [78]); red circles, estimated extinction times of 16 megafaunal genera in mainland Australia[79]; aqua area, last glacial period; vertical dashed green line, last glacial maximum; vertical solid black line, first koala population declines 40,000 years ago. Dark colored lines are estimated from genome data; lighter lines, plots inferred from 100 bootstrap replicates. A mutation rate of 1.45×10^{-8} mutations per site per generation and 7-year generation time were assumed. **b**, Right, principal component (PC) analysis (including 95% inertia ellipses) of 1,200 SNPs in 49 wild koalas from throughout Australia. Left, geographic clustering of wild koalas in eastern Australia in relation to proposed biogeographic barriers[68,72], highlighting known historic barriers to gene flow, the Brisbane and Clarence River Valleys, but also suggesting a role for the Hunter Valley. The cluster of genetically similar southern koalas reflects a recent history of widespread translocation[8]. **c**, Average inbreeding coefficient (*F*) (calculated by TrioML[80,81]) of 49 wild koalas. Qld, Queensland; SE, southeast; NSW, New South Wales. *P* values arising from linear modeling represent significant differences in mean *F* between regions (***$P < 0.001$; **$P < 0.01$). There is a high correlation between geographic distance and genetic distance (Mantel test: $r^2 = 0.4898$), indicating that genetic rescue between populations is feasible. Center lines, median; box limits, upper and lower quartiles. Upper whisker = min(max(*x*), Q_3 + 1.5 × IQR), lower whisker = max(min(*x*), Q_1 - 1.5 × IQR); i.e., upper whisker = upper quartile + 1.5 × box length, lower whisker = lower quartile - 1.5 × box length; circles, outliers. Linear modeling indicated that mean *F* differed significantly between several regions (Mid-coast NSW-Southern Australia, $P = 0.000524$; Qld-Southern NSW, $P = 0.00237$; Qld-Southern Australia, $P = 0.00000107$; SE Qld-Southern Australia, $P = 0.006596$).

issues described above. Since it is not possible to generalize management, it is imperative that decisions are informed by empirical data relevant to each bioregion.

Analysis of the koala genome provided the unique opportunity to combine historical evolutionary data with high-resolution contemporary population genomic markers to address these management challenges. To infer the ancient demographic history of the species, we analyzed the long-read reference genome and short-read data from two other koalas, using the pairwise sequentially Markovian coalescent (PSMC) method[65] (Fig. 3a, Supplementary Fig. 10 and Methods). The data show that the modern koala, which appeared in the fossil record 350,000 years ago[2], underwent an initial increase in population, followed by a rapid and widespread decrease in population size ~30,000–40,000 years ago. This is consistent with fossil evidence of rapid declines in multiple Australian species, including the extinct megafauna, 40,000–50,000 years ago[66] and 30,000–40,000 years ago[67]. The koala was thus one of a number of species affected by decline during this time that did not ultimately become extinct[67].

Distinct PSMC profiles of the koalas from two geographic areas and their failure to coalesce suggests some regional differences in koala populations, including impediments to gene flow (Fig. 3a). Regional differentiation was also detected in analyses of mtDNA[68,69], although over a shorter time scale.

We analyzed populations of recent koala samples using 1,200 SNPs derived from targeted capture libraries mapped to the koala genome (Supplementary Note). We found notable levels of genetic diversity with limited fine-scale differentiation consistent with long-term connectivity across regions. We found evidence of low genetic diversity in southern koalas, consistent with a recent history of sequential translocations[8,68,70,71] (Fig. 3b,c). At a continental scale, we show biogeographic barriers to gene flow associated with the Brisbane Valley and Clarence River, as identified by mtDNA studies[68,72], and find a barrier associated with the Hunter Valley, which was not previously known in koalas (Fig. 3b). Levels of inbreeding varied across regions (Fig. 3c), but the northern populations most under threat in New South Wales and Queensland show high levels of genetic diversity.

The information generated here provides a foundation for a conservation management strategy to maintain gene flow regionally while incorporating the genetic legacy of biogeographic barriers. Furthermore, the contrast in genome-wide levels of diversity between southern and northern populations highlights the detrimental consequences of the unmonitored use of small isolated populations as founders for reestablishing and/or rescuing of populations on genome-wide levels of genetic diversity. Low levels of genetic diversity in southern koalas have been associated with genetic abnormalities consistent with inbreeding depression, including testicular abnormalities[73].

Now that we understand the consequences of past translocations, and the existing genetic structure, it is clear that maintaining and facilitating gene flow via habitat connectivity will be the most effective means of ensuring genetically healthy koala populations over the long term. However, where more intensive measures such as translocation are required to rescue genetically depauperate southern populations, these tools and data provide the basis for decisions that maximize benefits while minimizing risks[74,75]. Future utility of these SNPs will also include tracking of individual pedigrees in captive koala populations and in those wild populations being intensively monitored.

The koala genome offers insights into historic and contemporary population dynamics, providing evolutionary and genetic context for a species that is the focus of considerable management actions and resources. By providing a deeper understanding of disease dynamics and population genetic processes, including the maintenance and monitoring of gene flow, this genomic information will enable the development of strategies necessary to preserve the species, from the preservation of habitat corridors through to the genetic rescue of isolated populations. As members of government advisory committees, some of the authors have initiated inclusion of genomic information into the New South Wales Koala Strategy. This will be used to inform koala management in the state with the goal of securing koalas in the wild for the future.

Discussion

The koala genome provides the highest quality marsupial genome to date. This assembly has enabled insights into the colonization of the koala genome by an exogenous retrovirus and revealed the architecture of the immune system, necessary to study and treat emerging diseases that threaten koala populations. A greater understanding of genetic diversity across the species will guide the selection of individuals from genetically healthy northern populations to augment genetically restricted populations in the south, bearing in mind that chlamydia has not been detected on some offshore islands, so risk assessment should be carried out before embarking on translocations. Sequencing the genome has advanced our understanding of the unique biology of the koala, including detoxification pathways and innovations in taste and smell to enable food choices in an obligate folivore. Long-term survival of the species depends on understanding the impacts of disease and management of genetic diversity, as well as the koala's ability to source moisture and select suitable foraging trees. This is particularly important given the koala's narrow food range, which makes it especially vulnerable to a changing climate. The genome provides a springboard for conservation of this biologically unique and iconic Australian species.

URLs. FALCON assembly algorithm, https://github.com/PacificBiosciences/FALCON-integrate/; FALCON (v 0.3.0), http://falconframework.org/; RepeatMasker (v 4.0.3), http://www.repeatmasker.org/; RepeatModeler, http://www.repeatmasker.org/RepeatModeler/; RepBase (v 2015-08-07), http://www.girinst.org/repbase/; MAKER, http://www.yandell-lab.org/software/maker.html; Trinity (v 2.3.2), https://github.com/trinityrnaseq/trinityrnaseq/; SNAP, http://archive.broadinstitute.org/mpg/snap/; GeneMark, http://opal.biology.gatech.edu/GeneMark/; Augustus, http://bioinf.uni-greifswald.de/augustus/; NCBI Blast (v 2.3.0), https://blast.ncbi.nlm.nih.gov/Blast.cgi; OrthoMCL (v 2.0.9), http://orthomcl.org/orthomcl/; MAFFT (v 7.2.71), https://mafft.cbrc.jp/alignment/software/; TreeBeST (v 1.9.2), http://treesoft.sourceforge.net/treebest.shtml; HyPhy, https://veg.github.io/hyphy-site/; Datamonkey, http://www.datamonkey.org/; STAR, http://star.mit.edu/genetics/; featureCounts, http://bioinf.wehi.edu.au/featureCounts/; DESeq2, https://bioconductor.org/packages/release/bioc/html/DESeq2.html; SARTools, https://github.com/PF2-pasteur-fr/SARTools/; Dotter, https://sonnhammer.sbc.su.se/Dotter.html; GATK (v 3.3-0-g37228af), https://software.broadinstitute.org/gatk/; KAT comp, https://github.com/TGAC/KAT/; BUSCO (v 2), http://busco.ezlab.org/; Trimmomatic (v 0.36 PE), http://www.usadellab.org/cms/?page=trimmomatic; Bowtie2 (v 2.2.4), http://bowtie-bio.sourceforge.net/bowtie2/index.shtml; MACS2 (v 2.0.10.20131216), https://github.com/taoliu/MACS/; R (v 3.2.5), https://www.r-project.org/; gplots (v 3.0.1), https://cran.r-project.org/web/packages/gplots/index.html; bedtools (v 2.25.0), http://bedtools.readthedocs.io/en/latest/; kSamples (v 1.2-4), https://cran.r-project.org/web/packages/kSamples/index.html; ggbiplot (v 0.55), https://github.com/vqv/ggbiplot/; Tandem Repeats Finder, https://tandem.bu.edu/trf/trf.html; seqLogo, https://bioconductor.org/packages/release/bioc/html/seqLogo.html; RNAfold, http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi; UniProt/Swiss-Prot, http://www.uniprot.org/; dammit!, https://dammit.readthedocs.io/en/refactor-1.0/; Transfuse, https://github.com/cboursnell/transfuse/; GMAP, http://research-pub.gene.com/gmap/; Trim Galore!, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; Kallisto, https://pachterlab.github.io/kallisto/; Sleuth, https://pachterlab.github.io/sleuth_walkthroughs/trapnell/analysis.html; All-vsl-all BLASTP (version 2.2.30+), https://blast.ncbi.nlm.nih.gov/Blast.cgi; MUSCLE (v 3.8.31), https://www.drive5.com/muscle/; HMMER suit (v 3.1b1 May 2013), http://hmmer.org/; FASTASEARCH (v 36.8.8), https://www.ebi.ac.uk/Tools/sss/fasta/; Integrative Genomics Viewer (IGV) (v 2.3.97), https://github.com/ssadedin/IGV-CRAM/; MEGA (v 7.0.18), https://www.megasoftware.net/; RAxML (v 8.2.11), https://sco.h-its.org/exelixis/web/software/raxml/index.html; Burrows-Wheeler aligner (v 0.7.15), http://bio-bwa.sourceforge.net/; Samtools (v 1.3), http://www.htslib.org/; Geneious (v 10.2.3), https://www.geneious.com/; Coancestry, https://www.zsl.org/science/software/coancestry/; PLINK (v 1.07), http://zzz.bwh.harvard.edu/plink/.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0153-5.

