GigaScience

Best genome sequencing strategies for annotation of complex immune gene families in wildlife

--Manuscript Draft--

Manuscript Number:	GIGA-D-22-00064
Full Title:	Best genome sequencing strategies for annotation of complex immune gene families in wildlife
Article Type:	Research
Funding Information:	Australian Research Council (DP180102465) — Prof Katherine Belov Australian Research Council (CE200100012) — Prof Katherine Belov
Abstract:	Background The biodiversity crisis and increasing impact of wildlife disease on animal and human health provides impetus for studying immune genes in wildlife. Despite the recent boom in genomes for wildlife species, immune genes are poorly annotated in non-model species owing to their high level of polymorphism and complex genomic organisation. Our research over the past decade and a half on Tasmanian devils and koalas highlights the importance of genomics and accurate immune annotations to investigate disease in wildlife. Given this, we have increasingly been asked the minimum levels of genome quality required to effectively annotate immune genes in order to study immunogenetic diversity. Here we set out to answer this question by manually annotating immune genes in five marsupial genomes and one monotreme genome to determine the impact of sequencing data type, assembly quality and automated annotation on accurate immune annotation. Results Genome quality is directly linked to our ability to annotate complex immune gene families, with long reads and scaffolding technologies required to reassemble immune gene clusters and elucidate evolution, organisation and true gene content of the immune repertoire. Draft quality genomes generated from short-reads with HiC or 10x Chromium linked-reads were unable to achieve this. Despite mammalian BUSCOv5 scores of up to 94.1 % amongst the six genomes, automated annotation pipelines incorrectly annotated up to 60% of manually annotated immune genes regardless of assembly quality or method of automated annotation. Conclusions Our results demonstrate that long-reads and scaffolding technologies, alongside manual annotation, are required to accurately study the immune gene repertoire of wildlife species. Keywords: immune gene, genome, quality, annotation, MHC, wildlife, disease
Corresponding Author:	Carolyn Hogg AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Emma Peel
First Author Secondary Information:	
Order of Authors:	Emma Peel
	Luke Silver
	Parice Brandies
	Ying Zhu

	Yuanyuan Cheng
	Carolyn Hogg
	Katherine Belov
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	No
If not, please give reasons for any omissions below. as follow-up to "**Experimental design and statistics** Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript? "	The data included in this manuscript uses published genomic data to show complexities of immune gene annotation with varying degrees of genome quality. The method design is around genome assembly and annotation and all the relevant metrics are included in the manuscript.
Resources A description of all resources used, including antibodies, cell lines, animals	Yes

and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Best genome sequencing strategies for annotation of complex immune gene families in wildlife

Emma Peel[1], Luke Silver[1], Parice Brandies[1], Ying Zhu[2], Yuanyuan Cheng[1], Carolyn J Hogg[1] & Katherine Belov[1]*

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

[2] Sichuan Provincial Academy of Natural Resource Sciences, Chengdu, Sichuan, China

emma.peel@sydney.edu.au ORCID: 0000-0002-2335-8983

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

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

so_zy2003@126.com

yuanyuan.cheng@sydney.edu.au ORCID: 0000-0002-1747-9308

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

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

Abstract

Background

The biodiversity crisis and increasing impact of wildlife disease on animal and human health provides impetus for studying immune genes in wildlife. Despite the recent boom in genomes for wildlife species, immune genes are poorly annotated in non-model species owing to their high level of polymorphism and complex genomic organisation. Our research over the past decade and a half on Tasmanian devils and koalas highlights the importance of genomics and accurate immune annotations to investigate disease in wildlife. Given this, we have increasingly been asked the minimum levels of genome quality required to effectively annotate immune genes in order to study immunogenetic diversity. Here we set out to answer this question by manually annotating immune genes in five marsupial genomes and one monotreme genome to determine the impact of sequencing data type, assembly quality and automated annotation on accurate immune annotation.

Results

Genome quality is directly linked to our ability to annotate complex immune gene families, with long reads and scaffolding technologies required to reassemble immune gene clusters and elucidate evolution, organisation and true gene content of the immune repertoire. Draft quality genomes generated from short-reads with HiC or 10x Chromium linked-reads were unable to achieve this. Despite mammalian BUSCOv5 scores of up to 94.1 % amongst the six genomes, automated annotation pipelines incorrectly annotated up to 60% of manually annotated immune genes regardless of assembly quality or method of automated annotation.

Conclusions

Our results demonstrate that long-reads and scaffolding technologies, alongside manual annotation, are required to accurately study the immune gene repertoire of wildlife species.

Keywords: immune gene, genome, quality, annotation, MHC, wildlife, disease

48 # Background

49 Globally we are facing a biodiversity crisis, with 25% of known plant and animal species under threat
50 and one million species facing extinction [1]. Disease is one of many drivers of global wildlife decline
51 and extinction, with recent devastating examples such as chytridiomycosis in amphibians [2], white
52 nose syndrome in bats [3] and devil facial tumour disease (DFTD) in Tasmanian devils (*Sarcophilus*
53 *harrisii*) [4]. Habitat loss, fragmentation and climate change lead to population decline and subsequent
54 loss of genetic diversity, which increases susceptibility of populations to new and existing disease
55 threats [5].

56 Genomics is increasingly applied in conservation [6] facilitated by a boom in genomes for wildlife
57 species [7-10], with over 4,000 vertebrate genomes currently accessioned with the National Center
58 for Biotechnology Information (NCBI) (March 2022). Genomics in conservation typically involves
59 technologies such as reduced representation sequencing which capture single nucleotide
60 polymorphisms (SNPs) with a bias towards neutral regions of the genome [11, 12]. This can be used
61 to investigate population genetic metrics such as heterozygosity, inbreeding and relatedness to inform
62 conservation management. This is a cost-effective approach for conservation and has been used in a
63 range of taxa to inform conservation actions, for examples see Tasmanian devils [13], gorillas (*Gorillia*
64 *gorilla gorilla* and *Gorilla beringei graueri*) [14], helmeted honeyeaters (*Lichenostomus melanops*
65 *cassidix*) [15] and bilbies (*Macrotis lagotis*) [16].

66 The COVID-19 pandemic is one of many examples which highlight the ever-increasing importance of
67 understanding wildlife immunity and disease to better understand and manage disease spill over [17].
68 In the case of wildlife threatened by disease, conservation questions are more challenging to answer
69 and typically involve immunogenetic diversity which relies on accurate immune gene annotations.
70 Immune genes are some of the most polymorphic regions of the genome, owing to the need to
71 generate diversity in response to ever-changing pathogenic pressures [18, 19]. Diversity within these
72 gene families is generated through gene duplication, gene copy number variation, SNPs and rapid

evolution, resulting in a complex genomic organisation and high level of pseudogenization [18]. Generally, immune genes are encoded within clusters in the genome, especially highly duplicated families such as the major histocompatibility complex (MHC) and natural killer cell (NK) receptors [20]. Given these factors, accurate assembly and annotation of genomic regions encoding immune genes can be challenging [21-23], especially in wildlife.

Automated annotation pipelines such as MAKER [24] and Fgenesh++ [25] are accurate at identifying the majority of protein-coding genes within a genome [26, 27]. However, they are less effective at characterising complex and highly variable gene families such as immune genes [28, 29] which are misassembled even in the high-quality human genome [21]. As such, manual annotation and curation of immune genes is required, which is conducted for model organism genomes accessioned with Ensembl [30]. Wildlife are not currently included in this scope, and hence immune genes are poorly annotated, or not annotated at all, in many species.

Advances in sequencing technology means chromosome-length genomes are now achievable for a range of species [8]. Use of multiple sequencing, scaffolding, chromatin conformation and optical mapping technologies leads to accurate assembly of complex and variable genomic regions, such as immune genes [8]. However, the high input sample quantity and quality requirements are not always feasible for wildlife [31]. This leads to the use of lower-input short-read sequencing to generate a draft-quality genome assembled into scaffolds. However, short-read sequencing is well known to be incompetent at resolving highly repetitive and complex gene regions [32, 33]. While scaffolding technologies can improve contiguity of these assemblies, complex and variable regions often remain fragmented. The need to balance budget, sample and genome assembly quality against accurate immune gene annotation is essential to answer questions around disease and immunity.

Over the past decade and a half our research has focused on immunity and disease in two iconic marsupial species; the Tasmanian devil and koala (*Phascolarctos cinereus*). During this period, we have worked with bacterial artificial chromosome (BAC) and complementary DNA (cDNA) libraries and draft

genomes of varying qualities. Our research, and that of others, has been crucial for understanding, managing and preventing disease-induced decline [4, 34-36]. As the cost of sequencing has dropped, and the appreciation of the power of genetics and genomics for population management has increased, we have increasingly been asked about the minimum levels of genome quality required to be able to effectively annotate immune genes in order to study levels of diversity in wild populations. Here we set out to answer that question.

Tasmanian devils are threatened by DFTD, a contagious cancer which has decimated over 80% of the population since it was first documented in 1996 [4]. The Tasmanian devil reference genome was sequenced using illumina short-reads in 2012 [37], generating a 3.17 Gbp genome with a scaffold N50 of 1.8 Mbp and contig N50 of 20kbp. The Major Histocompatibility Complex was not able to be annotated in the draft genome due to the high levels of fragmentation, scattered across at least 15 scaffolds. But manual annotation was possible alongside transcriptomes [38-40] and targeted sequencing of MHC-positive BAC clones [38, 41-45]. Development of MHC markers led to determination of gene copy number and nucleotide variation amongst the devil population, revealing devils have low MHC diversity, much of which is shared with DFTD [43, 46]. The low histocompatibility barriers, coupled with downregulation of tumour MHC expression, allows DFTD to transmit between individuals and evade the host immune response [44]. Recent MHC genotyping using long-read sequencing enabled the identification of full-phased MHC alleles and separation of highly similar alleles (1bp difference), resulting in the identification of new functional MHC diversity within the devil population [47].

The koala is another iconic Australian marsupial where disease is a major contributing factor to population decline [48]. Chlamydiosis is one of many threatening processes affecting koalas, a disease caused by infection with the intracellular bacterium *Chlamydia pecorum* [48]. A chromosome-length koala reference genome was sequenced in 2018 using Pacfici Biosciences (PacBio) long-reads, Illumina short-reads and BioNano optical maps [49]. This generated a 3.19 Gbp assembly with a scaffold N50

123 of 480 Mbp and contig N50 of 11.4 Mbp [49], a 400-fold increase in scaffold contiguity compared to
124 the Tasmanian devil genome assembly [37]. This high-quality koala genome enabled accurate
125 annotation of immune gene families, including the first complete reconstruction of MHC and T cell
126 receptor gene clusters from a genome sequence in marsupials [35, 50-52]. Preliminary genome
127 resequencing identified that variants within IFNγ, TNFα and MHC genes are essential for clearance of
128 *Chlamydia* in koalas [34]. MHC genotype has also been linked to disease susceptibility and severity in
129 different koala populations [53, 54].

130 Understanding the role of immunogenetics in DFTD and chlamydiosis formed the basis for the
131 development of vaccines in devils [55] and koalas [56] respectively. Several chlamydial vaccines have
132 been developed for koalas over the past decade, culminating in a multivalent vaccine that induces a
133 strong and protective immune response which may be therapeutic [34, 56]. Current DFTD vaccines in
134 devils similarly hinge upon the host immune response and are based on IFNγ-treated DFTD cells which
135 result in MHC expression on the cell surface resulting in immune recognition and clearance of the
136 tumour [55].

