

GigaScience

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

Manuscript Number:	GIGA-D-22-00064R2	
Full Title:	Best genome sequencing strategies for annotation of complex immune gene families in wildlife	
Article Type:	Research	
Funding Information:	Australian Research Council (CE200100012)	Prof Katherine Belov
	Australian Research Council (DP180102465)	Prof Katherine Belov
Abstract:	<p>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.</p> <p>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 59% of manually annotated immune genes regardless of assembly quality or method of automated annotation.</p> <p>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.</p> <p>Keywords: immune gene, genome, quality, annotation, MHC, wildlife, disease</p>	
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:	
Response to Reviewers:	Amendments made as requested to the data availability section
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<p>Experimental design and statistics</p> <p>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.</p> <p>Have you included all the information requested in your manuscript?</p>	No
<p>If not, please give reasons for any omissions below.</p> <p>as follow-up to "Experimental design and statistics</p> <p>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.</p> <p>Have you included all the information requested in your manuscript?</p> <p>"</p>	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	Yes

<p>A description of all resources used, including antibodies, cell lines, animals 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.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
<p>Availability of data and materials</p> <p>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.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

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

3 Emma Peel^{1,2}, Luke Silver¹, Parice Brandies¹, Ying Zhu³, Yuanyuan Cheng¹, Carolyn J Hogg^{1,2} &
4 Katherine Belov^{1,2*}

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

7 ² Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science,
8 University of Sydney, NSW 2006, Australia

9 ³ Sichuan Provincial Academy of Natural Resource Sciences, Chengdu, Sichuan, China

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

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

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

13 so_zy2003@126.com

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

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

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

17

18

19

20

21

22

23

24

25

26 **Abstract**

27 **Background**

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

38 **Results**

39 Genome quality is directly linked to our ability to annotate complex immune gene families, with long
40 reads and scaffolding technologies required to reassemble immune gene clusters and elucidate
41 evolution, organisation and true gene content of the immune repertoire. Draft quality genomes
42 generated from short-reads with HiC or 10x Chromium linked-reads were unable to achieve this.
43 Despite mammalian BUSCOv5 scores of up to 94.1 % amongst the six genomes, automated annotation
44 pipelines incorrectly annotated up to 59% of manually annotated immune genes regardless of
45 assembly quality or method of automated annotation.

46 **Conclusions**

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

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

50 Background

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

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

68 The COVID-19 pandemic is one of many examples which highlight the ever-increasing importance of
69 understanding wildlife immunity and disease to better understand and manage disease spill over [17].
70 In the case of wildlife threatened by disease, conservation questions are more challenging to answer
71 and typically involve immunogenetic diversity which relies on accurate immune gene annotations.
72 Immune genes in mammals can be classified into six major families based on their evolutionary history
73 and function: T cell receptors (TCR), immunoglobulins (IG), major histocompatibility complex (MHC),
74 natural killer (NK) receptors, toll-like receptors (TLR) and cytokines. Mammals utilise two antigen

75 recognition systems: TCR and IG expressed by T lymphocytes and B lymphocytes respectively. TCR and
76 IG are encoded in large clusters within the genome, each of which contain few constant sequences
77 that define the receptor sub-type, and multiple highly duplicated variable segments that recognise
78 and bind antigens. The number and sequence polymorphism of IG and TCR V segments varies
79 significantly between mammalian species [18-20]. Another major family of immune genes is the major
80 histocompatibility complex which contains three classes of genes (class I, II and III). MHC class I and II
81 genes encode cell-surface receptors which bind and present self- and pathogen-derived antigens to T
82 lymphocytes, activating the adaptive immune response. Class I and II genes evolve via duplication and
83 can be highly polymorphic, hence gene number differs between species [21, 22]. Natural killer (NK)
84 cells directly kill virus-infected and cancerous cells and are an important component of innate
85 immunity. Their activity is mediated via cell-surface receptors encoded by genes classified into two
86 functionally similar but structurally dissimilar families; the leukocyte receptor complex (LRC) and
87 natural killer complex (NKC). These families are encoded in separate clusters within the genome, and
88 as they evolve via gene duplication, gene number varies significantly between species [23]. TLRs are
89 membrane-spanning receptors expressed by immune and non-immune cells which bind pathogen-
90 associated molecular patterns (PAMP), activating the innate and adaptive immune response.
91 Compared to other immune genes, TLRs gene number and sequence is relatively conserved across
92 mammals [24]. Lastly, cytokines are small proteins secreted by numerous cell types which direct the
93 immune response. Cytokines can be classified into multiple families including interferons (IFN),
94 tumour necrosis factors (TNF) and interleukins (IL), and gene content within each family varies
95 between mammals [25].

96 Immune genes are some of the most polymorphic regions of the genome, owing to the need to
97 generate diversity in response to ever-changing pathogenic pressures [26, 27]. Diversity within these
98 gene families is generated through gene duplication, gene copy number variation, SNPs and rapid
99 evolution, resulting in a complex genomic organisation and high level of pseudogenization [26].
100 Generally, immune genes are encoded within repetitive clusters in the genome, especially highly

101 duplicated families such as the MHC and NK receptors [28]. Given these factors, accurate assembly
102 and annotation of genomic regions encoding immune genes can be challenging [29-31], especially in
103 wildlife.