Received: 16 February 2018; Accepted: 30 April 2018;
Published online: 02 July 2018

References

1. Meredith, R. W., Krajewski, C., Westerman, M. & Springer, M. S. Relationships and divergence times among the orders and families of Marsupialia. *Mus. North. Ariz. Bull.* **65**, 383–406 (2009).
2. Black, K. H., Price, G. J., Archer, M. & Hand, S. J. Bearing up well? Understanding the past, present and future of Australia's koalas. *Gondwana Res.* **25**, 1186–1201 (2014).
3. Gleadow, R. M., Haburjak, J., Dunn, J. E., Conn, M. E. & Conn, E. E. Frequency and distribution of cyanogenic glycosides in *Eucalyptus* L'Hérit. *Phytochemistry* **69**, 1870–1874 (2008).
4. Nagy, K. & Martin, R. Field metabolic rate, water flux, food consumption and time budget of koalas, *Phascolarctos cinereus* (Marsupialia: Phascolarctidae) in Victoria. *Aust. J. Zool.* **33**, 655–665 (1985).
5. Woinarski, J. C., Burbidge, A. A. & Harrison, P. L. Ongoing unraveling of a continental fauna: decline and extinction of Australian mammals since European settlement. *Proc. Natl. Acad. Sci. USA* **112**, 4531–4540 (2015).
6. Adams-Hosking, C. et al. Use of expert knowledge to elicit population trends for the koala (*Phascolarctos cinereus*). *Divers. Distrib.* **22**, 249–262 (2016).
7. McAlpine, C. et al. Conserving koalas: a review of the contrasting regional trends, outlooks and policy challenges. *Biol. Conserv.* **192**, 226–236 (2015).
8. Martin, R. & Handasyde, K. A. *The Koala: Natural History, Conservation and Management.* (UNSW Press: Sydney, New South Wales, Australia (1999).
9. Hrdina, F. & Gordon, G. The koala and possum trade in Queensland, 1906–1936. *Aust. Zool.* **32**, 543 (2004).
10. Menkhorst, P. Hunted, marooned, re-introduced, contracepted: a history of koala management in Victoria. in *Too Close for Comfort: Contentious Issues in Human–Wildlife Encounters* (eds. Lunney, D. et al.) 73–92 (Royal Zoological Society of NSW, Mosman, New South Wales, Australia, 2008).
11. Seymour, A. M. et al. High effective inbreeding coefficients correlate with morphological abnormalities in populations of South Australian koalas (*Phascolarctos cinereus*). *Anim. Conserv.* **4**, 211–219 (2001).
12. Simmons, G., Clarke, D., McKee, J., Young, P. & Meers, J. Discovery of a novel retrovirus sequence in an Australian native rodent (*Melomys burtoni*): a putative link between gibbon ape leukemia virus and koala retrovirus. *PLoS One* **9**, e106954 (2014).
13. Alfano, N. et al. Endogenous gibbon ape leukemia virus identified in a rodent (*Melomys burtoni* subsp.) from Wallacea (Indonesia). *J. Virol.* **90**, 8169–8180 (2016).
14. Tarlinton, R. E., Meers, J. & Young, P. R. Retroviral invasion of the koala genome. *Nature* **442**, 79–81 (2006).
15. Xu, W. et al. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. *Proc. Natl. Acad. Sci. USA* **110**, 11547–11552 (2013).
16. Taylor-Brown, A. & Polkinghorne, A. New and emerging chlamydial infections of creatures great and small. *New Microbes New Infect.* **18**, 28–33 (2017).
17. Hayman, D. Marsupial cytogenetics. *Aust. J. Zool.* **37**, 331–349 (1989).
18. Deakin, J. E. et al. Anchoring genome sequence to chromosomes of the central bearded dragon (*Pogona vitticeps*) enables reconstruction of ancestral squamate macrochromosomes and identifies sequence content of the Z chromosome. *BMC Genomics* **17**, 447 (2016).
19. Brown, J.D. & O'Neill, R.J. The evolution of centromeric DNA sequences. *Encyclopedia of Life Sciences* https://doi.org/10.1002/9780470015902.a0020827.pub2 (Wiley, Hoboken, NJ, USA, 2014).
20. Carone, D. M. et al. A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. *Chromosoma* **118**, 113–125 (2009).
21. Earnshaw, W. C. & Rothfield, N. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma* **91**, 313–321 (1985).
22. O'Neill, R. J. W., O'Neill, M. J. & Graves, J. A. M. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**, 68–72 (1998).
23. Nagaki, K. et al. Sequencing of a rice centromere uncovers active genes. *Nat. Genet.* **36**, 138–145 (2004).
24. Zhang, Y. et al. Structural features of the rice chromosome 4 centromere. *Nucleic Acids Res.* **32**, 2023–2030 (2004).
25. Carbone, L. et al. Centromere remodeling in *Hoolock leuconedys* (Hylobatidae) by a new transposable element unique to the gibbons. *Genome Biol. Evol.* **4**, 648–658 (2012).
26. Grant, J. et al. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* **487**, 254–258 (2012).
27. Hobbs, M. et al. A transcriptome resource for the koala (*Phascolarctos cinereus*): insights into koala retrovirus transcription and sequence diversity. *BMC Genomics* **15**, 786 (2014).
28. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
29. Foley, W. J. & Moore, B. D. Plant secondary metabolites and vertebrate herbivores–from physiological regulation to ecosystem function. *Curr. Opin. Plant Biol.* **8**, 430–435 (2005).
30. Eschler, B. M., Pass, D. M., Willis, R. & Foley, W. J. Distribution of foliar formylated phloroglucinol derivatives amongst Eucalyptus species. *Biochem. Syst. Ecol.* **28**, 813–824 (2000).
31. Pass, G. J., McLean, S., Stupans, I. & Davies, N. Microsomal metabolism of the terpene 1,8-cineole in the common brushtail possum (*Trichosurus vulpecula*), koala (*Phascolarctos cinereus*), rat and human. *Xenobiotica* **31**, 205–221 (2001).
32. Ngo, S. N. T., McKinnon, R. A. & Stupans, I. Cloning and expression of koala (*Phascolarctos cinereus*) liver cytochrome P450 CYP4A15. *Gene* **376**, 123–132 (2006).
33. Myburg, A. A. et al. The genome of *Eucalyptus grandis*. *Nature* **510**, 356–362 (2014).
34. Kirischian, N., McArthur, A. G., Jesuthasan, C., Krattenmacher, B. & Wilson, J. Y. Phylogenetic and functional analysis of the vertebrate cytochrome P450 2 family. *J. Mol. Evol.* **72**, 56–71 (2011).
35. Nelson, D. R. The cytochrome P450 homepage. *Hum. Genomics* **4**, 59–65 (2009).
36. Miners, J. O. & Birkett, D. J. Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. *Br. J. Clin. Pharmacol.* **45**, 525–538 (1998).
37. Davies, N. M. & Skjodt, N. M. Clinical pharmacokinetics of meloxicam. A cyclo-oxygenase-2 preferential nonsteroidal anti-inflammatory drug. *Clin. Pharmacokinet.* **36**, 115–126 (1999).
38. Kimble, B. et al. In vitro hepatic microsomal metabolism of meloxicam in koalas (*Phascolarctos cinereus*), brushtail possums (*Trichosurus vulpecula*), ringtail possums (*Pseudocheirus peregrinus*), rats (*Rattus norvegicus*) and dogs (*Canis lupus familiaris*). *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **161**, 7–14 (2014).
39. Blanshard, W. & Bodley, K. Koalas. in *Medicine of Australian Mammals* (eds. Vogelnest, L. & Woods, R.) 307–327 (Csiro Publishing, Melbourne, Victoria, Australia, 2008).
40. Villalba, J. J., Provenza, F. D. & Bryant, J. Consequences of the interaction between nutrients and plant secondary metabolites on herbivore selectivity: benefits or detriments for plants? *Oikos* **97**, 282–292 (2002).
41. Kratzing, J. E. The anatomy and histology of the nasal cavity of the koala (*Phascolarctos cinereus*). *J. Anat.* **138**, 55–65 (1984).
42. Moore, B. D., Foley, W. J., Wallis, I. R., Cowling, A. & Handasyde, K. A. Eucalyptus foliar chemistry explains selective feeding by koalas. *Biol. Lett.* **1**, 64–67 (2005).
43. Freeland, W.J. & Janzen, D.H. Strategies in herbivory by mammals: the role of plant secondary compounds. *Am. Nat.* **108**, 269–289 https://doi.org/10.1086/282907 (1974).
44. McBride, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc. Natl. Acad. Sci. USA* **104**, 4996–5001 (2007).
45. Watson, K. J. et al. Expression of aquaporin water channels in rat taste buds. *Chem. Senses* **32**, 411–421 (2007).
46. Rosen, A. M., Roussin, A. T. & Di Lorenzo, P. M. Water as an independent taste modality. *Front. Neurosci.* **4**, 175 (2010).
47. Gilbertson, T. A., Baquero, A. F. & Spray-Watson, K. J. Water taste: the importance of osmotic sensing in the oral cavity. *J. Water Health* **4**, 35–40 (2006).
48. Meyerhof, W. et al. The molecular receptive ranges of human TAS2R bitter taste receptors. *Chem. Senses* **35**, 157–170 (2010).
49. Hayakawa, T., Suzuki-Hashido, N., Matsui, A. & Go, Y. Frequent expansions of the bitter taste receptor gene repertoire during evolution of mammals in the Euarchontoglires clade. *Mol. Biol. Evol.* **31**, 2018–2031 (2014).
50. Li, D. & Zhang, J. Diet shapes the evolution of the vertebrate bitter taste receptor gene repertoire. *Mol. Biol. Evol.* **31**, 303–309 (2014).
51. Li, R. et al. The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
52. Johnston, S. D., McGowan, M. R., O'Callaghan, P., Cox, R. & Nicolson, V. Studies of the oestrous cycle, oestrus and pregnancy in the koala (*Phascolarctos cinereus*). *J. Reprod. Fertil.* **120**, 49–57 (2000).
53. Morris, K. M. et al. Characterisation of the immune compounds in koala milk using a combined transcriptomic and proteomic approach. *Sci. Rep.* **6**, 35011 (2016).
54. Department of the Environment. *Phascolarctos cinereus* (combined populations of Queensland, New South Wales and the Australian Capital Territory) in Species Profile and Threats Database (Department of the Environment, Canberra, Australian Capital Territory, 2016).
55. Polkinghorne, A., Hanger, J. & Timms, P. Recent advances in understanding the biology, epidemiology and control of chlamydial infections in koalas. *Vet. Microbiol.* **165**, 214–223 (2013).