137 In this study, our aim was to determine the impact of sequence data type, assembly quality and
138 automated annotation on accurate immune annotation. To achieve this, we manually annotated
139 immune genes in the genomes of five marsupials and one monotreme. These include recent published
140 genome assemblies of five marsupials; koala (*Phascolarctos cinereus*) [49, 57, 58], woylie (*Bettongia*
141 *penicillata*) [59], common wombat (*Vombatus ursinus*) [57, 58], brown antechinus (*Antechinus*
142 *stuartii*) [60] and numbat (*Myrmecobius fasciatus*) [61], and previous immune gene annotations from
143 one monotreme, the platypus [33]. These six genomes differ in quality, from scaffold assemblies
144 generated using only 10x Chromium linked-reads (numbat, antechinus), short-read with high-
145 throughput chromosome conformation capture (HiC) (wombat), long and short-read (woylie), to high-
146 quality chromosome-length genomes generated using multiple data types (koala and platypus) (Table
147 1). In addition, we assess the accuracy of automated immune gene annotation by Fgenesh++, MAKER

148 and NCBI pipelines in these non-model species. Although this is not a perfect comparison given
149 species-specific immune gene expansion/contraction, it provides a guide of the impact of genome
150 quality on immune gene annotation. Here we show that high quality chromosome-length genomes
151 are necessary for accurate immune annotation in the context of wildlife disease.

152 Analyses

153 Immune genes were annotated in the koala, woylie, wombat, antechinus, and numbat genomes and
154 transcriptomes using similarity-based search methods such as BLAST [62] and HMMER [63] with
155 known marsupial immune gene sequences as queries. This resulted in the manual characterisation of
156 over 2,700 immune genes amongst the five species, from six immune gene families or groups: toll-like
157 receptors (TLR), T cell receptors (TCR), immunoglobulins (IG), major histocompatibility complex
158 (MHC), natural killer (NK) cell receptors and cytokines (Table 2). Platypus immune gene families have
159 previously been annotated [33, 64-75], some of which had already been mapped within the current
160 genome assembly (MHC and TCR) [33] and the remainder were mapped in this study. Genomic
161 coordinates of all immune genes annotated in this study are available in Additional file 1. A
162 comprehensive summary of results for each immune gene family are available in Additional file 2.

163 Table 1. Assembly metrics for the five marsupial and one monotreme genome used in this study.

	Koala [49, 57, 58]	Woylie [59]	Wombat [57, 58]	Antechinus [60]	Numbat [61]	Platypus [33]
Data types	PacBio RS II Illumina BioNano HiC (DNAzoo) RNAseq (16 transcriptomes)	PacBio HiFi Illumina RNAseq (4 transcriptomes)	Illumina HiC (DNAzoo)	10x Chromium RNAseq (12 transcriptomes)	10x Chromium RNAseq (3 transcriptomes)	PacBio 10x Chromium BioNano HiC (Phase genomics & Dovetail) RNAseq (19 transcriptomes)
Genome size (Gbp)	3.19	3.39	3.34	3.31	3.42	2.13
GC (%)	39.05	38.64	38.89	36.20	36.3	46.23
No. scaffolds	1,318	1,116	633,737	30,876	112,299	322
No. contigs	1,935	3,016	685,859	106,199	219,447	834

Scaffold N50 (Mbp)	480.11	6.94	576.1	72.7	0.223	83.33
Contig N50 (Mbp)	11.4	1.995	0.07	0.08	0.038	15.1
Gaps (%)	0.01	0.403	0.54	2.75	3.52	0.81
Complete mammalian BUSCOv5.2.2	94.1%	94.1%	89.3%	92.5%	76.4%	83.0%

Table 2. Number of annotated immune genes in each of the five marsupials and one monotreme in this study.

	Koala	Woylie	Wombat	Antechinus	Numbat	Platypus
TLR	10	10	10	10	10	10
TCR constant	10	12	10	11	9	19
TCR variable	103	122	95	126	104	252
IG constant	15	20	10	7	6	14
IG variable	289	226	98	145	121	118
MHC I	19	17	5	7	3	6
MHC II	16	23	7	14	8	5
MHC III	37	37	41	32	34	58
Ext. MHC & framework genes	27	28	31	25	32	20
NKC	17	17	11	11	17	122
LRC	25	60	32	49	38	4
Extended LRC	6	24	9	15	11	11
Cytokines	83	76	76	68	70	49
Total	657	672	435	520	463	678

Table 2 legend. Includes complete and partial gene sequences. A more detailed comparison of immune genes annotated in this study, with those identified in other marsupials and humans is available in Supplementary Table 2 within Additional file 2.

Overall, the immune gene repertoire of the koala, woylie, wombat, antechinus, and numbat was similar to other marsupials [50, 76], with marsupial-specific genes and eutherian orthologs identified. Relatively conserved immune genes such as TLRs and constant regions of TCR and IG, as well as polymorphic genes such as MHC and NK receptors, were identified in all five species. Numerous koala immune gene sequences have been characterised previously due to their involvement in chlamydiosis

176 and koala retrovirus which threaten populations [48]. These include MHC [49, 77-79], IG [50], TCR
177 [49], NK receptors [51] and selected cytokines [50, 80-83] (Supplementary Table S2 in Additional file
178 2). We mapped the location of these genes within the current version of the genome, and identified
179 additional new sequences within the LRC, IG and cytokine families (Table 2, Supplementary Table S2
180 in Additional file 2). Immune genes unique to the marsupial lineage were also characterised in the five
181 species studied here. These included MHC class II genes DA, DB and DC, TLR1/6 and TCRμ. Large
182 marsupial-specific gene expansions within the LRC NK receptors were characterised in all five species,
183 as well as reduced gene content within the NKC cluster of NK receptors. Consistent with other
184 marsupials investigated to date Igδ was not found in any of the five assemblies [84]. A detailed outline
185 of immune genes annotated in this study compared to those of other marsupials and humans is
186 provided in Supplementary Table S2 within Additional file 2.

187 Automated versus manual immune gene annotation

188 For woylie, wombat, antechinus, numbat, and platypus genomes, we assessed how well our manual
189 immune gene annotation aligned with automated annotations by Fgenesh++ (woylie, antechinus, and
190 numbat), MAKER (wombat) and the NCBI pipeline (platypus). The koala was not included in the
191 comparison as the genome available on DNAzoo
192 (https://www.dnazoo.org/assemblies/Phascolarctos_cinereus) has not been annotated. Inclusion of
193 the platypus NCBI annotation ensures that any differences in automated and manual immune gene
194 annotation are not due to deficiencies within the Fgenesh++ annotation pipeline, as the woylie,
195 antechinus and numbat genomes were all annotated with Fgenesh++ using the same parameters.

196 Automated annotation pipelines failed to characterise the complete immune repertoire of the
197 platypus or any of the four marsupial species (Figure 1). Only 24.65%, 21.32%, 21.32%, 29.66%, 30.97%
198 of immune genes were correctly annotated by the automated pipeline in platypus, woylie, wombat,
199 antechinus, and numbat respectively, defined as ≥90% overlap in genomic coordinates of immune
200 genes between our manual annotations and the automated annotations (Figure 1). Interestingly, more

201 immune genes were correctly annotated by the automated software in the low-quality wombat,
202 antechinus, and numbat genomes than the high-quality platypus and woylie genomes. This inverse
203 relationship between genome quality and proportion of correctly annotated immune genes is likely
204 related to the characterisation of additional divergent and polymorphic genes such as MHC class I and
205 II in woylie and platypus, which could not be identified by automated or manual annotation in the
206 wombat, antechinus, and numbat due to genome fragmentation (Table 3). The platypus and all four
207 marsupial genomes displayed a high proportion of immune genes which were very poorly annotated
208 by automated pipelines (≤10% overlap between immune gene coordinates from manual versus
209 automated annotation); 34.48%, 47.46%, 69.98%, 31.79% and 31.17% for platypus, woylie, wombat,
210 antechinus, and numbat respectively (Figure 1). Most of these genes comprised immunoglobulin and
211 T cell receptor variable gene segments, and species-specific gene expansions in NKC and LRC families,
212 indicating the difficulty in automated annotation of these regions (Figure 3).

213 Figure 1. Percentage overlap of genomic coordinates between manual and automated annotations of
214 immune genes in five genomes.

215 Figure 1 legend. Colours indicate proportion of immune genes with 0 to 100% overlap between manual
216 and automated annotations, with 0 indicating manually annotated genes with no overlap of genomic
217 coordinates with the automated annotation.

218 Relationship between genome quality and manual immune gene annotation

219 Manual annotation of immune genes across the koala, woylie, wombat, antechinus and numbat
220 genomes, and mapping of previous annotations to the new platypus genome, highlighted a clear
221 relationship between immune gene fragmentation and genome quality (Figure 2). Overall, the high-
222 quality koala, platypus and woylie genomes all contained complete immune gene family clusters,
223 which were highly fragmented in the lower quality wombat, antechinus, and numbat genomes.
224 Fragmentation was particularly evident within families which contain genes that do not share
225 orthology to those in eutherians, such as LRC NK receptors and TCRμ, and highly duplicated families

226 such as MHC (Figure 3). To investigate this relationship further, we calculated the number of scaffolds
227 which encoded 50% (L50) and 90% (L90) of manually annotated immune genes in each of the five
228 species studied (Figure 2).

229 Figure 2. L50 and L90 immune gene metric for six genomes, compared to $\log_{10}$ contig N50.

230 The platypus, koala and woylie had an L90 of 10, 9 and 36 respectively, which suggests immune gene
231 families were highly contiguous within all three genomes (Figure 2). Complete coding sequences were
232 identified for 98% and 95% of immune genes in koala and woylie respectively. In addition, 90% of
233 annotated immune genes were located on scaffolds greater than 33.3 Mbp, 75 Mbp and 1 Mbp in
234 platypus, koala, and woylie respectively. Complex multi-gene immune families such as MHC, NK
235 receptors and TCR were highly intact in all three species. The koala and woylie MHC regions were both
236 located on a single scaffold (Figure 3). Class I and II genes were interspersed, and flanked by class III,
237 framework and extended class I and II gene clusters, which reflected the MHC organisation of other
238 marsupials (Figure 3) [49, 85]. Unlike marsupials, the platypus MHC is encoded within a
239 pseudoautosomal region of two sex chromosomes. MHC class I and II genes were interspersed in a
240 single cluster on chromosome X3, and class III, extended class I and II, and framework genes located
241 in a single cluster on chromosome X5 (Figure 3) [33]. Large gene expansions within the LRC NK
242 receptors were encoded on a single scaffold in koala and six scaffolds in woylie (Figure 3). The number
243 and type of monotreme NK receptor genes differs to marsupials, as they have a large expansion within
244 the NKC gene cluster and reduction within the LRC gene cluster [66]. More than 80% of platypus NKC
245 genes were located in a single cluster on chromosome 17, with LRC genes located on 5 different
246 chromosomes [66]. Fragmentation of the LRC cluster is not a factor of genome quality but reflects the
247 evolutionary history of this immune family [66]. The four TCR loci (α/δ, β, γ and μ) were encoded in
248 single clusters on three chromosomes in platypus and single scaffolds in koala. The TCR loci were
249 fragmented across up to three scaffolds in woylie. This includes genes known to flank these loci in

250 other marsupials, which enabled resolution of TCR locus organisation in these species, and confirmed
251 gene synteny across marsupials, human and mouse as identified previously [85].

252 Figure 3. Genomic organisation and gene content of the LRC (A) and MHC region (B) in five genomes.

253 Figure 3 legend. The number of genes within each cluster are given, as well as scaffold counts of
254 orphan genes (genes on single scaffolds). In A, LRC genes are purple, extended LRC genes are teal. In
255 B, MHC class I genes are red, class II blue, class III green, extended class I pink, extended class II yellow
256 and framework genes orange. Large distances between genes are given below the scaffold, otherwise
257 the distance between genes and/or clusters was within the expected range for each family. Figure
258 created with BioRender.com.