56. Rhodes, J. R. et al. Using integrated population modelling to quantify the implications of multiple threatening processes for a rapidly declining population. *Biol. Conserv.* **144**, 1081–1088 (2011).
57. Morris, K. et al. The koala immunological toolkit: sequence identification and comparison of key markers of the koala (*Phascolarctos cinereus*) immune response. *Aust. J. Zool.* **62**, 195–199 (2014).
58. Morris, K. M. et al. Identification, characterisation and expression analysis of natural killer receptor genes in *Chlamydia pecorum* infected koalas (*Phascolarctos cinereus*). *BMC Genomics* **16**, 796 (2015).
59. Cheng, Y. et al. Characterisation of MHC class I genes in the koala. *Immunogenetics* **70**, 125–133 (2018).
60. Jones, E. A., Cheng, Y., O'Meally, D. & Belov, K. Characterization of the antimicrobial peptide family defensins in the Tasmanian devil (*Sarcophilus harrisii*), koala (*Phascolarctos cinereus*), and tammar wallaby (*Macropus eugenii*). *Immunogenetics* **69**, 133–143 (2017).
61. Burton, M. J. et al. Pathogenesis of progressive scarring trachoma in Ethiopia and Tanzania and its implications for disease control: two cohort studies. *PLoS Negl. Trop. Dis.* **9**, e0003763 (2015).
62. Derrick, T., Roberts, C., Last, A. R., Burr, S. E. & Holland, M. J. Trachoma and ocular chlamydial infection in the era of genomics. *Mediators Inflamm.* **2015**, 791847 (2015).
63. Stoye, J. P. Koala retrovirus: a genome invasion in real time. *Genome Biol.* **7**, 241 (2006).
64. Hobbs, M. et al. Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. *Sci. Rep.* **7**, 15838 (2017).
65. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
66. Roberts, R. G. et al. New ages for the last Australian megafauna: continent-wide extinction about 46,000 years ago. *Science* **292**, 1888–1892 (2001).
67. Field, J., Wroe, S., Trueman, C. N., Garvey, J. & Wyatt-Spratt, S. Looking for the archaeological signature in Australian megafaunal extinctions. *Quat. Int.* **285**, 76–88 (2013).
68. Neaves, L. E. et al. Phylogeography of the koala, (*Phascolarctos cinereus*), and harmonising data to inform conservation. *PLoS One* **11**, e0162207 (2016).
69. Tsangaras, K. et al. Historically low mitochondrial DNA diversity in koalas (*Phascolarctos cinereus*). *BMC Genet.* **13**, 92 (2012).
70. Taylor, A. C., Graves, J. A., Murray, N. D. & Sherwin, W. B. Conservation genetics of the koala (*Phascolarctos cinereus*). II. Limited variability in minisatellite DNA sequences. *Biochem. Genet.* **29**, 355–363 (1991).
71. Taylor, A. C. et al. Conservation genetics of the koala (*Phascolarctos cinereus*): low mitochondrial DNA variation amongst southern Australian populations. *Genet. Res.* **69**, 25–33 (1997).
72. Dennison, S. et al. Population genetics of the koala (*Phascolarctos cinereus*) in north-eastern New South Wales and south-eastern Queensland. *Aust. J. Zool.* **64**, 402–412 (2017).
73. Cristescu, R. et al. Inbreeding and testicular abnormalities in a bottlenecked population of koalas (*Phascolarctos cinereus*). *Wildl. Res.* **36**, 299–308 (2009).
74. Frankham, R. et al. Predicting the probability of outbreeding depression. *Conserv. Biol.* **25**, 465–475 (2011).
75. Frankham, R. et al. *Genetic Management of Fragmented Animal and Plant Populations* (Oxford University Press, Oxford, 2017).
76. Hansen, J., Sato, M., Russell, G. & Kharecha, P. Climate sensitivity, sea level and atmospheric carbon dioxide. *Philos. Trans. A Math. Phys. Eng. Sci.* **371**, 20120294 (2013).
77. O'Connell, J. F. & Allen, J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *J. Archaeol. Sci.* **56**, 73–84 (2015).
78. Clarkson, C. et al. Human occupation of northern Australia by 65,000 years ago. *Nature* **547**, 306–310 (2017).
79. Saltré, F. et al. Climate change not to blame for late Quaternary megafauna extinctions in Australia. *Nat. Commun.* **7**, 10511 (2016).
80. Wang, J. Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genet. Res.* **89**, 135–153 (2007).
81. Wang, J. COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Mol. Ecol. Resour.* **11**, 141–145 (2011).
82. Warren, W. C. et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175–183 (2008).
83. Mikkelsen, T. S. et al. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).
84. Renfree, M. B. et al. Genome sequence of an Australian kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. *Genome Biol.* **12**, R81 (2011).
85. Murchison, E. P. et al. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* **148**, 780–791 (2012).

Acknowledgements

R.N.J. and the Australian Museum acknowledge the Australian Museum Foundation, Bioplatforms Australia, New South Wales Environmental Trust grant 2014/RD/0015, Australian Research Council LIEF Grant LE160100154, the University of Sydney HPC service and Amazon Web Services for support; and C. Staples from Featherdale Wildlife Park, C. Flanagan from Port Macquarie Koala Hospital, J. Hangar, E. Hynes, J. Reed, S. Ingleby, A. Divljan and S. Ginn for assistance with sample acquisition. K.B. acknowledges support from the Australian Research Council and Bioplatforms Australia. M.R.W. and the Ramaciotti Centre for Genomics acknowledge support from the Australian Research Council, from the Australian Government NCRIS scheme via Bioplatforms Australia, the New South Wales State Government RAAP scheme and the University of New South Wales. W.H. and W.J.N. were supported by strategic BBSRC funding (Institute Strategic Programme Grant BB/J004669/1) and by the NBI Computing Infrastructure for Science (CiS) group. A.D.G., K.M.H and K.T. were supported by grant R01GM092706 from the National Institute of General Medical Sciences (NIGMS) and A.D.G. had additional support from Morris Animal Foundation grant D14ZO-94. T.N.H., Z.D. and R.J.O. were supported by awards from the National Science Foundation 1613806 and the facilities within the Center for Genome Innovation at the University of Connecticut. C.E.H. thanks CSIRO National Research Collections Australia funding. K.B. and A.P. thank the veterinary staff at Australia Zoo Wildlife Hospital, Currumbin Wildlife Hospital and Moggill Koala Hospital for their assistance in the collection of samples for the koala conjunctival transcriptome study. T.H. acknowledges the Kyoto University Research Administration Office (KURA) for support and was financed by JSPS KAKENHI grant number 16K18630 and the Sasakawa Scientific Research Grant from the Japan Science Society. A.P. and P.T. acknowledge financial support from the Australian Research Council and A.G. financial support via Australian Research Council Discovery Grant DP110104377. C.M.W. is supported by a University of Sydney research fellowship from the estate of Mabs Melville. All authors thank Bioplatforms Australia and Pacific Biosciences. The authors thank T. Haydon for valuable editorial input; S. Potter for expert technical assistance; and R. Gleadow, C. Frere, D. Lunney and D. Alvarez-Ponce for valuable discussions on content.

Author contributions

R.N.J., K.B., P.T. and M.R.W. designed the original concept and scientific objectives and oversaw the project and analyses. R.N.J., D.C., M.D.B.E., A.K.G., D.O., A.K. and P.T. acquired samples for sequencing. T.L.R., M.R.W., Z.C., D.O., G.J.E. and F.D.P. performed library preparation, genome sequencing, assembly and annotation. S.Y.W.H. performed PSMC analysis. A.Y.-Y.C. characterized repetitive sequences. R.J.O., T.N.H. and Z.D. characterized centromeric and telomeric regions. C.M.W. and M.B.R. annotated and analyzed reproductive and developmental genes. K.M.M. annotated and analyzed lactation genes. T.H. and D.C. annotated and analyzed *TAS1R* and *TAS2R* genes. H.R.P. annotated and analyzed *OR* genes. D.C. annotated and analyzed aquaporin genes. K.B., Y.C., P.A.B., E.A.J., D.O. and E.P. annotated and analyzed *MHC*, *Ig*, *TCR*, *NK* and antimicrobial genes. A.P., K.B., D.O. and D.M. analyzed the ocular RNA-seq data. P.A.B., B.W., C.E.G., P.T., K.B. and A.P investigated candidate genes for chlamydia vaccine response. P.T., M.R.W., R.S., M.H., A.K., A.D.G. and K.T. characterized retrovirus insertions and wrote the KoRV sections of the manuscript. J.G.B., S.D., M.D.B.E., G.J.F., L.E.N., R.N.J., B.W. and C.J.H. contributed to analyses and interpretation of exon capture sequence data. P.D.W., S.A.W. and H.R.P. annotated and analyzed *RSX* data. W.J.N., C.E.G., Y.C., W.H., F.D., M.G., K.M.E., B.W. and C.R. analyzed *CYP* genes. C.E.G. and C.M.W. analyzed *V1R* genes. J.E.D., A.G. and H.R.P. constructed super-scaffolds. J.A.M.G., V.A., F.D., C.J.H., K.M.H., A.P., B.W., D.C., M.H., D.E.A.-P., P.A.B., L.E.N., C.E.G., S.A.W. and C.E.H. provided constructive feedback on data analysis and interpretation. R.N.J., P.T., K.B., M.R.W., A.P., M.D.B.E. and G.J.F. obtained funding and other resources. R.N.J. and K.B. wrote the manuscript with input from all other authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0153-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.N.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

General methods. A full description of the Methods can be found in the Supplementary Note. No statistical methods were used to predetermine sample size.

Genome sequencing and assembly of the koala reference genome. *Sequencing.* Samples were obtained as part of veterinary care at the Port Macquarie Koala Hospital and Australia Zoo Wildlife Hospital, and from the Australian Museum Tissue Collection. Sample collection was performed in accordance with methods approved by the Australian Museum Animal Ethics Committee (permit numbers 11–03 and 15–05). "Pacific Chocolate" (Australian Museum registration M.45022), a female from Port Macquarie in northeast New South Wales, was sampled immediately after euthanasia by veterinary staff at the Port Macquarie Koala Hospital (27 June 2012), following unsuccessful treatment of severe chlamydiosis. Two koalas from southeast Queensland—a female, "Bilbo" (Australian Museum registration M.47724), from Upper Brookfield, and a male, "Birke", from Birkdale—were sampled following euthanasia due to severe chlamydiosis (20 August 2015) and severe injuries (26 August 2012), respectively. High molecular weight (HMW) DNA was extracted from heart tissue for Pacific Chocolate and kidney tissue for Birke using the DNeasy Blood and Tissue kit (Qiagen), with RNaseA (Qiagen) treatment. HMW DNA from Bilbo was extracted for PacBio sequencing from spleen tissue using Genomic-Tip 100/G columns (Qiagen), DNA Buffer set (Qiagen) and RNaseA (Qiagen) treatment. Fifteen SMRTbell libraries were prepared (RCG) as per the PacBio 20-kb template preparation protocol, with an additional damage repair step performed after size selection. A minimum size cutoff of 15 or 20 kb was used in the size selection stage using the Sage Science BluePippin system. The libraries were sequenced on the Pacific Biosciences RS II platform (Pacific Biosciences) employing P6 C4 chemistry with either 240 min or 360 min movie lengths. A total of 272 SMRT Cells were sequenced to give an estimated overall coverage of 57.3× based on a genome size of 3.5 Gbp. A TruSeq DNA PCR free library was constructed with a mean library insert size of 450 bp. 400,473,997 paired-end reads were generated yielding a minimum coverage of 34×. HMW gDNA was sequenced on an Illumina 150bpPE HiSeq X Ten sequencing run (Illumina)

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

After filtering low-quality and duplicate reads, approximately 57.3-fold long-read coverage was used for assembly. The primary contigs from the FALCON v 0.3.0 assembly (representing homozygous regions of the genome) yielded genome version phaCin_unsw_v4.1. This comprised 3.19 Gb, including 1,906 contigs with an N50 of 11.6 Mb and sizes ranging up to 40.6 Mb. The heterozygous regions of the genome (representing the alternative contigs from the assembly) were a total of 230 Mb, with an N50 of 48.8 Kb (Supplementary Table 2). Approximately 30-fold coverage of Illumina short reads was used to polish the assembly with Pilon[86].

BUSCO analysis on the draft assembly was run against the mammalian ortholog database with the –long parameter on all genomes under comparison. This initial analysis showed the assembly only reached about 60% of genome completeness, suggesting a high number of indels in the draft genome. The genome polishing tool Pilon[86] was employed to improve draft assembly from FALCON. About 30× of 150 bp paired-end Illumina X Ten short reads from Bilbo was used as an input for this polishing process, which was run on a compute cluster provided by Intersect Australia Limited.

We implemented the method of Deakin et al.[18] for super-scaffolding. Briefly, tables of homologous genes were generated using the physical order of genes on the chromosomes of gray short-tailed opossum and tammar wallaby as references and koala phaCin_unsw_v4.1 (Bilbo) as target (Supplementary Table 4).

Analysis of centromeric regions and repeat structure. Repeat content was called using RepeatMasker with combined RepBase libraries (v 2015-08-07) and RepeatModeller calls generated from the genome assemblies. The resulting calls were then filtered using custom Python scripts to remove short fragments (see "Code availability") and combine tandem or overlapping repeat calls. To characterize the centromeric regions of the genome, chromatin immunoprecipitation (ChIP) was performed using the Invitrogen MAGnify Chromatin Immunoprecipitation System (Revision 6). Repeat content of the centromeric regions was determined using RepBase annotated marsupial repeats and output from RepeatModeller analysis of koala. RepeatMasker was used to locate repeats. Candidate centromeric segments were identified using two sliding window analyses, with window sizes of 200 kb and 20 kb and step sizes of 100 kb and 10 kb, respectively. Small tandem repeats were discovered in koala *RSX* sequence using the Tandem Repeat Finder program[87], using +2, −3, and −7 as scores for match, mismatch and gap opening, respectively. Alignments of consensus repeat units with the *RSX* sequence were processed to obtain nucleotide frequency at each position.

Genome annotation and gene family analysis. Annotations were generated using the automated genome annotation pipeline MAKER[88,89]. We masked repeats in the assembly by providing MAKER with a koala-specific repeat library generated with RepeatModeler[90], against which RepeatMasker (v 4.0.3)[91] queried genomic contigs. Gene annotations were made using a protein database combining the UniProt/Swiss-Prot[92] protein database, all sequences for human (*Homo sapiens*), gray short-tailed opossum (*Monodelphis domestica*), Tasmanian devil (*Sarcophilus harrisii*) and tammar wallaby (*Notamacropus eugenii*) from the NCBI protein database[93], and a curated set of marsupial and monotreme immune genes[94]. We downloaded all published koala mRNAseq reads from SRA (PRJNA230900, PRJNA327021) and reassembled de novo male, female and mammary transcriptomes using the default parameters of Trinity v 2.3.2[95]. Each assembly was filtered such that contigs accounting for 90% of mapped reads were passed to MAKER as homologous transcript evidence. Ab initio gene predictions were made using the programs SNAP[96], Genemark[97] and Augustus[98]. Three iterative runs of MAKER were used to produce the final gene set.

Gene families were called using NCBI Blast (2.3.0) OrthoMCL (2.0.9)[99]. The protein sequences of genes belonging to orthogroups identified by OrthoMCL were aligned using MAFFT (7.2.71)[100] and the gene tree was inferred using TreeBeST (1.9.2)[101] providing a species tree to guide the phylogenetic reconstruction. Custom scripts (see "Code availability") were applied to identify families with expansion within the koala, Diprotodontia, Australidelphia and marsupial lineages.

Sequence evolution. Sequence evolution on specific gene families was conducted on the cytochrome P450 (*CYP*), vomeronasal receptor (*V1R*), olfactory receptor (*OR*), aquaporin and taste receptor genes (Supplementary Note). Genes involved in koala development and reproduction and lactation were also characterized (Supplementary Note). Koala *MHC*, *TCR* and *IGG* genes were annotated and analyzed for expression between diseased and healthy animals (Supplementary Note). Evidence of selection across *CYP* and *V1R* genes was evaluated (Supplementary Note) using multispecies alignments ($N = 152$ and 8 sequences, respectively) in HyPhy[102], hosted by the Datamonkey webserver[103].

RNA-seq analysis of koala conjunctival tissue samples. Conjunctival tissue samples were collected from 26 koalas euthanized due to injury or disease by veterinarians at Australia Zoo Wildlife Hospital, Currumbin Wildlife Hospital and Moggill Koala Hospital. The collection protocol was approved by the University of the Sunshine Coast Animal Ethics Committee (AN/S/15/36). Health assessments of the eye were performed by an experienced veterinarian and classified as either 'healthy' ($N = 13$) or 'diseased' ($N = 13$) based on evidence of gross pathology consistent with ocular chlamydiosis[55]. Conjunctival tissue samples from each animal were placed directly in RNALater (Qiagen, Germany) buffer overnight at 4 °C before storing at −80 °C for later use. RNA was extracted using an RNeasy Mini Kit (Qiagen, Germany) according to the manufacturer's instructions, with an on-column DNase treatment to eliminate contaminating DNA from the sample. The concentration and quality of the isolated RNA was determined using a NanoDrop ND-1000 160 Spectrophotometer and Agilent BioAnalyzer (Agilent, USA). Library construction and sequencing were performed by the Ramaciotti Centre (UNSW, Kensington, NSW) with TruSeq stranded mRNA chemistry on a NextSeq500 (Illumina, USA). Reads were mapped to the phCin_unsw_v4.1 assembly using the default parameters of STAR[104] and counts summed over features using featureCounts[105]. Differentially expressed genes were called using DESeq2[106] as implemented in the SARTools package[107].