259 Fragmentation of immune genes in the wombat genome differed between immune families, with an
260 L90 of 56 (Figure 2). 22% of scaffolds encoding immune genes were shorter than 100Kb and partial
261 coding sequences were identified for 7% of annotated immune genes. The MHC region was relatively
262 contiguous in the wombat, with 92% of genes encoded on a single scaffold (Figure 3). Although, a
263 number of MHC genes were encoded as orphan genes to the main MHC cluster, indicating this family
264 is misassembled in the wombat genome. In addition, some MHC genes could not be identified in the
265 wombat genome, while only single copies could be identified for others which are known to be
266 duplicated in all other marsupials studied to date (Additional file 2). While this reduced MHC gene
267 content in the wombat may reflect the true MHC gene repertoire of this species, it is likely MHC genes
268 could not be annotated due to assembly error. The LRC cluster was highly fragmented across 16
269 scaffolds (Figure 3), of which more than 80% encoded a single gene and were less than 10kb in length.
270 Extended LRC and LRC genes were interspersed, likely due to mis-assembly of the region as these
271 genes should be located in separate clusters as observed in koala and woylie (Figure 3). TCRα, β and γ
272 loci were encoded on individual scaffolds, however TCRμ was fragmented across 10 scaffolds, with
273 34% of genes located on individual scaffolds of less than 15Kb. While the TCRβ locus was encoded in

274 a single cluster in the wombat, half of the locus was in the reverse orientation. This organisation is
275 unusual amongst mammalian TCR and is likely a result of the HiC scaffolding.

276 Immune gene families were highly fragmented in the antechinus and numbat genomes, with an L90
277 of 156 and 218 respectively (Figure 2). 29% and 43% of immune genes were located on scaffolds less
278 than 100Kb, and partial coding sequences were identified for 5.7% and 10.8% of immune genes, in
279 antechinus and numbat respectively. Complex multi-gene families such as MHC, NK receptors and TCR
280 were highly fragmented, with individual genes or exons located on short scaffolds. While 86% of MHC
281 genes were located on a single scaffold in antechinus (Figure 3), genome fragmentation prevented the
282 identification of additional MHC genes, hence the true MHC gene content could not be determined.
283 The numbat MHC region was highly fragmented across 52 scaffolds, 63% of which were less than
284 100Kb in length (Figure 3). Large gene expansions of LRC NK receptors were fragmented across 34
285 scaffolds in antechinus and numbat, of which 67% (antechinus) and 35% (numbat) were less than
286 10Kb, and 76% of scaffolds encoded individual LRC genes in both species (Figure 3). Similar to wombat,
287 extended LRC and LRC genes were interspersed, likely a mis-assembly as these genes should be
288 encoded within separate clusters as observed in koala and woylie. All four TCR loci were fragmented
289 in numbat, and all except TCRα in antechinus, with individual loci encoded across up to 6 scaffolds in
290 numbat and 19 in antechinus. Low contiguity within genomic regions encoding immune gene families
291 in the antechinus and numbat limited investigation of genomic organisation, synteny and evolution in
292 these species.

293 Discussion

294 By manually annotating immune genes in five marsupial and one monotreme genome of varying
295 qualities, we have confirmed that genome quality is directly linked to our ability to annotate complex
296 immune gene families. Without long reads and scaffolding technologies, immune genes are scattered
297 across many individual scaffolds and gene family organisation and evolution cannot be elucidated. We

298 conclude that a kitchen sink approach, that uses long-read data combined with HiC technology, to
299 generate a high-quality genome assembly is required to investigate immunity and disease in wildlife.

300 The immune gene repertoire of the koala, woylie, wombat, antechinus and numbat was similar to
301 other marsupials such as Tasmanian devil [38, 41, 45], tammar wallaby (*Macropus eugenii*) [68, 86-89]
302 and grey short-tailed opossum (*Monodelphis domestica*) [76]. The platypus immune gene repertoire
303 has been characterised previously [33], and we identified their location within the current genome
304 assembly. Fewer MHC genes were identified in the wombat, antechinus, and numbat, compared to
305 the platypus, koala, and woylie (Table 2, Supplementary Table S2 in Additional file 2). This is likely due
306 to poor read assembly within this highly variable and duplicated region of the genome, rather than a
307 true reduction in MHC gene content within these three species. The assembly of a complete MHC
308 cluster in the platypus, koala and woylie is due to the ability of long reads to span duplicated and
309 variable sequences, which enables assembly algorithms to accurately reconstruct this complex region
310 of the genome.

311 Automated annotation poorly characterises immune genes in non-model species

312 Despite mammalian BUSCOv5 scores of up to 94.1% amongst the six genomes in this study, indicating
313 that the genomes were "functionally complete", on average 42% of immune genes were not
314 accurately annotated and up to 61% of genes were not annotated by the automated software
315 Fgenesh++ and MAKER, nor the NCBI pipeline, compared to our manual annotations (Figure 3). The
316 majority of immune genes incorrectly annotated or missing from the automated annotations were
317 variable segments of TCR and IG, or divergent genes such as MHC with low or no BLAST homology to
318 nucleotide or protein databases. Gene models generated by automated annotation software are
319 hypotheses based on supporting evidence such as RNAseq data, which was used as evidence for the
320 automated annotation in the four marsupial and platypus genomes. While immune transcripts were
321 identified in the transcriptomes from these species, RNAseq data did not provide enough evidence to
322 support gene models for ~40% of immune genes within the genome. Some immune genes may not

323 have been expressed in the tissue sequenced, were expressed at low levels, or were fragmented. For
324 human and mouse, comprehensive and curated gene sets such as GENCODE and RefSeq are available
325 to guide gene model predictions, comprising data from more than 10,000 RNA experiments and
326 decades of dedicated work in this field [90, 91]. Given time, budget and sample constraints for wildlife,
327 these curated gene sets are not available, hence RNAseq evidence is incomplete resulting in deficient
328 gene models by automated annotation software.

329 It is not surprising that TCR and IG V segments were not automatically annotated in the genomes in
330 this study. These genes are notoriously difficult to characterise and are manually annotated in the
331 human and mouse genome on Ensembl using the International Immunogenetics Information System
332 (IMGT) database [30, 92]. Alignment of mature IG and TCR sequences from RNAseq data to the
333 genome results in poor automated annotation, as V segments utilize different sequence signal splice
334 sites to introns, which are not recognized by the open reading frame prediction algorithms. V
335 sequences from three marsupials and two monotremes are available in IMGT, however as non-model
336 species, they are not included in the scope for manual annotation by Ensembl or NCBI, so these
337 important functional features are not annotated.

338 Our results highlight the importance of manual annotation of complex and variable immune genes,
339 and caution reliance on BUSCO metrics to assess functional completeness of a genome. If this pattern
340 is observed more widely across non-model species and other complex gene families, functionally
341 important genes may not be accurately represented in genome annotations, which will flow on to
342 downstream applications [28, 93]. While automated annotation is required to keep pace with the
343 rapid sequencing of genome assemblies, manual gene characterisation is still the gold standard for
344 genome annotation [90] and is conducted for the human, mouse, zebrafish and rat genomes on
345 Ensembl [94]. For non-model species, manual annotation is conducted by individual research groups
346 following genome assembly accession with NCBI or Ensembl, which conduct in-house automated
347 annotation for some but not all species [95, 96]. These highly valuable manual gene annotations are

348 not incorporated into the Ensembl annotation release but are often listed in the supplementary
349 materials of multiple individual publications. NCBI does have some capacity to incorporate manual
350 changes to existing annotation records [97]. Changes to multiple annotations, such as adding new
351 genes as is the case in this study, require the genome to be re-annotated, which is not feasible for all
352 research groups. In addition, it is not a requirement for manual changes to annotations to be tracked
353 between genome versions, hence this information could easily be lost. Given NCBI and Ensembl
354 annotations are widely used by the scientific community, these institutions should consider
355 incorporating manual gene annotations into the annotation record or provide scope for permanently
356 storing this valuable data alongside the respective assembly.

357 Genome quality correlates with immune gene fragmentation

358 As expected, we found that genome quality directly correlates with likelihood that an immune gene
359 family was assembled and annotated correctly. Immune genes fragment as genome quality declines
360 (Figure 2 and 3). This highlights the importance of long reads and HiC scaffolding to re-assemble
361 complex gene families (platypus, koala, woylie), which are poorly assembled in short read and linked-
362 read assemblies (wombat, antechinus, numbat). Figure 4 provides a graphical representation of the
363 impact of different sequencing technologies on the assembly and fragmentation of immune gene
364 clusters. When the average read or contig length is shorter than the gene length, the assembly
365 algorithm is unable to reconstruct genes, which are fragmented across multiple short contigs [93]. The
366 average immune gene in this study was ~10 kbp in length. Long reads greater than 10 kbp in the
367 platypus, koala and woylie genomes were able to span these genes, whereas the ~150 bp short reads
368 in the wombat, antechinus and numbat genomes were insufficient to re-assemble the entire gene,
369 resulting in gene fragments on short scaffolds. Gene families with copy number variation such as MHC
370 and NK receptors are notoriously difficult to assemble and annotate [18, 21], so it is not surprising
371 these gene families were highly fragmented in the antechinus and numbat genomes. Gene copies
372 within these families can contain almost identical domains, may be pseudogenes and are encoded in
373 clusters within the genome [28]. For example, koala NK LRC genes share up to 96% amino acid

374 sequence identity and are encoded within a single cluster. For these reasons, assembly and annotation
375 of MHC and NK receptors have been used to illustrate improvements in assembly quality. For example,
376 MHC class I genes were located on a single contig in a recent release of the human genome [21],
377 however the highly repetitive MHC class II locus remains unresolved [21].

378 Figure 4. Impact of different sequencing technologies on the assembly of immune gene clusters such
379 as the MHC.

380 Figure 4 legend. The impact of long-read (A – platypus, koala and woylie), short-read (B – wombat)
381 and 10x Chromium linked read (C – antechinus and numbat) sequencing technologies, alone or in
382 combination with HiC scaffolding (i – koala & platypus, and ii – wombat), on the assembly of complex
383 and repetitive immune gene clusters such as the MHC. Colour gradient represents gene orientation
384 (A) Long read sequencing generates reads which span complex and repetitive sequences, resulting in
385 long contigs and scaffolds which contain multiple immune genes with complete coding sequences. (B)
386 Short-read sequencing generated reads which are unable to span immune genes, hence reads are
387 assembled into multiple short contigs which end when the algorithm is unable to assemble a repetitive
388 and complex immune gene sequence. (C) In linked-read sequencing, individual DNA molecules are
389 partitioned into gel beads and identical barcodes attached, then sequenced using short-read
390 technology resulting in read clouds [98]. As no individual read within the cloud spans the entire length
391 of the DNA molecule, the algorithm is unable to assemble repetitive and complex sequences, resulting
392 in multiple short contigs similar to a short-read assembly. Short contigs in B and C result in
393 fragmentation of immune genes, leading to false pseudogenization and "missing" genes. (i) HiC
394 sequencing provides contact information for DNA sequences located in close proximity within the
395 nucleus, as frequency decreases with increasing linear distance within the genome assembly [99]. This
396 contact information can be used to cluster, order and orient contigs into chromosome-size scaffolds
397 [100]. Long contigs scaffolded with HiC result in near-complete reconstruction of immune gene
398 clusters. (ii) Short contigs scaffolded with HiC generates what appears to be long scaffolds, however

399 complex immune gene clusters are incomplete. As multiple HiC contacts can span the length of the
400 contig, the correct contig orientation is not apparent leading to inversions and mis-placed contigs
401 during scaffolding. This leads to incorrect orientation of genes, which can cause pseudogenization
402 and/or gene fragmentation. Manual immune gene annotation reveals that the true gene complement
403 of the immune cluster is not contained within the scaffolded sequence. Figure created with
404 BioRender.com.