Koala retrovirus (KoRV). We searched for KoRV sequences within the scaffolds of the phaCin_unsw v4.1 assembly of the Bilbo genome sequence, and also within alternative contig sequences before their correction by Pilon (since we noticed that in a few cases KoRV sequences were removed in the course of the sequence polishing process). KoRV sequences were found by using the program blastn[108] to search with KoRV genome reference sequences (GenBank AF151794 and AB721500). Search results were converted to BED format and the KoRV and recKoRV components of each read were merged with the program mergeBed. KoRV insertions within genes were identified using the program intersectBed[109]. Pre-integration allelic sequences were found by using blastn[108] to search the phaCin_unsw v4.1 genome sequence assembly with sequences flanking KoRV/recKoRV integrations as queries. In two cases the expected allelic sequence was not present in the Bilbo genome, but was found by searching the genome of another koala (Pacific Chocolate). To check the expected relationship between pairs of allelic sequences, we inspected dot plot alignments of representative sequences (not shown) created with the program dotter[110].

Koala population genomics: historical population size. Demographic history was inferred from the diploid sequence of each of the three koalas, using a pairwise sequential Markovian coalescent (PSMC) method[65]. We conducted a range of preliminary analyses and found that PSMC plots were not sensitive to the values chosen for the maximum number of iterations (N), the number of free atomic time intervals (p), the maximum time to the most recent common ancestor (t), and the initial value of ρ. Based on these investigations, our final PSMC analyses of the three genome sequences used values of $N=25$, $t=5$, $\rho=1$ and $p=4+25\times2+4+6$. The number of atomic time intervals is similar to that recommended for analyses of modern human genomes[65], which are similar in size to the koala genomes. We determined the variance in estimates of N_e using 100 bootstrap replicates. Replicate analyses in which we varied the values of p, t and ρ produced PSMC plots that were broadly similar to those using our chosen 'optimal' settings (Supplementary Fig. 10).

The plots of demographic history were scaled using a generation length of 7 years, corresponding to the midpoint of the range of 6 to 8 years estimated for the koala[111] and the midpoints of the estimates of the human mutation rate (1.45×10^{-8} mutations per site per generation; summarized by ref. [112]) and mouse mutation rate (5.4×10^{-9} mutations per site per generation[113]) were applied in the absence of a mutation rate estimate for koala (Supplementary Fig. 10). The koala mutation rate is likely to be closer to that of humans, based on greater similarity in genome size, life history, and effective population size, relative to mouse[112].

Koala population genomics: contemporary population analysis. Forty-nine koalas were sampled throughout the distribution using a hierarchical approach to allow examination of genetic relationships at a range of scales, from familial to range-wide. All individuals were sequenced using a target capture approach described in ref. [114], with a kit targeting 2,167 marsupial exon sequences. Illumina sequence reads were quality-filtered and trimmed (see ref. [114] for details) and mapped to the koala genome (Bowtie2, v2.2.4[115]). A panel of 4,257 SNP sites was identified (using GATK version 3.3-0-g37228af[116]) that showed expected levels of relatedness and differentiation among the sampled individuals. A panel of 1,200 SNPs (obtained by mapping to targets, filtering, and selecting one SNP per target) showed fine-scale regional differentiation consistent with evolutionary history and recent population management (Fig. 3).

Statistics and reproducibility. In Fig. 1e, points shown indicate the mean empirical Bayes factor (EBF) for sites under selection; error bars, 95% confidence interval. In Fig. 1f–h, 95% confidence intervals are calculated as 1.96 × s.e.m. (sample size is sequence depth, as indicated by red bars in Fig. 1c).

In Fig. 3c, center lines indicate median and box limits indicate upper and lower quartiles. Upper whisker = min(max(x), Q_3 + 1.5 × IQR), lower whisker = max(min(x), Q_1 – 1.5 × IQR); i.e., upper whisker = upper quartile + 1.5 × box length, lower whisker = lower quartile – 1.5 × box length. Circles indicate outliers. Linear modeling indicated that mean F differed significantly between several regions (Midcoast New South Wales–Southern Australia, $P=0.000524$; Queensland–Southern New South Wales, $P=0.00237$; Queensland–Southern Australia, $P=0.00000107$; Southeast Queensland–Southern Australia, $P=0.006596$).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. (1) Custom scripts to identify gene families with expansion within the koala, Diprotodontia, Australidelphia and marsupial lineages; (2) custom scripts to identify refined repeat calls; and (3) code used to generate SNP genotypes from exon capture data are available at https://github.com/DrRebeccaJ/KoalaGenome.

Data availability. The *Phascolarctos cinereus* BioSamples are as follows: Bilbo 61053, SAMN06198159; Pacific Chocolate, SAMEA91939168; Birke. SAMEA103910665. Koala Genome Consortium Projects for the Koala Whole Genome Shotgun project and genome assembly are registered under the umbrella BioProject PRJEB19389 (union of PRJEB5196 and PRJNA359763).

Transcriptome data are submitted under PRJNA230900 (adrenal, brain, heart, lung, kidney, uterus, liver and spleen) and PRJNA327021 (milk and mammary gland). Illumina short-read data for Birke is submitted under PRJEB19982.

The Bilbo 61053 assembly described in this paper is version MSTS01000000 and consists of sequences MSTS01000001–MSTS01001906. For the Bilbo assembly Illumina X Ten reads are submitted under PRJEB19457 and PacBio reads under PRJEB19889.

ChIP-seq data have been deposited under BioProject PRJNA415832 and GEO GSE111153.

References

86. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
87. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
88. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
89. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012).
90. Smit, A., Hubley, R. & Green, P. RepeatModeler Open-1.0. 2008–2015 (2014).
91. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015 (2015).
92. Boutet, E. et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. in *Plant Bioinformatics: Methods and Protocols* (ed. Edwards, D.) 23–54 (2016).
93. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
94. Wong, E. S., Papenfuss, A. T. & Belov, K. Immunome database for marsupials and monotremes. *BMC Immunol.* **12**, 48 (2011).
95. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
96. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
97. Borodovsky, M. & Lomsadze, A. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Curr. Protoc. Bioinformatics* **4**, 4.5.1–4.5.17 (2011).
98. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
99. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
100. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
101. Vilella, A. J. et al. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
102. Pond, S.L.K. & Muse, S.V. HyPhy: hypothesis testing using phylogenies. in *Statistical Methods in Molecular Evolution* 125–181 (Springer, New York, 2005).
103. Delport, W., Poon, A. F., Frost, S. D. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455–2457 (2010).
104. Dobin, A. & Gingeras, T. R. Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics* **11**, 11.14.1–11.14.19 (2015).
105. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
106. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
107. Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. & Dillies, M.-A. SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. *PLoS One* **11**, e0157022 (2016).
108. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
109. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
110. Sonnhammer, E. L. & Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**, GC1–GC10 (1995).
111. Phillips, S. S. Population trends and the koala conservation debate. *Conserv. Biol.* **14**, 650–659 (2000).
112. Lynch, M. et al. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
113. Uchimura, A. et al. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res.* **25**, 1125–1134 (2015).
114. Bragg, J. G., Potter, S., Bi, K. & Moritz, C. Exon capture phylogenomics: efficacy across scales of divergence. *Mol. Ecol. Resour.* **16**, 1059–1068 (2016).
115. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
116. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

A2.3 SUPPLEMENTARY

Supplementary information for “Adaptation and conservation insights from the koala genome” including supplementary figures, text, large data tables and other files are available at: https://www.nature.com/articles/s41588-018-0153-5#Sec31

APPENDIX 3: MHC GENES AND MATE CHOICE

A3.1 BACKGROUND

The following book chapter describes the importance of major histocompatibility complex (MHC) gene diversity and explains the hypothesised mechanisms by which these genes act to influence mate choice in vertebrates:

Brandies, PA, Grueber, CE, Hogg, CJ & Belov, K 2019, 'MHC Genes and Mate Choice', in Choe, J (ed.), Encyclopedia of Animal Behaviour, 2 edn, Elsevier, Massachusetts, USA.

I was invited to write this book chapter during my PhD following publication of my B.S. (Adv) Honours work which examined MHC-based mate choice in a captive koala population. Catherine E. Grueber, Carolyn J. Hogg and Katherine Belov assisted with drafting the chapter.

A3.2 BOOK CHAPTER

Due to copyright restrictions only the first page of the book chapter is presented below. A copy of the complete book chapter is available from authors upon request.

MHC Genes and Mate Choice

Parice A Brandies, University of Sydney, Sydney, NSW, Australia
Catherine E Grueber, University of Sydney, Sydney, NSW, Australia; and San Diego Zoo Global, San Diego, CA, United States
Carolyn J Hogg and Katherine Belov, University of Sydney, Sydney, NSW, Australia

Abstract

Genetic mate choice evolves to minimize inbreeding and/or maximize offspring genetic quality. Genes of the major histocompatibility complex (MHC) have been found to influence mate choice decisions in a wide variety of organisms presumably due to their role in adaptive immunity. This article explores the biological mechanisms of MHC-based mate choice and discusses three major MHC-based hypotheses including A) quantity of alleles, B) genetic compatibility and C) advantage of particular alleles. The context-dependent nature of these hypotheses is examined using examples from the literature.