405 HiC scaffolding of contigs derived from platypus and koala long reads resulted in complete and
406 accurate reassembly of immune gene clusters in both genomes (Figure 4 A). Conversely, HiC
407 scaffolding of contigs from wombat short reads resulted in immune gene fragmentation (Figure 4 B),
408 reflected in the high immune gene L90 for the wombat genome (Figure 2). Both the koala and wombat
409 genomes were scaffolded with DNAzoo HiC data using the same 3D-DNA pipeline [57, 58, 101]. This
410 result underscores the importance of assessing annotations when determining genome quality, as the
411 wombat genome is classified as chromosome-length yet is highly fragmented within functionally
412 important genomic regions. Input genome assembly contiguity is known to influence HiC scaffolding
413 ordering and orientation errors [102], despite claims that HiC scaffolding with 3D-DNA generates
414 chromosome-length scaffolds from US$1,000 short read contigs [57]. Problems with HiC scaffolding
415 within repetitive and duplicated regions are well documented [23, 102, 103], which is exacerbated by
416 short contigs [102]. Modelling of human genome scaffolding performance using 3D-DNA revealed
417 scaffold chimeras, ordering and orientation errors increased as contig length decreased [102]. While
418 the koala and platypus genomes used as input to HiC scaffolding benefited from polishing with short
419 read data and optical mapping [49], HiC scaffolding is insufficient to recover the majority of immune
420 clusters from a fragmented genome.

421 The 3D-DNA pipeline orientates contigs within scaffolds by maximizing contact frequency between
422 contig ends [58]. Short contigs, such as those from the wombat, would have multiple contacts that
423 span the length of the contig. This means both true and false contig orientations would have a similar

424 frequency, resulting in errors such as the partial inversion of the TCRB locus which is likely false
425 (Additional file 2). At a gene level, these errors lead to the misplacement of genes on short scaffolds
426 outside the main immune cluster and false pseudogenisation (Figure 4 B). Long contigs, such as those
427 from the koala, would have fewer contacts that span the length of the contig, hence the true
428 orientation of the contig would be clear from the higher contact frequency at the correct joining end.
429 The combination of long contigs which span repetitive and highly heterozygous regions with HiC
430 scaffolding maximizes contiguity within immune gene clusters (Figure 4 A).

431 10x Chromium linked-read sequencing was insufficient to accurately re-assemble immune gene
432 clusters in our study (Figure 4 C). Complete marsupial immune gene clusters can span hundreds of
433 kilobases to megabases, as shown by annotation of the complete MHC, NK receptor and TCR regions
434 in the koala (Additional file 2). DNA molecules input to 10x library preparation were on average 74
435 kbp and 23 kbp in antechinus and numbat respectively. For most immune gene clusters, these
436 molecules would not span an entire cluster, nor even multiple immune genes in the case of the
437 numbat. This is reflected in our results, where smaller immune clusters such as the 70 kbp TRG locus
438 were intact in the antechinus, while no cluster was intact in the numbat. Interestingly, the antechinus
439 MHC cluster appears to be intact (Figure 3), however manual annotation revealed multiple genes were
440 "missing" within the scaffold and instead were located on individual short scaffolds. Even in humans,
441 10x linked reads are unable to resolve repetitive sequences which are larger than the input DNA
442 molecule [104]. Molecule length is influenced by input DNA quality [105], which was >40 kbp for both
443 antechinus [60] and numbat. DNA from human blood and cell lines routinely achieve molecule lengths
444 greater than 100 Mbp [106]. Given the challenges surrounding sampling of wildlife, this outcome
445 would be unlikely for many wildlife genomics projects using 10x linked read sequencing.

446 Regardless of input DNA molecule length, 10x libraries are still subject to the limitations of short-read
447 sequencing regarding assembly of complex sequences. Antechinus and numbat 10x libraries were
448 sequenced as short ~150 bp reads, hence while reads can be assigned back to the corresponding input

449 DNA molecule, no single read spans the molecule length. Gaps between the reads make *de novo*
450 assembly of repetitive and complex immune sequences difficult, often resulting in termination of
451 contig extension and gene fragments scattered across short scaffolds [104, 107, 108]. These gene
452 fragments can be misinterpreted as pseudogenes owing to loss of up/downstream coding regions
453 (Figure 4C). For example, antechinus and numbat NK LRC genes share up to 97% and 98% amino acid
454 sequence identity amongst the 91 and 70 immunoglobulin superfamily (IGSF) domains identified in
455 each species respectively. The LRC should be encoded within a single cluster, as in the koala genome
456 (Figure 3). Instead, the antechinus and numbat LRC clusters are fragmented across 33 and 34 scaffolds
457 respectively.

458 As the global biodiversity crisis deepens, the need to sequence eukaryotic life while it remains is
459 imperative [1, 7, 8]. High quality genomes, using a combination of long-read and HiC, have recently
460 been generated for a number of wildlife species [8], which have been used to answer questions
461 involving chromosome evolution [109], comparative genomics [110] and runs of homozygosity [111]
462 amongst others. Our results show that high-quality genomes are also necessary to study immune
463 genes in wildlife.

464 Draft quality *de novo* genomes, in this study the antechinus and numbat (linked reads), have limited
465 capacity for usefully informing immunogenetics studies as only partial sequences will be identified for
466 most immune genes. A scaffold-quality genome, in this study the woylie (long-reads) or wombat
467 (short-reads with HiC), would be suitable for immune marker development targeting most immune
468 gene families, and studying TCR and IG diversity. Long-reads will provide contiguity within duplicated
469 MHC and NK families, which should reassemble into complete clusters. HiC data may resolve some
470 immune gene clusters from a short-read assembly, however, may introduce errors as discussed
471 earlier. Finally, the kitchen sink approach, in this study the platypus and koala genomes (multiple data
472 types), will accurately assemble immune gene clusters, which is essential for investigating genomic
473 organisation, synteny and evolution. In the context of wildlife disease, it may be necessary to wait for

474 an opportunistic sample from an individual that is euthanised, or acquire ethics to euthanise an
475 individual, in order to obtain sufficient sample quantity and quality to generate high-quality
476 chromosome-length genome assemblies and associated transcriptomes [7, 9, 112-114].

477 Potential implications

478 The biodiversity crisis and increasing impact of wildlife disease on animal and human health provides
479 impetus for studying immune genes in wildlife. Genomes are now available for many wildlife species,
480 however utility of these assemblies for annotating complex immune gene families is unknown. We
481 have provided an assessment of complex immune gene annotation across genomes of varying quality,
482 using immune genes in five marsupials and one monotreme as an example. Genome quality directly
483 influenced the reassembly of immune gene clusters, and ability to investigate evolution, organisation,
484 and true gene content of the immune repertoire. A high-quality genome generated from long-reads
485 with HiC accurately assembles immune gene clusters. However, draft-quality genomes generated
486 from short-reads with HiC or 10x Chromium linked-reads were unable to achieve this. Aside from
487 genome quality, manual annotation of immune genes is required to cover the shortfall in deficient
488 gene models used by automated annotation software. Our results highlight the limitations of different
489 sequencing technologies and established workflows for genome annotation and quality assessment,
490 when applied to non-model species and the investigation of wildlife disease and immunity.

491 Methods

492 Five published marsupial genomes, koala [49, 57, 58], woylie [59], wombat [57], antechinus [60] and
493 numbat [61] (Table 1), and one monotreme genome, platypus [33], were selected for this study based
494 on use of different sequencing technologies (alone and in combination) and variation in assembly
495 quality. These include assemblies generated using multiple data types (koala and platypus), long and
496 short-reads (woylie), short-reads and HiC (wombat) or 10x Chromium linked-reads (antechinus and
497 numbat). BUSCO scores were generated by uploading the six genome assemblies to the Galaxy web

498 platform [115], where the public server at galaxy.org was used to run BUSCOv5.2.2 [27] against the
499 mammalian database.

500 Immune genes were annotated in the koala (phaCin_unsw_v4.1_HiC) [49, 57, 58], antechinus
501 (anrechinusM_pseudohap2.1) [60], woylie (mBetpen1.pri.20210916) [59], wombat (vu-2k) [57, 58]
502 and numbat genome (mMyrfas1.pri.20210917) [61] using multiple search strategies. BLAST was used
503 to search genome assemblies, associated annotation files and/or transcriptomes using published
504 marsupial, monotreme and eutherian immune gene sequences as queries, with default parameters
505 and an e-value threshold of 10 so as not to exclude any potential gene candidates. HMMERv3.2 [116]
506 was also used to identify putative genes within immune families that are known to contain
507 duplications in other marsupials, such as NK receptors. Hidden markov models (HMM) were
508 constructed using ClustalW alignments of published marsupial and eutherian immune gene sequences
509 constructed in BioEditv7.2.5 [117], which were then used to search all genomes and transcriptomes
510 using HMMER v3.2 with an e-value threshold of 10. For variable segments of T cell receptor and
511 Immunoglobulin families, recombination signal sequences (RSS) downloaded from the IMGT database
512 [92] and published koala sequences [49], were aligned using ClustalW in BioEditv7.2.5 [117]and used
513 to construct HMM. These RSS HMM were then used to search each genome using HMMERv3.2 [116],
514 to identify conserved RSS which flank each variable segment. For NK receptors, putative NKC and LRC
515 sequences from BLAST+v2.7.1 [62] and HMMERv3.2 [116] searches were queried against the swissprot
516 nonredundant database, and any sequences with top hits to swissprot NK genes, marsupial-specific
517 NK genes or the protein families database (Pfam) [118]immunoglobulin domain PF00047 or C-type
518 lectin domain PF00059 HMM model were retained. IGSF domains within putative NK sequences from
519 each species were identified using the simple modular architecture research tool (SMART) database
520 [119], and IGSF domains within 5 kbp were considered exons of a single LRC gene. Putative immune
521 genes were named following the appropriate nomenclature for each family, with duplicated genes
522 named according to their genomic location from the 5’ to 3’ end of the locus. For each immune gene
523 family, amino acid sequences from all five species, in addition to other marsupial, monotreme and

524 eutherian sequences, were aligned using ClustalW in BioEditv7.2.5 [117]. This alignment was then
525 used to construct neighbour-joining phylogenetic trees in MEGAXv10.2.4 [120] using the p-distance
526 method, pairwise deletion and 1000 bootstrap replicates.

527 Additional files

528 File name: Additional file 1

529 File format: .xls

530 Title of data: Supplementary Table S1

531 Description of data: Genomic coordinates of manually annotated immune genes in the koala, woylie,
532 wombat, antechinus and numbat genomes. The genomic coordinates of published platypus immune
533 genes used in this study are also included.

534 File name: Additional file 2

535 File format: .doc

536 Title of data: Supplementary results

537 Description of data: A comprehensive comparison of manually annotated immune genes in this
538 study to those in other marsupials and humans is provided in Supplementary Table 2. For each
539 immune gene family characterised in this study, a summary of results and phylogenetic analysis is
540 provided. This includes genes encoding toll-like receptors, natural killer receptors, cytokines
541 (interferons, interleukins and tumour necrosis factors), T cell receptor constant and variable regions
542 (all five chains in marsupials and monotremes), immunoglobulin constant and variable regions
543 (heavy and light chains) and major histocompatibility complex class I, II and III genes. Additional file 2
544 contains 7 tables and 14 figures.