Keywords

Adaptive immunity; Genetic compatibility; Genetic diversity; Heterozygosity; Inbreeding avoidance; Kin recognition; Major histocompatibility complex (MHC); MHC alleles; MHC recognition; Olfactory cues; Sexual selection

The Genetic Basis of Mate Choice

Mate choice mechanisms evolve when the choosier sex experiences fitness advantages by selecting mates based on particular traits. These traits may be direct benefits such as parental care, protection from a predator and access to resources, or indirect benefits that improve reproductive success or offspring quality. Consequently, mate choice exists across a vast range of organisms and has implications for our understanding of the evolution, biology and conservation of species.

Mate choice can be measured empirically, but the mechanisms underlying mate choice are not always well understood. Progress in molecular genetics technology has allowed the study of mate choice to shift its focus from behavioral observations to understanding mating preferences at the molecular level. Mate choice based on genetic factors can evolve as a mechanism of inbreeding avoidance and/or to maximize offspring genetic quality. Matings between relatives can reduce individual fitness due to increased offspring homozygosity, resulting in exposure of deleterious recessive alleles and inbreeding depression. Selecting distantly-related mates increases heterozygosity of offspring and avoids inbreeding depression. However, in order for this type of mate choice to emerge, potential mating partners must be able to discriminate kin from non-kin. One molecular mechanism of kin recognition is thought to be mediated by the major histocompatibility complex (MHC).

Major Histocompatibility Complex

The major histocompatibility complex, MHC, plays a vital role in the adaptive immune response across a number of vertebrate taxa including fish, reptiles, birds, mammals and humans. Each MHC allele encodes a molecule that recognizes and binds particular self or non-self-peptides. The MHC thereby allows the immune system to distinguish local and foreign antigens and to initiate an immune response against invading microbes (Balakrishnan and Adams, 1995). Within the MHC gene family there are two main classes of molecules: MHC class I and MHC class II. Class I molecules present virus-derived peptides to CD8+ T-lymphocytes, which kill virus-infected cells. Class II molecules present peptides from extracellular bacteria and larger parasites to CD4+ T-lymphocytes, which stimulate B cells and the production of antibodies.

The Importance of MHC Diversity

As MHC expression is co-dominant, both maternal and paternal alleles are expressed. Often, only one or few MHC alleles provide resistance to a particular pathogen (Suri *et al.*, 2003), so having greater heterozygosity at MHC loci increases an individual's ability to respond to a larger range of pathogens. For example, Penn *et al.* (2002) found that heterozygous mice (*Mus domesticus*) are more resistant to multiple-strain *Salmonella* infections than homozygous mice. In line with the expectation that a broad repertoire of MHC alleles should improve individual immunity, the genomic region that contains MHC-coding genes shows evidence of complex evolution in many species, including gene duplication and gene conversion. The result is a highly polygenic, polymorphic MHC. A side-effect of immune-driven diversity of MHC genes, and the genes' ability to distinguish self from non-self, is a potential mechanism by which prospective mates can be discriminated.

https://doi.org/10.1016/B978-0-12-809633-8.90061-X

Encyclopedia of Animal Behavior, Second Edition, 2019, 331–336

APPENDIX 4: EXEMPLARS OF DOCUMENTATION FOR BIOINFORMATIC PIPELINES

A4.1 BACKGROUND

Throughout my PhD I collaborated with the Australian BioCommons to create accessible documentation that could assist with the development, deployment and/or optimisation of key community-endorsed bioinformatics tools and workflows in Australia. In this appendix I provide examples of this documentation for one genome assembly workflow and one genome annotation workflow that I tested and optimised in collaboration with the Australian BioCommons and Pawsey Supercomputing Centre. These workflows were employed in Chapters 4 and 5 of this thesis. The documentation has been made available to researchers through the Australian BioCommons Github: https://github.com/AustralianBioCommons

A4.2 ASSEMBLING PACBIO HIFI DATA WITH IPA AND PURGE_DUPS

The following documentation provides a quick start tutorial and computational guidelines for researchers wanting to employ the improved phased assembler (IPA) and purge_dups tools to assemble a genome with PacBio HiFi reads.

IPA on Nimbus @ Pawsey Supercomputing Centre

Accessing workflow

The scripts for using IPA on Nimbus, have been made available below in the Quick start tutorial.

Quick start tutorial

Below is a tutorial for installing and running IPA on Nimbus.

Install the improved phased assembler via bioconda

Note:

- the instructions below will install the latest version of miniconda and IPA.
- to install a specific version of IPA, use the following install script: `conda install pbipa=1.1.2`

1. Download miniconda: `curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh`
2. Install miniconda: `sh Miniconda3-latest-Linux-x86_64.sh`
    - **Note**: You can choose where to install Miniconda during the installation steps. Depending on how many programs and dependencies you need to install, the installation folder can get quite large in size so it is recommended to install Miniconda on your attached storage drive rather than on the root drive.
3. Create a conda environment and add the required channels: `conda create -n ipa`
4. Activate the conda environment: `conda activate ipa`
5. Add the required channels: `conda config --add channels defaults && conda config --add channels bioconda && conda config --add channels conda-forge`
6. Install the improved phased assembler in the environment: `conda install pbipa`
7. Verify the installation: `ipa validate`
    - **Note**: If there is a samtools error during installation you may need to run `conda install samtools=1.9 --force-reinstall`

Install Purge_Dups and required dependencies:

8. Inside the ipa conda environment install minimap 2: `conda install minimap2`

9. Install purge dups:

```
wget https://github.com/dfguan/purge_dups/archive/master.zip
unzip master.zip
cd purge_dups-master/src && make
export PATH=$PATH:/path/to/purge_dups-master/bin/
```

Run the separate elements of the workflow

Note The `nohup` command is used to run commands in the background so any interruption to the terminal session will not result in the pipeline being stopped. Feel free to use `screen` or any other method instead.

Improved Phased Assembler

1. Create an output directory on your attached storage drive: `mkdir species_ipa_out`

2. Create a temporary directory on your attached storage drive: `mkdir species_ipa_temp`

3. Run the pipeline in the background:

```
nohup ipa local --nthreads 16 --njobs 4 --run-dir
/path/to/species_ipa_out/ --tmp-dir /path/to/species_ipa_temp/ -i
/path/to/input.ccs.bam > species_ipa.log 2>&1&
```

- **Note**: To resume an existing run, run the exact same command with `--resume` on the end. See the IPA documentation for a full list of accepted input CCS file types.

Purge_Dups

4. Convert ccs.bam file to fastq.gz:

```
nohup samtools bam2fq -0 multiple_movies.ccs.fastq.gz -T np,rq -@ 64
multiple_movies.ccs.bam > bam2fq.out 2>&1&
```

5. Align PacBio fastqs to hifi assembly:

```
nohup minimap2 -t 64 -xmap-pb -I 6G final.p_ctg.fasta
multiple_movies.ccs.fastq.gz | gzip -c - > multiple_movies.ccs.paf.gz
2> align.log &
```

- **Note**: Ensure -I is set to a number larger than the size of your hifi assembly in Gb

6. Create required outputs for purge_dups:

```
pbcstat multiple_movies.ccs.paf.gz > pbcstat.log 2>&1
calcuts PB.stat > cutoffs 2> calcults.log
```

7. Split assembly into contigs and perform self alignment:

```
split_fa final.p_ctg.fasta > final.p_ctg.fasta.split 2> split.log
nohup minimap2 -t 64 -xasm5 -DP final.p_ctg.fasta.split
```

```
final.p_ctg.fasta.split | gzip -c - >
final.p_ctg.fasta.split.self.paf.gz 2> selfalign.log &
```

8. Identify duplications: `nohup purge_dups -2 -T cutoffs -c PB.base.cov final.p_ctg.fasta.split.self.paf.gz > dups.bed 2> purge_dups.log &`

9. Create purged assembly: `nohup get_seqs dups.bed final.p_ctg.fasta > get_seqs.log 2>&1&`

10. Merge the purged haplotigs with the original ipa alternate haplotigs: `cat final.a_ctg.fasta hap.fa > alternate.fa`

11. Perform steps 5-9 on the alternate assembly to get the final set of alternate haplotigs

Optimisation required

One extra conda command was required to fix an issue with Samtools dependency during IPA install: try running `conda install samtools=1.9 --force-reinstall`

Infrastructure usage and benchmarking

Exemplar 1: Greater bilby reference genome assembly

- Estimated Genome Size: ~3.5GB
- PacBio HiFi Coverage = ~15x
- Input movie.ccs.bam file (CCS step completed by sequencing company) = ~30GB
- Genome Assembly Walltime: 3hr IPA, 1hr purge_dups

VM Specifications

- Operating System: Ubuntu: 18.04
- RAM: 256GB
- CPUs: 64
- Attached Storage: 3TB

Output Bilby Genome Specifications

- Genome Size: 3.65GB
- No. Scaffolds: 8,173
- No. Contigs: 8,203
- Scaffold N50: 0.71MB
- Contig N50: 0.71MB

Acknowledgements / citations / credits

Workflow components

This workflow documentation refers to, and makes use of, IPA and PurgeDups tools, which were developed by others. Please see the documentation for the individual tools for further information regarding acknowledgements and citations.