545 # Data availability

546 The published woylie and numbat genome and global transcriptome assemblies are available through
547 Amazon Web Services Open Datasets Program https://registry.opendata.aws/australasian-
548 genomics/, NCBI under BioProject accession PRJNA763700 and GigaDB for woylie and PRJNA786364
549 and GigaDB [121] for numbat. The published koala genome assembly and annotation
550 (phaCin_unsw_v4.1_HiC.fasta) are available from the DNAzoo website
551 https://www.dnazoo.org/assemblies/Phascolarctos_cinereus. The published wombat genome
552 assembly and annotation (vu-2k.fasta) are also available from the DNAzoo website
553 https://www.dnazoo.org/assemblies/Vombatus_ursinus. The published antechinus genome
554 assembly and annotation (anrechinusM_pseudohap2.1.fasta) are available from NCBI under
555 BioProject accession PRJNA664282 and GigaDB [122], and published platypus genome assembly and
556 annotation (mOrnAna1.pri.v4) under BioProject accession PRJNA489114. Genomic coordinates for all
557 immune gene sequences annotated in this study are available in Additional file 1. Supporting
558 information for this study is available in Additional file 2.

559 # Declarations

560 ## List of abbreviations

561 Bacterial artificial chromosome (BAC), basic local alignment search tool (BLAST), benchmarking single
562 copy gene orthologs (BUSCO), complementary DNA (cDNA), devil facial tumour disease (DFTD), giga-
563 base-pair (Gpb), high-throughput chromosome conformation capture (HiC), hidden markov model
564 (HMM), immunoglobulin (IG), immunoglobulin superfamily (IGSF), interferon (IFN), international
565 immunogenetic information system (IMGT), kilo-base-pair (kbp), leukocyte receptor complex (LRC),
566 major histocompatibility complex (MHC), mega-base-pair (Mbp), National Center for Biotechnology
567 Information (NCBI), natural killer complex (NKC), natural killer receptor (NK), Pacific Biosciences
568 (PacBio), protein families database (Pfam), recombination signal sequence (RSS), simple modular

569 architecture research tool (SMART), single nucleotide polymorphisms (SNPs), T cell receptor (TCR)
570 and toll-like receptor (TLR).

571 Consent for publication
572 Not applicable

573 Competing interests
574 The authors declare that they have no competing interests

575 Funding
576 This work has been funded by the Australian Research Council Centre of Excellence for Innovations in
577 Peptide and Protein Science (CE200100012) and Discovery Project (DP180102465). LS was supported
578 by LP180100244 and PB was supported by an Australian Postgraduate award. YZ is supported by the
579 China Scholarship Council.

580 Authors' contributions
581 LS assembled and annotated the woylie genome and transcriptomes, PB assembled and annotated
582 the numbat genome and transcriptomes, EP assisted with both. EP, PB, LS, YC and YZ annotated
583 immune genes. KB, CJH and EP designed the study. EP drafted the manuscript, all authors read and
584 commented on drafts of the manuscript and have approved the submission.

585 Acknowledgements
586 The authors would also like to acknowledge the generous contribution of the Presbyterian Ladies'
587 College Sydney and Bioplatforms Australia.

588 References

589 1. Diaz S, Settle J, Brondizio ES, Ngo HT, Gueze M, Agard J, et al. *IPBES (2019): Summary for*
590 *policymakers of the global assessment report on biodiversity and ecosystem servies of the*
591 *Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. 2019.
592 Bonn, Germany.
593 2. Scheele BC, Pasmans F, Skerratt LF, Berger L, Martel A, Beukema W, et al. Amphibian fungal
594 panzootic causes catastrophic and ongoing loss of biodiversity. Science. 2019;363
595 6434:1459. doi:10.1126/science.aav0379.
596 3. Hoyt JR, Kilpatrick AM and Langwig KE. Ecology and impacts of white-nose syndrome on
597 bats. Nature Reviews Microbiology. 2021;19 3:196-210. doi:10.1038/s41579-020-00493-5.
598 4. Woods GM, Lyons AB and Bettiol SS. A Devil of a Transmissible Cancer. Tropical Medicine
599 and Infectious Disease. 2020;5 2:50.

5. Rohr JR, Civitello DJ, Halliday FW, Hudson PJ, Lafferty KD, Wood CL, et al. Towards common ground in the biodiversity–disease debate. Nature Ecology & Evolution. 2020;4 1:24-33. doi:10.1038/s41559-019-1060-6.
6. Hohenlohe PA, Funk WC and Rajora OP. Population genomics for wildlife conservation and management. Molecular Ecology. 2021;30 1:62-82. doi:https://doi.org/10.1111/mec.15720.
7. Lewin H, Robinson G, Kress WJ, Baker W, Coddington J, Crandall K, et al. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences (PNAS). 2018;115 17:4325-33. doi:10.1073/pnas.1720115115.
8. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592 7856:737-46. doi:10.1038/s41586-021-03451-0.
9. Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP and Myers E. Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. Annual Review of Animal Biosciences. 2018;6 1:23-46. doi:10.1146/annurev-animal-022516-022811.
10. Zhang G. Bird sequencing project takes off. Nature. 2015;522 7554:34-. doi:10.1038/522034d.
11. Peterson BK, Weber JN, Kay EH, Fisher HS and Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLOS ONE. 2012;7 5:e37135. doi:10.1371/journal.pone.0037135.
12. Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, et al. A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. Plant Methods. 2010;6 1:16. doi:10.1186/1746-4811-6-16.
13. McLennan EA, Grueber CE, Wise P, Belov K and Hogg CJ. Mixing genetically differentiated populations successfully boosts diversity of an endangered carnivore. Animal Conservation. 2020;23 6:700-12. doi:https://doi.org/10.1111/acv.12589.
14. Scally A, Yngvadottir B, Xue Y, Ayub Q, Durbin R and Tyler-Smith C. A Genome-Wide Survey of Genetic Variation in Gorillas Using Reduced Representation Sequencing. PLOS ONE. 2013;8 6:e65066. doi:10.1371/journal.pone.0065066.
15. Robledo-Ruiz DA, Pavlova A, Clarke RH, Magrath MJL, Quin B, Harrisson KA, et al. A novel framework for evaluating in situ breeding management strategies in endangered populations. Molecular Ecology Resources. 2021;n/a n/a doi:https://doi.org/10.1111/1755-0998.13476.
16. Lott MJ, Wright BR, Kemp LF, Johnson RN and Hogg CJ. Genetic Management of Captive and Reintroduced Bilby Populations. The Journal of Wildlife Management. 2020;84 1:20-32. doi:https://doi.org/10.1002/jwmg.21777.
17. Irving AT, Ahn M, Goh G, Anderson DE and Wang L-F. Lessons from the host defences of bats, a unique viral reservoir. Nature. 2021;589 7842:363-70. doi:10.1038/s41586-020-03128-0.
18. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map of the extended human MHC. Nature Reviews Genetics. 2004;5 12:889-99. doi:10.1038/nrg1489.
19. Robinson J, Waller MJ, Parham P, Groot Nd, Bontrop R, Kennedy LJ, et al. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Research. 2003;31 1:311-4. doi:10.1093/nar/gkg070.
20. Trowsdale J and Parham P. Mini-review: Defense strategies and immunity-related genes. European Journal of Immunology. 2004;34 1:7-17. doi:https://doi.org/10.1002/eji.200324693.
21. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology. 2018;36 4:338-45. doi:10.1038/nbt.4060.

22. Ming L, Wang Z, Yi L, Batmunkh M, Liu T, Siren D, et al. Chromosome-level assembly of wild Bactrian camel genome reveals organization of immune gene loci. Molecular Ecology Resources. 2020;20 3:770-80. doi:https://doi.org/10.1111/1755-0998.13141.
23. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nature Genetics. 2017;49 4:643-50. doi:10.1038/ng.3802.
24. Holt C and Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12 1:491. doi:10.1186/1471-2105-12-491.
25. Solovyev V, Kosarev P, Seledsov I and Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biology. 2006;7 Suppl 1:S10. doi:10.1186/gb-2006-7-s1-s10.
26. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Cold Spring Harbor Laboratory, 2020.
27. Seppey M, Manni M and Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer New York; 2019. p. 227-45.
28. Mudge JM and Harrow J. The state of play in higher eukaryote gene annotation. Nature Reviews Genetics. 2016;17 12:758-72. doi:10.1038/nrg.2016.119.
29. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, et al. EGASP: the human ENCODE genome annotation assessment project. Genome Biology. 2006;7 Suppl 1:S2. doi:10.1186/gb-2006-7-s1-s2.
30. Ensembl: Annotation of immunoglobulin and T cell receptor genes. https://m.ensembl.org/info/genome/genebuild/ig_tcr.html (2021). Accessed 19th July 2021.
31. Hogg CJ, Ottewell K, Latch P, Rossetto M, Biggs J, Gilbert A, et al. Threatened Species Initiative: Empowering conservation action using genomic resources. Proceedings of the National Academy of Sciences. 2022;119 4:e2115643118. doi:10.1073/pnas.2115643118.
32. Gordon D, Huddleston J, Chaisson Mark JP, Hill Christopher M, Kronenberg Zev N, Munson Katherine M, et al. Long-read sequence assembly of the gorilla genome. Science. 2016;352 6281:aae0344. doi:10.1126/science.aae0344.
33. Zhou Y, Shearwin-Whyatt L, Li J, Song Z, Hayakawa T, Stevens D, et al. Platypus and echidna genomes reveal mammalian biology and evolution. Nature. 2021; doi:10.1038/s41586-020-03039-0.
34. Quigley BL and Timms P. The Koala Immune Response to Chlamydial Infection and Vaccine Development—Advancing Our Immunological Understanding. Animals. 2021;11 2:380.
35. Madden D, Whaite A, Jones E, Belov K, Timms P and Polkinghorne A. Koala immunology and infectious diseases: How much can the koala bear? Developmental & Comparative Immunology. 2018;82:177-85. doi:https://doi.org/10.1016/j.dci.2018.01.017.
36. Peel E and Belov K. Lessons learnt from the Tasmanian devil facial tumour regarding immune function in cancer. Mammalian Genome. 2018; doi:10.1007/s00335-018-9782-3.
37. Murchison EP, Schulz-Trieglaff OB, Ning Z, Alexandrow LB, Bauer MJ, Fu B, et al. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. Cell. 2012;148:780-91.
38. Morris B, Cheng Y, Warren W, Papenfuss AT and Belov K. Identification and analysis of divergent immune gene families within the Tasmanian devil genome. BMC Genomics. 2015;16 1.
39. Murchison EP, Tovar C, Hsu AL, Bender HS, Kheradpour P, Rebbeck CA, et al. The Tasmanian devil transcriptome reveals schwann cell origins in a clonally transmissible cancer. Science. 2010;327:84-7.