Community

Infrastructure deployment, optimisation and testing information is provided by the Australasian Wildlife Genomics Group, The University of Sydney and was supported by the University of Sydney and the Pawsey Supercomputing Centre.

A4.3 GENOME ANNOTATION WITH FGENESH++

The following documentation provides a quick start tutorial and computational guidelines for researchers wanting to employ the Fgenesh++ workflow for genome annotation.

Fgenesh++ on Nimbus @ Pawsey Supercomputing Centre

Accessing tool/workflow

The scripts for using FGENESH++ on Nimbus, have been made available below in the Quick start tutorial.

Quickstart tutorial

Setting up the Protein Database

Note: The most recent NCBI "nr_animals_par" and "nr_plants" protein database curated by Softberry and associated files are already available at `/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/FGENESHPIPE/`. These databases should be suitable for all animal and plant species respectively. Follow the steps below when working with a new or custom NR protein database and also refer to Appendix 1 of the FGENESH++ documentation.

1. Create a blast database from the protein fasta file: `makeblastdb -in custom_proteins -dbtype prot -max_file_sz 2GB`
2. Index the blast database: `/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/FGENESHPIPE/scripts/make_nr_indexed.pl -f custom_proteins -i custom_proteins.ind`

Note: All protein sequences must be a minimum of 6 amino acids otherwise the pipeline will error.

Setting up the mRNA Evidence Files

FGENESH++ requires 3 mRNA evidence files:

1. `species.cdna` = mRNA sequences in fasta format
2. `species.pro` = respective protein sequences in fasta format

3. `species.dat` = a text file which contains the start codon position of the protein sequence within the respective mRNA file, the stop codon position and the chromosome this transcript should be annotated to (or "na" for when the chromosome is unknown)

Note: Each line in the `species.dat` file should correspond to the respective fasta sequence in the `species.cdna` and `species.pro` files. This means the number of fasta files in both of these files should be equal to one another and to the number of lines in the `species.dat` file

FGENESH++ also has strict requirements for the formatting of the mRNA evidence files including:

- Only mRNA sequences containing complete protein sequences can be included, i.e. The sequences must contain both a start codon and stop codon
    - **Note**: Final stop codons `'*'` should be removed from the protein sequences in the `species.pro` fasta file
- Any mRNA Sequences on the reverse strand should be reverse complemented so that the start codon position in the `species.dat` file is always lower than the stop codon position
- The mRNA sequence headers in the `species.cdna` file must have the following fasta header format: `>seq_id <chromosome> <ATG_coord> <STOP_coord> ## <mRNA comments> ## <mRNA_length>` e.g. `>TRINITY_DN100007_c0_g1_i1 na 275 574 ## mRNA assembled by Trinity ## 1202`

To check that you have formatted your 3 mRNA evidence files correctly use the script provided by Softberry:

`/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/check_cdna_dat_pro_files.pl species.cdna species.dat species.pro`

If you wish to use a custom transcriptome dataset, you will first need to run transdecoder (or another ORF predictor) on the transcriptome assembly. You will then need to extract the protein and mRNA sequences corresponding to the longest ORF per "gene", create the .dat file, and adjust the transcript header as above. Alternatively, if you wish to use a RefSeq mRNA datset, please refer to Appendix 2 of the FGENESH++ documentation.

Note: We strongly recommend using a reference-guided approach for custom transcriptomes rather than a *de novo* approach where possible. It is also strongly recommended to only use 1 representative transcript per "gene" (i.e. do not include multiple isoforms in mRNA evidence files) as Fgenesh++ does not handle differential splicing and this also signficiantly increases walltime. Transcriptome data can later be mapped to the annotations if splicing information is required.

Setting up the Genome Sequence Files for Parallel Execution of FGENESH++

FGENESH++ does not run in parallel by default which can result in extremely long wall times. Hence, it is recommended to split the genome up into separate sequence lists to run the program in parallel. First, you need to convert your multi-fasta genome file (and your respective masked file) to single fasta sequences and generate lists of these sequences:

```
#Make output directory
mkdir outputdir

#Split reference multifasta into single fasta files and create list of output files
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/split_multi_fasta.pl ref.fasta -
name seq_id -dir /full/path/to/outputdir -mklist scaffolds.list

#Split masked reference multifasta into single fasta files and create list of output
files
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/split_multi_fasta.pl
ref.fasta.masked -name seq_id -dir /full/path/to/outputdir -mklist
masked_scaffolds.list -ext fasta.masked
```

Then, you can split these lists up into multiple sub-lists to be run in parallel. How you do this will depend on the number of sequences in your fasta file:

Option 1: no. sequences in genome < number of cores

If there are less sequences in the genome than the number of cores on the machine (e.g. for near chromosome length assemblies), then split each sequence into a separate file:

```
#Split list of scaffolds into N sublists (where N = the number of total sequences)
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N.pl scaffolds.list -n N

#Split list of masked scaffolds into N sublists (where N = the number of total
sequences)
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N.pl masked_scaffolds.list -n N
```

Option 2: many sequences in genome (100s)

If there are quite a few sequences in the genome (e.g. hundreds), then split the genome up into as many sublists as there are cores on the machine with roughly equal total sequence length/sizes in each sublist:

```
#Create a file of scaffold lengths in decending order and sort the list of scaffolds
into this order
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl scaffolds.list
scaffolds.list.sorted scaffolds_len.txt.sorted

#Create a file of masked scaffold lengths in decending order and sort the list of
masked scaffolds into this order
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl masked_scaffolds.list
masked_scaffolds.list.sorted masked_scaffolds_len.txt.sorted

#Split the scaffold list into N sublists (where N = the number of cores on the
machine) so that each list contains roughly the same total sequence lengths
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl scaffolds_len.txt.sorted -n
N -name scaffolds.list

#Split the masked scaffold list into N sublists (where N = the number of cores on the
machine) so that each list contains roughly the same total sequence lengths
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl
masked_scaffolds_len.txt.sorted -n N -name masked_scaffolds.list
```

Option 3: many sequences in genome (1000s)

If there are many sequences in the genome (thousands or more), then first sort sequences by length, take the first 100-1000 longest sequences and split these up into as many sublists as there are cores on the machine, then take the next 5000-10000 intermediate sequences and split these up into as many sublists as there are cores on the machine, and finally take the remaining short sequences and split these up into as many sublists as there are cores on the machine. The genome will then be run in several iterations (long first, then intermediate when long completes, then short):

```
#Note: These commands get the first 500 longest scaffolds, then the next 7500
intermediate scaffolds and then the remaining small scaffolds. You can change these
values to whatever you like as long as the number of scaffolds remains in the
```

```
recommended range of long: 100-1000, intermediate:5000-10000, short:remaining.

#Divide all scaffolds into long, intermediate and short sequence lists
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl scaffolds.list
scaffolds.list.sorted scaffolds_len.txt.sorted
head -n 500 scaffolds.list.sorted > sequence_lists/long_scaffolds.list
tail -n +501 scaffolds.list.sorted | head -n 7500 >
sequence_lists/intermediate_scaffolds.list
tail -n +8001 scaffolds.list.sorted > sequence_lists/short_scaffolds.list

#Divide all masked scaffolds into long, intermediate and short sequence lists
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl masked_scaffolds.list
masked_scaffolds.list.sorted masked_scaffolds_len.txt.sorted
head -n 500 masked_scaffolds.list.sorted > sequence_lists/masked_long_scaffolds.list
tail -n +501 masked_scaffolds.list.sorted | head -n 7500 >
sequence_lists/masked_intermediate_scaffolds.list
tail -n +8001 masked_scaffolds.list.sorted >
sequence_lists/masked_short_scaffolds.list

#Move into the sequence lists directory
cd sequence_lists/

#Create a file of long scaffold lengths in decending order
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl long_scaffolds.list
long_scaffolds.list.sorted long_scaffolds_len.txt.sorted

#Split the long scaffold list into N sublists (where N = the number of cores on the
machine) so that each list contains roughly the same total sequence lengths
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl
long_scaffolds_len.txt.sorted -n N -name long_scaffolds.list

#Create a file of intermediate scaffold lengths in decending order
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl
intermediate_scaffolds.list intermediate_scaffolds.list.sorted
intermediate_scaffolds_len.txt.sorted

#Split the intermediate scaffold list into N sublists (where N = the number of cores
on the machine) so that each list contains roughly the same total sequence lengths
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl
intermediate_scaffolds_len.txt.sorted -n N -name intermediate_scaffolds.list

#Create a file of short scaffold lengths in decending order
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl short_scaffolds.list
short_scaffolds.list.sorted short_scaffolds_len.txt.sorted

#Split the short scaffold list into N sublists (where N = the number of cores on the
machine) so that each list contains roughly the same total sequence lengths
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl
short_scaffolds_len.txt.sorted -n N -name short_scaffolds.list

#Create a file of long masked scaffold lengths in decending order
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl
```

```
masked_long_scaffolds.list masked_long_scaffolds.list.sorted
masked_long_scaffolds_len.txt.sorted

#Split the long masked scaffold list into N sublists (where N = the number of cores on
the machine) so that each list contains roughly the same total sequence lengths
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl
masked_long_scaffolds_len.txt.sorted -n N -name masked_long_scaffolds.list

#Create a file of intermediate masked scaffold lengths in decending order
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl
masked_intermediate_scaffolds.list masked_intermediate_scaffolds.list.sorted
masked_intermediate_scaffolds_len.txt.sorted

#Split the intermediate masked scaffold list into N sublists (where N = the number of
cores on the machine) so that each list contains roughly the same total sequence
lengths
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl
masked_intermediate_scaffolds_len.txt.sorted -n N -name
masked_intermediate_scaffolds.list

#Create a file of short masked scaffold lengths in decending order
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/make_sorted_seq_list.pl
masked_short_scaffolds.list masked_short_scaffolds.list.sorted
masked_short_scaffolds_len.txt.sorted

#Split the short masked scaffold list into N sublists (where N = the number of cores
on the machine) so that each list contains roughly the same total sequence lengths
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/scripts_split_seq_list/split_list_N_size.pl
masked_short_scaffolds_len.txt.sorted -n N -name masked_short_scaffolds.list
```

Setting up the Config File

Edit the below configuration file to provide the correct paths to your required input files i.e. the correct matrix file for your species, the correct parameters file (mammals vs non-mammals) the correct protein database (animals vs plants vs custom) and the mRNA evidence files. The number 1 represents “on” and the number 0 represents “off”. If both protein and mRNA evidence is turned off, the pipeline will only make *ab initio* predictions.