699 40. Hewavisenti RV, Morris KM, O'Meally D, Cheng Y, Papenfuss AT and Belov K. The
700 identification of immune genes in the milk transcriptome of the Tasmanian devil (*Sarcophilus*
701 *harrisii*). PeerJ. 2016;4:1569.
702 41. Cheng Y, Stuart A, Morris K, Taylor R, Siddle HV, Deakin JE, et al. Antigen-presenting genes
703 and genomic copy number variations in the Tasmanian devil MHC. BMC Genomics.
704 2012;13:87.
705 42. Morris KM, Wright B, Grueber CE, Hogg C and Belov K. Lack of genetic diversity across
706 diverse immune genes in an endangered mammal, the Tasmanian devil (*Sarcophilus harrisii*).
707 Molecular Ecology. 2015;24:3860-72.
708 43. Siddle HV, Kreiss A, Eldridge MDB, Noonan E, Clarke CJ, Pyecroft S, et al. Transmission of a
709 fatal clonal tumor by biting occurs due to depleted MHC diversity in a threatened
710 carnivorous marsupial. PNAS. 2007;104 41:16221-6.
711 44. Siddle HV, Kreiss A, Tovar C, Yuen CK, Chen Y, Belov K, et al. Reversible epigenetic down-
712 regulation of MHC molecules by devil facial tumour disease illustrates immune escape by a
713 contagious cancer. PNAS. 2013;110 13:5103-8.
714 45. Siddle HV, Sanderson CE and Belov K. Characterization of major histocompatability complex
715 class I and II genes from the Tasmanian devil (*Sarcophilus harrisii*). Immunogenetics.
716 2007;59:753-60.
717 46. Cheng Y, Sanderson CE, Jones M and Belov K. Low MHC class II diversity in the Tasmanian
718 devil. Immunogenetics. 2012;64:525-33.
719 47. Cheng Y, Grueber C, Hogg CJ and Belov K. Improved high-throughput MHC typing for non-
720 model species using long-read sequencing. Molecular Ecology Resources. 2021;n/a n/a
721 doi:https://doi.org/10.1111/1755-0998.13511.
722 48. Quigley BL and Timms P. Helping koalas battle disease – Recent advances in Chlamydia and
723 koala retrovirus (KoRV) disease understanding and treatment in koalas. FEMS Microbiology
724 Reviews. 2020; doi:10.1093/femsre/fuaa024.
725 49. Johnson RN, O'Meally D, Chen Z, Etherington GJ, Ho SYW, Nash WJ, et al. Adaptation and
726 conservation insights from the koala genome. Nature Genetics. 2018;50 8:1102-11.
727 doi:10.1038/s41588-018-0153-5.
728 50. Morris K, Prentis PJ, O'Meally D, Pavasovic A, Brown AT, Timms P, et al. The koala
729 immunological toolkit: sequence identification and comparison of key markers of the koala
730 (*Phascolarctos cinereus*) immune response. Australian Journal of Zoology. 2014;62:195-9.
731 51. Morris KM, Matthew M, Waugh C, Ujvari B, Timms P, Polkinghorne A, et al. Identification,
732 characterisation and expression analysis of natural killer receptor genes in *Chlamydia*
733 *pecorum* infected koalas (*Phascolarctos cinereus*). BMC Genomics. 2015;16 1.
734 52. Morris KM, O'Meally D, Zaw T, Song X, Gillett A, Molloy MP, et al. Characterisation of the
735 immune compounds in koala milk using a combined transcriptomic and proteomic approach.
736 Scientific Reports. 2016;6:e35011.
737 53. Lau Q, Griffith JE and Higgins DP. Identification of MHCII variants associated with chlamydial
738 disease in the koala (*Phascolarctos cinereus*). PeerJ. 2014;2:443.
739 54. Robbins A, Hanger J, Jelocnik M, Quigley BL and Timms P. Koala immunogenetics and
740 chlamydial strain type are more directly involved in chlamydial disease progression in koalas
741 from two south east Queensland koala populations than koala retrovirus subtypes. Scientific
742 Reports. 2020;10 1:15013. doi:10.1038/s41598-020-72050-2.
743 55. Owen RS and Siddle HV. Devil Facial Tumours: Towards a Vaccine. Immunological
744 Investigations. 2019;48 7:719-36. doi:10.1080/08820139.2019.1624770.
745 56. Phillips S, Quigley BL and Timms P. Seventy years of *Chlamydia* vaccine research - limitations
746 of the past and directions for the future. Frontiers in Microbiology. 2019;10 70.
747 57. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, et al. The Juicebox
748 Assembly Tools module facilitates *de novo* assembly of mammalian genomes with
749 chromosome-length scaffolds for under $1000. bioRxiv. 2018:254797. doi:10.1101/254797.

750 58. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo
751 assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds.
752 Science advances. 2017;356:92-5.
753 59. Peel E, Silver L, Brandies P, Hogg CJ and Belov K. A reference genome for the critically
754 endangered woylie, *Bettongia penicillata ogilbyi*. Gigabyte. 2021;1
755 doi:https://doi.org/10.46471/gigabyte.35.
756 60. Brandies PA, Tang S, Johnson RSP, Hogg CJ and Belov K. The first *Antechinus* reference
757 genome provides a resource for investigating the genetic basis of semelparity and age-
758 related neuropathologies. Gigabyte. 2020;2020:0. doi:10.46471/gigabyte.7.
759 61. Peel E, Silver L, Brandies PA, Hayakawa T, Belov K and Hogg CJ. Genome assembly of the
760 numbat (*Myrmecobius fasciatus*), the only termitivorous marsupial. Gigabyte. 2022;
761 doi:https://doi.org/10.46471/gigabyte.47.
762 62. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
763 architecture and applications. BMC bioinformatics. 2009;10 421.
764 63. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical
765 significance estimation. PLOS Computational Biology. 2008;4 5:e1000069.
766 64. Belov K, Lam MKP, Hellman L and Colgan DJ. Evolution of the major histocompataibility
767 complex: isolation of the class II beta cDNAs from two monotremes, the platypus and the
768 short-beaked echidna. Immunogenetics. 2003;55:402-11.
769 65. Belov K and Hellman L. Immunoglobulin genetics of *Ornithorhynchus anatinus* (platypus) and
770 *Tachyglossus aculeatus* (short-beaked echidna). Comparative Biochemistry and Physiology.
771 2003;136:811-9.
772 66. Wong ESW, Sanderson CE, Deakin JE, Whittington CM, Papenfuss AT and Belov K.
773 Identification of natural killer cell receptor clusters in the platypus genome reveals an
774 expansion of C-type lectin genes. Immunogenetics. 2009;61:565-79.
775 67. Dohm JC, Tsend-Ayush E, Reinhardt R, Grutzner F and Himmelbauer H. Distribution and
776 pseudoautosomal localization of the major histocompatability complex in monotremes.
777 Genome Biology. 2007;8:R175-R.16.
778 68. Papenfuss AT, Feng Z, Krasnec K, Deakin JE, Baker ML and Miller RD. Marsupials and
779 monotremes possess a novel family of MHC class I genes that is lost from the eutherian
780 lineage. BMC Genomics. 2015;16:535.
781 69. Nowak MA, Parra ZE, Hellman L and Miller RD. The complexity of expressed kappa light
782 chains in egg-laying mammals. Immunogenetics. 2004;56:555-63.
783 70. Johansson J, Aveskogh M, Munday BL and Hellman L. Heavy chain V region diversity in the
784 duck-billed platypus (*Ornithorhynchus anatinus*): long and highly variable complementary-
785 determining region 3 compensates for limited germline diversity. The Journal of
786 Immunology. 2002;168:5155-62.
787 71. Johansson J, Salazar JN, Aveskogh M, Munday B, Miller RD and Hellman L. High variability in
788 complementarity-determining regions compensates for a low number of V gene families in
789 the λ light chain locus of the platypus. European Journal of Immunology. 2005;35 10:3008-
790 19. doi:https://doi.org/10.1002/eji.200425574.
791 72. Parra ZE, Arnold T, Nowak MA, Hellman L and Miller RD. TCR gamma chain diversity in the
792 splen of the duckbill platypus (*Ornithorhynchus anatinus*). Developmenal and Comparative
793 Immunology. 2006;30:699-71.
794 73. Parra ZE, Lillie M and Miller RD. A model for the evolution of the mammalian T-cell receptor
795 alpha/delta and mu loci based on evidence from the duckbill platypus. Molecular Biology
796 and Evolution. 2012;29 10:3205-14.
797 74. Wang X, Parra ZE and Miller RD. Platypus TCRmu provides insight into the origins and
798 evolution of a uniquely mammalian TCR locus. The Journal of Immunology. 2011;187:5246-
799 54.

75. Wong ESW, Papenfuss AT, Miller RD and Belov K. Hatching time for monotreme immunology. Australian Journal of Zoology. 2006;57:185-98.
76. Belov K, Sanderson CE, Deakin JE, Wong ESW, Assange D, McColl KA, et al. Characterization of the opossum immune genome provides insight into the evolution of the mammalian immune system. Genome Research. 2007;17:982-91.
77. Lau Q, Jobbins SE, Belov K and Higgins DP. Characterisation of four major histocompatibility complex class II genes of the koala (*Phascolarctos cinereus*). Immunogenetics. 2013;65:37-46.
78. Jobbins SE, Sanderson CE, Griffith JE, Krockenberger MB, Belov K and Higgins DP. Diversity of MHC class II *DAB1* in the koala (*Phascolarctos cinereus*). Australian Journal of Zoology. 2012;60 1:1-9. doi:https://doi.org/10.1071/ZO12013.
79. Cheng Y, Polkinghorne A, Gillett A, Jones EA, O'Meally D, Timms P, et al. Characterisation of MHC class I genes in the koala. Immunogenetics. 2018;70:125-33.
80. Matthew M, Beagley KW, Timms P and Polkinghorne A. preliminary characterisation of tumour necrosis factor alpha and interleukin-10 responses to *Chlamydia pecorum* infection in the koala (*Phascolarctos cinereus*). PLOS one. 2013;8 3:e59958.
81. Matthew M, Pavasovic A, Prentis PJ, Beagley KW, Timms P and Polkinghorne A. Molecular characterisation and expression analysis of Interferon gamma in response to natural *Chlamydia* infection in the koala, *Phascolarctos cinereus*. Gene. 2013;527:570-7.
82. Mathew M, Waugh C, Beagley KW, Timms P and Polkinghorne A. Interleukin 17A is an immune marker for chlamydial disease severity and pathogenesis in the koala (Phascolarctos cinereus). Developmental & Comparative Immunology. 2014;46 2:423-9. doi:http://dx.doi.org/10.1016/j.dci.2014.05.015.
83. Maher IE, Griffith JE, Lau Q, Reeves T and Higgins DP. Expression profiles of the immune genes CD4, CD8 beta, IFN gamma, IL-4, IL-6 and IL-10 in mitogen-stimulated koala lymphocytes (*Phascolarctos cinereus*) by qRT-PCR. PeerJ. 2014;2.
84. Miller RD. Those other mammals: The immunoglobulins and T cell receptors of marsupials and monotremes. Seminars in Immunology. 2010;22:3-9.
85. Parra ZE, Baker ML, Hathaway J, Lopez AM, Trujillo J, Sharp A, et al. Comparative genomic analysis and evolution of the T cell receptor loci in the opossum *Monodelphis domestica*. BMC Genomics. 2008;9 111:1-19.
86. Siddle HV, Deakin JE, Coggill P, Whilming LG, Harrow J, Kaufman J, et al. The tammar wallaby major histocompatability complex shows evidence of past genomic instability. BMC Genomics. 2011;12:421.
87. Cheng Y, Siddle HV, Beck S, Eldridge MDB and Belov K. High levels of genetic variation at MHC class II DBB loci in the tammar wallaby (*Macropus eugenii*). Immunogenetics. 2009;61:111-8.
88. Zuccolotto P, Harrison GA and Deane EM. Cloning of marsupial T cell receptor alpha and beta constant region cDNAs. Immunology and Cell Biology. 2000;78 2:103-9.
89. Daly KA, Digby M, Lefevre C, Mailer S, Thomson P, Nicholas KR, et al. Analysis of the expression of immunoglobulins throughout lactation suggests two periods of immune transfer in the tammar wallaby (*Macropus eugenii*). Veterinary Immunology and Immunopathology. 2007;120:187-200.
90. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 2018;47 D1:D766-D73. doi:10.1093/nar/gky955.
91. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang Y-C, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biology. 2018;19 1:208. doi:10.1186/s13059-018-1590-2.

92. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Research. 2015;43 D1:D413-D22. doi:10.1093/nar/gku1056.
93. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biology. 2019;20 1:92. doi:10.1186/s13059-019-1715-2.
94. Ensembl: Manual gene annotation by Havana. https://m.ensembl.org/info/genome/genebuild/manual_havana.html (2021). Accessed 19th July 2021.
95. National Center for Biotechnology Information: The NCBI Eukaryotic Genome Annotation Pipeline. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/ (2021). Accessed 19th July 2021.
96. Ensembl: Gene annotation in Ensembl. https://m.ensembl.org/info/genome/genebuild/index.html (2021). Accessed 19th July 2021.
97. National Center for Biotechnology Information: Updating information on GenBank genome records. https://www.ncbi.nlm.nih.gov/genbank/wgs_update/ (2021). Accessed 4th August 2021.
98. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct determination of diploid genome sequences. Genome Research. 2017;27 5:757-67. doi:10.1101/gr.214874.116.
99. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326 5950:289. doi:10.1126/science.1181369.
100. Luo J, Wei Y, Lyu M, Wu Z, Liu X, Luo H, et al. A comprehensive review of scaffolding methods in genome assembly. Briefings in Bioinformatics. 2021; doi:10.1093/bib/bbab033.
101. Durand NC, Robinson JT, Shamim MD, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualisation system for Hi-C contact maps with unlimited zoom. Cell Systems. 2016;3:99-101.
102. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLOS Computational Biology. 2019;15 8:e1007273. doi:10.1371/journal.pcbi.1007273.
103. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO and Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology. 2013;31 12:1119-25. doi:10.1038/nbt.2727.
104. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, et al. Resolving the full spectrum of human genome variation using Linked-Reads. Genome Research. 2019;29 4:635-45. doi:10.1101/gr.234443.118.
105. 10x Genomics: Molecule length calculation. https://support.10xgenomics.com/de-novo-assembly/software/pipelines/latest/output/moleculelen (2020). Accessed 22nd July 2021.
106. 10x Genomics: Supernova version 2.x performance. https://support.10xgenomics.com/de-novo-assembly/software/overview/latest/performance (2020). Accessed 22nd July 2021.
107. Ho SS, Urban AE and Mills RE. Structural variation in the sequencing era. Nature Reviews Genetics. 2020;21 3:171-89. doi:10.1038/s41576-019-0180-9.
108. Ott A, Schnable JC, Cheng-Ting Y, Wu L, Liu C, Heng-Cheng H, et al. Linked read technology for assembling large complex and polyploid genomes. BMC Genomics. 2018;19 doi:http://dx.doi.org/10.1186/s12864-018-5040-z.
109. Damas J, Corbo M and Lewin HA. Vertebrate Chromosome Evolution. Annual Review of Animal Biosciences. 2021;9 1:1-27. doi:10.1146/annurev-animal-020518-114924.
110. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. Nature. 2020;587 7833:240-5. doi:10.1038/s41586-020-2876-6.
111. Ceballos FC, Joshi PK, Clark DW, Ramsay M and Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nature Reviews Genetics. 2018;19 4:220-34. doi:10.1038/nrg.2017.109.

112. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. Journal of Heredity. 2009;100 6:659-74.
113. Kingan SB, Heaton H, Cudini J, Lambert CC, Baybayan P, Galvin BD, et al. A High-Quality De novo Genome Assembly from a Single Mosquito Using PacBio Sequencing. Genes. 2019;10 1:62. doi:10.3390/genes10010062.
114. Lawniczak M, Blaxter M, Johnson WE, Pettersson OV, Barker K and The Sample Collection and Processing Subcommittee. *Report on sample collection and processing standards*. March 2021 2021. Earth Biogenomne Project,.
115. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research. 2018;46 W1:W537-W44. doi:10.1093/nar/gky379.
116. Mistry J, Finn RD, Eddy SR, Bateman A and Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Research. 2013;41 12:e121-e. doi:10.1093/nar/gkt263.
117. Hall T. BioEdit v7.2.2 ed.: Ibis Biosciences, 2013.
118. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar Gustavo A, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Research. 2020;49 D1:D412-D9. doi:10.1093/nar/gkaa913.
119. Letunic I, Khedkar S and Bork P. SMART: recent updates, new developments and status in 2020. Nucleic Acids Research. 2020;49 D1:D458-D60. doi:10.1093/nar/gkaa937.
120. Kumar S, Stecher G, Li M, Knyaz C and Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular biology and evolution. 2018;35 6:1547-9. doi:10.1093/molbev/msy096.
121. Peel E, Silver L, Brandies PA, Hayakawa T, Belov K and Hogg CJ. Supporting data for "Genome assembly of the numbat (*Myrmecobius fasciatus*), the only termitivorous marsupial. GigaDB. 2022; doi:http://dx.doi.org/10.5524/100999.
122. Peel E, Silver L, Brandies PA, Hogg CJ and Belov K. Supporting data for "A reference genome for the critically endanfered woylie, *Bettongia penicillata ogilbyi*". GigaDB. 2021; doi:http://dx.doi.org/10.5524/100951.

Figure1

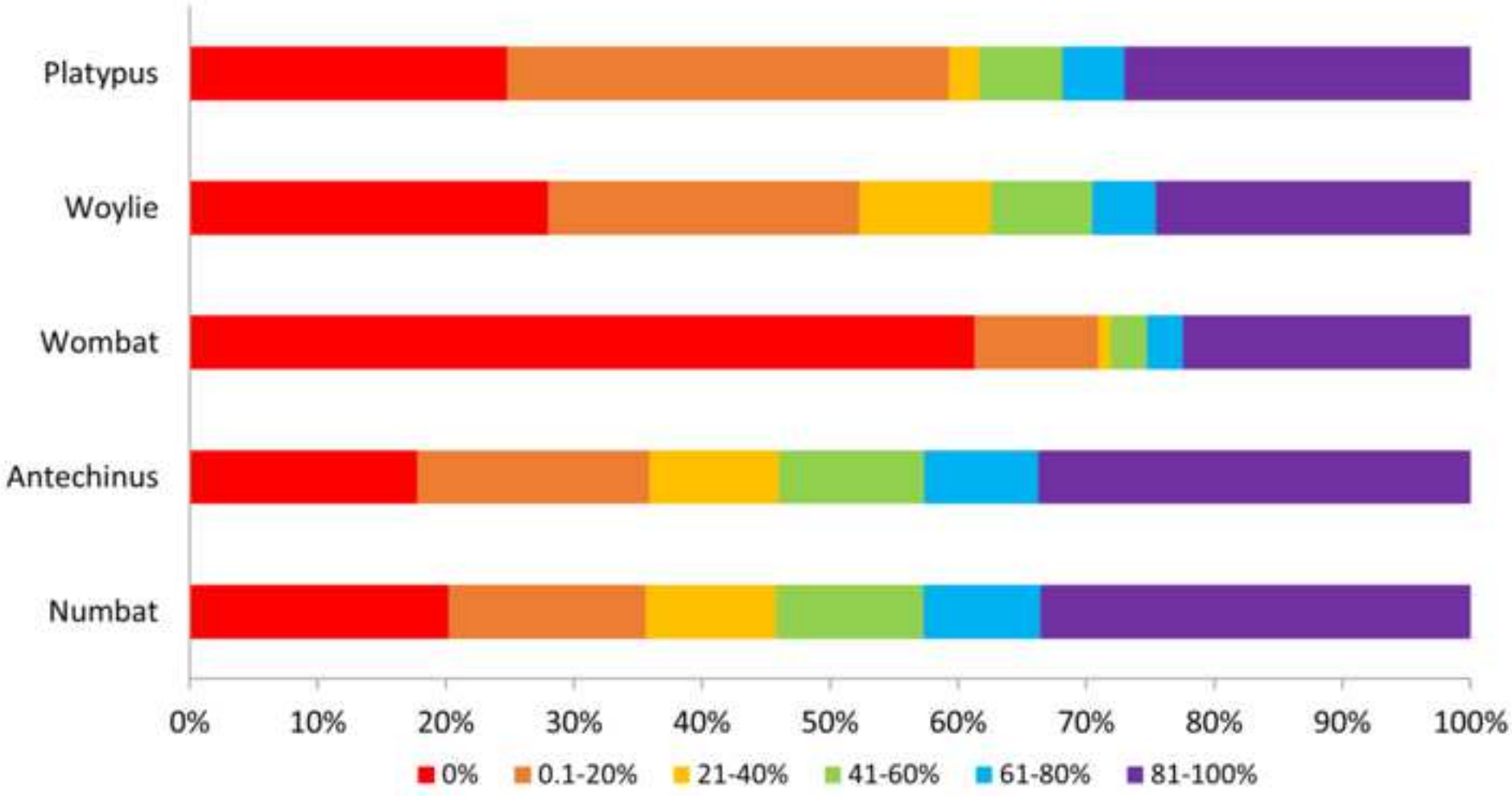


Figure2

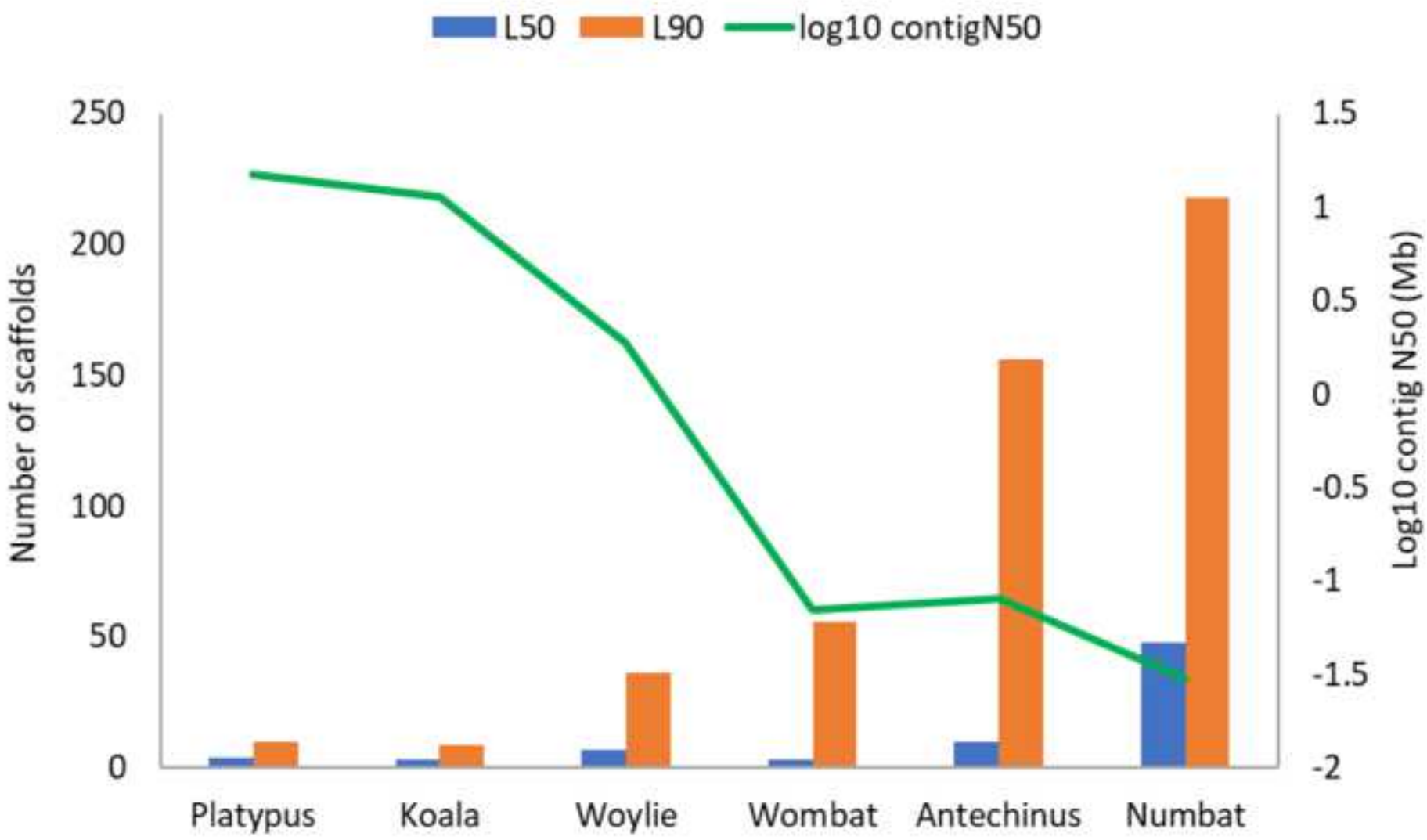


Figure3

Click here to access/download;Figure;Peel_Fig3.tiff

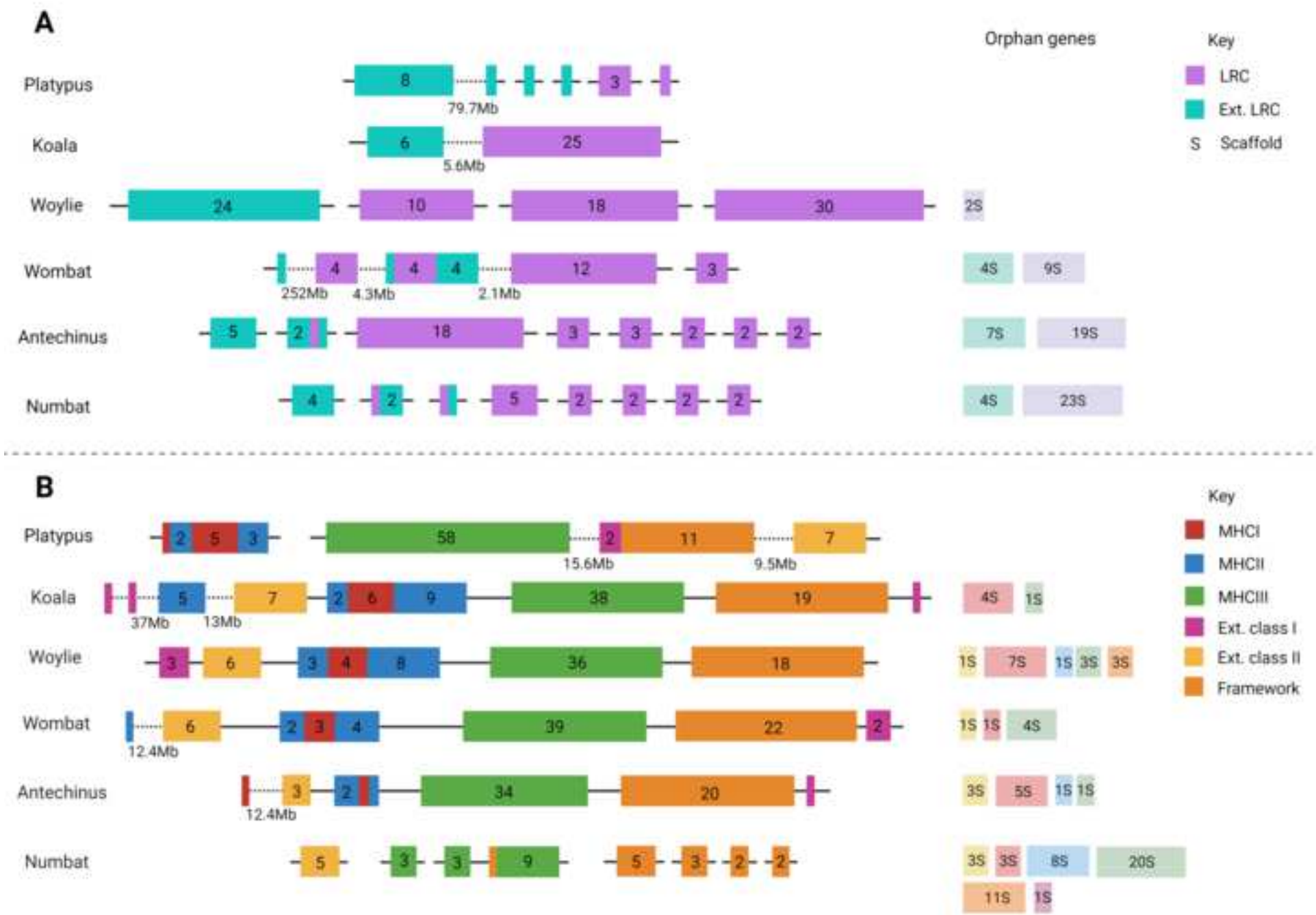
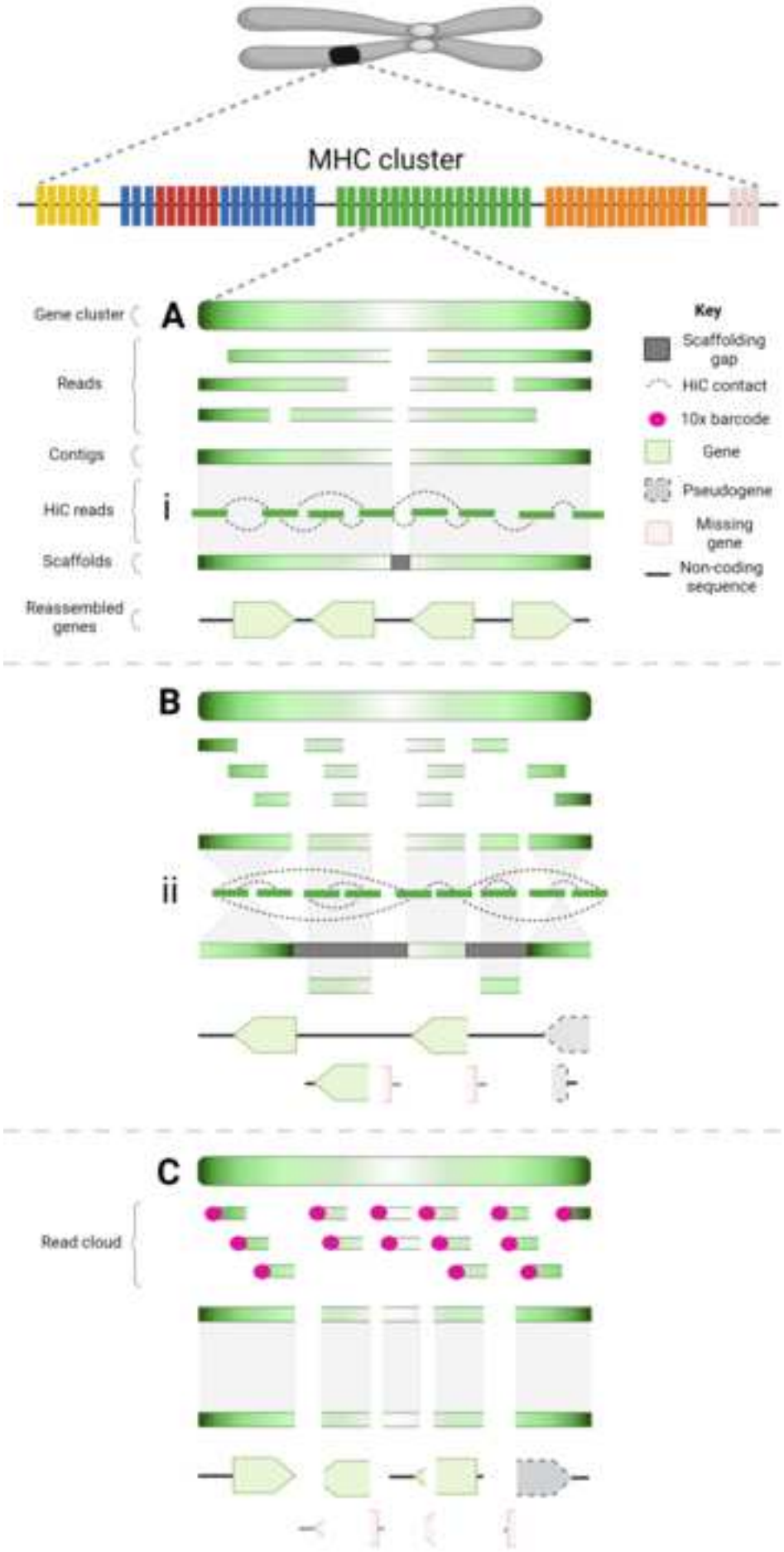


Figure4

Click here to access/download;Figure;Peel_Fig4.tiff



THE UNIVERSITY OF SYDNEY

Professor Katherine Belov AO *BSc (Hons) PhD*
Pro Vice-Chancellor Global Engagement

24th March 2022

Dr Scott Edmunds
Chief Editor
GigaScience

Dear Dr Edmunds,

Please find attached our manuscript "*Best genome sequencing strategies for annotation of complex immune gene families in wildlife*" which we are submitting as a research article for publication in GigaScience. The text of the manuscript totals 7139 words, with four figures, two tables and two additional files.

Globally we are in the midst of a biodiversity crisis and infectious diseases are a major driver of wildlife decline. The COVID-19 pandemic highlights the impact of wildlife disease on animal and human health, and provides impetus for studying immune genes in wildlife. Despite the recent increase in genomes for wildlife species, our understanding of immune genes in these species is limited owing to their high level of polymorphism and complex genomic organisation which makes assembly and annotation notoriously difficult.

Our research over the past decade and a half on Tasmanian devils and koalas highlights the importance of genomics and accurate immune annotation for wildlife disease investigations. As such, we are increasingly asked the minimum genome quality required to effectively annotate immune genes which underpin wildlife disease investigations. In this manuscript we aim to answer this question by manually annotating immune genes in five marsupial genomes and one monotreme genome of different qualities to determine the impact of sequencing strategy and automated annotation on accurate immune annotation.

We determined that high-quality chromosome-length genome assemblies generated using long-reads and scaffolding technologies are required to accurately annotate immune genes. Draft-quality genomes generated using short-reads and HiC technology, or 10x Chromium linked-read technology, resulted in highly fragmented immune genes which led to incorrect annotation and prevented interpretation of genomic organisation and gene family evolution.

While the six genomes in this study displayed BUSCO scores of up to 94.1% indicating functional completeness, we also show that automated annotation programs need improvement, as up to 60% of manually annotated immune genes were not accurately annotated by automated programs. Deficient annotations within functionally important immune gene families will flow through to downstream analysis and result in spurious results.

Office of the Pro Vice-Chancellor
Global Engagement
F23 Administration Building, Level 3
Corner of Eastern Avenue and City Road
The University of Sydney NSW 2006 Australia

T +61 2 9351 4461
E pvc.ge@sydney.edu.au

ABN 15 211 513 464
CRICOS 00026A

We have targeted GigaScience as this work provides a strong example for the importance of gold standard genome assemblies for studying wildlife disease and will appeal to researchers involved in sequencing, assembly, annotation and translation of genomics data. The content of this manuscript has not been published or submitted for publication elsewhere. The authors declare no competing interests, and all have approved the manuscript for submission. We hope you will agree that this work represents an important contribution to GigaScience.

Yours sincerely,

Professor Kathy Belov
Corresponding Author
On Behalf of co-authors Emma Peel, Luke Silver, Parice Brandies, Ying Zhu, Yuanyuan Cheng and Carolyn Hogg