Note: mRNA sequences can also be used for EST evidence but this only slightly improves splice site information. RNA-seq reads can also be used to improve splice site information but may result in extended walltimes. Please note that BOTH reads and EST evidence together is currently NOT supported, so only choose one.

```
#
# Location of data and options for eukaryotic genome annotation
#

# Organism-specific and pipeline parameters

GENE_PARAM = /home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/FGENESHPIPE/EXE_CFG/species.mpar.dat  # gene prediction parameters - replace
species with the required closely-related species
PIPE_PARAM = /home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/FGENESHPIPE/mammals.par  #
location of parameters files - select either mammals or non_mamm depending on your
study species
```

```
# Predict genes with GC donor splice sites or not

PREDICT_GC = 1

# Mapping known mRNAs

MAP_mRNAs  = 1                          # map known mRNA sequences to genome sequences
CDNA_FILE  = /path/to/species.cdna      # *.cdna file for known mRNAs
PROT_FILE  = /path/to/species.pro       # *.pro  file for known mRNAs
DAT_FILE   = /path/to/species.dat       # *.dat  file for known mRNAs

# Mapping ESTs

MAP_ESTS  = 0                           # map ESTs to genomic sequences
EST_FILE  = /path/to/rna_matches.fa     # file with ESTs

# Using reads

USE_READS = 0                         # use reads info to improve gene models
DIR_SITES = /path/to/reads_sites/     # directory with reads *.sites files

# Using known proteins for prediction
# (predict genes based on homology to known proteins)

USE_PROTEINS      = 1                           # 0 - no, 1 -yes
PROG_PROT         = 1                           # 1 - use prot_map, 2 - use blast

NUM_THREADS       = 1                           # number of processors for 'prot_map'
or 'blast'

PROTEIN_DB        = /home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/FGENESHPIPE/nr_animals_par          # protein DB - select either nr_animals_par
or nr_plants (or custom protein database) depending on your study species
PROTEIN_DB_INDEX  = /home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/FGENESHPIPE/nr_animals_par.ind      # protein DB index file
PROTEIN_DB_TAG    = NR                          # short name for protein DB

BLAST_AI_PROTEINS = 1         # find homologs for ab initio predicted genes (0 - no, 1
- yes)

# Location of BLAST+ or BLAST programs

# BLAST+

BLASTP  = /usr/bin/blastp     # blastp (protein vs. protein DB)
BLAST2  = /usr/bin/blastp     # blast 2 proteins

# BLAST
#
# BLASTP  = /home/blast-2.2.26/bin/blastall     # blastp (protein vs. protein DB)
# BLAST2  = /home/blast-2.2.26/bin/bl2seq       # blast 2 proteins

# Predicting genes in long introns of other genes

INTRONIC_GENES = 0                              # predict genes in long introns of
other genes
```

Running FGENESH++

Create a list of FGENESH++ commands (one for each sequence sublist that was generated earlier) and run each command on a different core using GNU parallel:

```
#Make results directory
mkdir results

#Ensure run_pipe.pl is in your current path
run_pipe.pl

#Generate list of commands to run in parallel
cd sequence_lists
for i in scaffolds_[0-9]*.list; do echo "run_pipe.pl /path/to/species.cfg -l
/path/to/sequence_lists/${i} -m /path/to/sequence_lists/masked_${i} -d
/path/to/results" >> commands.txt; done

#Run commands
cd ../
nohup parallel < commmands.txt 2>&1&
```

Note: If you split your genome up into long, intermediate and short scaffold sublists you will perform the above 3 separate times after one another, once for each set of commands.

Converting Output and Exporting Annotated Sequences

Combining FGENESH++ output for all sequences into one large output file

```
ls -1 results > files.list

/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/merge_res_files.pl -l files.list -
dir results -sort_by_number -o species_fgenesh.resn3
```

Note: you can use a different sort option to ensure the resn3 sequence file order matches the reference genome sequence order.

Run conversions to get output into a standard format (either gff3 or GenBank)

```
#GFF3 conversion
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/Fgenesh_2_gff3.v1.20/run_fgenesh_2_gff3_multi.pl species_fgenesh.resn3
species_fgenesh.gff3 -sort -print_exons

#GenBank conversion
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/Fgenesh_2_GenBank/fgenesh_2_genbank.pl -tata -polya -div:VRT -
org_code:GS -method:FGENESH++ header species_fgenesh.resn3 ref.fasta
species_fgenesh.gb
```

Note: Refer to the `/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/Fgenesh_2_GenBank/` folder for examples of headers and run `/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/Fgenesh_2_GenBank/fgenesh_2_genbank.pl` with no options for more information on the required flags

Export annotated mRNA, CDS and protein sequences

```
/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/get_mRNAs_proteins/get_mrnas_or_GC.pl species_fgenesh.resn3
ref.sorted.fasta species_fgenesh_mrna.fa -fix_id seq_name

/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-
linux/SCRIPTS/get_mRNAs_proteins/get_mrnas_or_GC.pl species_fgenesh.resn3
ref.sorted.fasta species_fgenesh_cds.fa -fix_id seq_name -CDS

/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/get_mRNAs_proteins/get_proteins.pl
species_fgenesh.resn3 species_fgenesh_proteins.fa -fix_id seq_name
```

Note: The combined resn3 file has to have sequences in the same order as the genome fasta file for mRNA and CDS extraction to work. If order is not the same (and the `merge_res_files.pl` script does not have the required sorting option) you will either need to re-order your genome file, or export the mRNA/CDS sequences from each separate .resn3 file and then concatenate them together. See `/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS/get_mRNAs_proteins/README.txt` for more information.

Other Information

Additional scripts that may perform other desired functions are available in the `/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/SCRIPTS` directory.

The output files will likely always require some additional downstream filtering. In particular, *ab initio* predictions may need to be filtered based on match coverage or percentage identity etc. All of these statistics are retained in the gene headers so can be used for downstream filtering.

Documentation for running the FGENESH++ pipeline and preparing input files is available online or via the README file at `/home/ubuntu/FGENESHPIPE_7.2.2-x86_64-linux/FGENESHPIPE/DOC/FGENESHPIPE_README_and_OUTPUT.txt`

The Softberry team are available to contact at softberry@softberry.com for any additional questions or queries.

Optimisation required

Optimisations to the pipeline are detailed above.

Infrastructure usage and benchmarking

Exemplar 1: *Antechinus* reference genome annotation

VM Specifications

- Operating System: Ubuntu: 18.04
- RAM: 256GB
- CPUs: 64
- Attached Storage: 125GB

Antechinus Genome Specifications

- Genome Size: 3.31Gb
- No. Scaffolds: 30,876
- No. Contigs: 106,199
- Scaffold N50: 72.7Mb

- Contig N50: 78Kb

Other Data

- Antechinus transcriptome mRNA files (created from trinity + trinotate output)
- 2020 NCBI Animal NR protein database (provided by Softberry)
- Tasmanian devil gene matrix file for species-specific gene prediction parameters (purchased from Softberry)

Core-Time Comparison

- Scaffolds ~3-8Mb run in ~2hr
- Total Walltime using all 64 cores = Just over 2 days
  - Long scaffolds = 28hr
  - Intermediate scaffolds = 3hr
  - Short scaffolds = 5hr
  - **Note:** Running FGENESH++ with protein evidence only (no mRNA evidence) results in only slightly reduced walltimes and slightly reduced BUSCO scores (though predictions may not be as accurate). So, if mRNA evidence is not available for species of interest – can use protein evidence only and still retrieve good results.

Antechinus Comparisons

Type	Total walltime	Total no. predictions	Transcriptome Predictions	Protein Predictions	*Ab Initio* Predictions	Complete (Single, Dup), Fragmented, Missing Mammalian BUSCOs (%)	Unannotated Scaffolds (% of genome length)
Longest Complete Transcripts	2 Days	55827	32747 (59%)	1286 (2%)	21794 (39%)	78.2 (76.0, 2.2), 13.6, 8.2	3.1%
No mRNA evidence (Protein evidence and ab initio predictions only)	2 days	43240	-	12216 (28%)	31024 (72%)	75.9 (74.0, 1.9), 21.1, 2.9	3.4%

- **Note:** Output BUSCO scores when using mRNA evidence are often quite dependent on the completeness of the input mRNA sequences. So to ensure best results, run BUSCO on input mRNA files first.

Acknowledgements / citations / credits

Thank you to the Australasian Wildlife Genomics Group, University of Sydney for their contribution to the testing and use of FGENESH++ and the associated documentation. The AWGG work was supported by the Australian Research Council (DP180102465) and Bioplatforms Australia Oz Mammals Genome Initiative.