104 Automated annotation pipelines such as MAKER [32] and Fgenesh++ [33] are accurate at identifying
105 the majority of protein-coding genes within a genome [34, 35]. However, they are less effective at
106 characterising complex and highly variable gene families such as immune genes [36, 37] which are
107 misassembled even in the high-quality human genome [29]. As such, manual annotation and curation
108 of immune genes is required, which is conducted for model organism genomes accessioned with
109 Ensembl [38]. Wildlife are not currently included in this scope, and hence immune genes are poorly
110 annotated, or not annotated at all, in many species.

111 Advances in sequencing technology means chromosome-length genomes are now achievable for a
112 range of species [8]. Use of multiple sequencing, scaffolding, chromatin conformation and optical
113 mapping technologies leads to accurate assembly of complex and variable genomic regions, such as
114 immune genes [8]. However, the high input sample quantity and quality requirements are not always
115 feasible for wildlife [39]. This leads to the use of lower-input short-read sequencing to generate a
116 draft-quality genome assembled into scaffolds. However, short-read sequencing is well known to be
117 incompetent at resolving highly repetitive and complex gene regions [40, 41]. While scaffolding
118 technologies can improve contiguity of these assemblies, complex and variable regions often remain
119 fragmented. The need to balance budget, sample and genome assembly quality against accurate
120 immune gene annotation is essential to answer questions around disease and immunity.

121 Over the past decade and a half our research has focused on immunity and disease in two iconic
122 marsupial species; the Tasmanian devil and koala (*Phascolarctos cinereus*). During this period, we have
123 worked with bacterial artificial chromosome (BAC) and complementary DNA (cDNA) libraries and draft
124 genomes of varying qualities. Our research, and that of others, has been crucial for understanding,
125 managing and preventing disease-induced decline [4, 42-44]. As the cost of sequencing has dropped,

126 and the appreciation of the power of genetics and genomics for population management has
127 increased, we have increasingly been asked about the minimum levels of genome quality required to
128 be able to effectively annotate immune genes in order to study levels of diversity in wild populations.
129 Here we set out to answer that question.

130 Tasmanian devils are threatened by DFTD, a contagious cancer which has decimated over 80% of the
131 population since it was first documented in 1996 [4]. The Tasmanian devil reference genome was
132 sequenced using Illumina short-reads in 2012 [45], generating a 3.17 Gbp genome with a scaffold N50
133 of 1.8 Mbp and contig N50 of 20kbp. The Major Histocompatibility Complex was not able to be
134 annotated in the draft genome due to the high levels of fragmentation, scattered across at least 15
135 scaffolds. But manual annotation was possible alongside transcriptomes [46-48] and targeted
136 sequencing of MHC-positive BAC clones [46, 49-53]. Development of MHC markers led to
137 determination of gene copy number and nucleotide variation amongst the devil population, revealing
138 devils have low MHC diversity, much of which is shared with DFTD [51, 54]. The low histocompatibility
139 barriers, coupled with downregulation of tumour MHC expression, allows DFTD to transmit between
140 individuals and evade the host immune response [52]. Recent MHC genotyping using long-read
141 sequencing enabled the identification of full-phased MHC alleles and separation of highly similar
142 alleles (1bp difference), resulting in the identification of new functional MHC diversity within the devil
143 population [55].

144 The koala is another iconic Australian marsupial where disease is a major contributing factor to
145 population decline [56]. Chlamydiosis is one of many threatening processes affecting koalas, a disease
146 caused by infection with the intracellular bacterium *Chlamydia pecorum* [56]. A chromosome-length
147 koala reference genome was sequenced in 2018 using Pacific Biosciences (PacBio) long-reads, Illumina
148 short-reads and BioNano optical maps [57]. This generated a 3.19 Gbp assembly with a scaffold N50
149 of 480 Mbp and contig N50 of 11.4 Mbp [57], a 400-fold increase in scaffold contiguity compared to
150 the Tasmanian devil genome assembly [45]. This high-quality koala genome enabled accurate

151 annotation of immune gene families, including the first complete reconstruction of MHC and TCR gene
152 clusters from a genome sequence in marsupials [43, 58-60]. Preliminary genome resequencing
153 identified that variants within IFN γ , TNF α and MHC genes are essential for clearance of *Chlamydia* in
154 koalas [42]. MHC genotype has also been linked to disease susceptibility and severity in different koala
155 populations [61, 62].

156 In this study, our aim was to determine the impact of sequence data type, assembly quality and
157 automated annotation on accurate immune annotation. To achieve this, we manually annotated
158 immune genes in the genomes of five marsupials and one monotreme. These include recent published
159 genome assemblies of five marsupials; koala [57, 63, 64], woylie (*Bettongia penicillata*) [65], common
160 wombat (*Vombatus ursinus*) [63, 64], brown antechinus (*Antechinus stuartii*) [66] and numbat
161 (*Myrmecobius fasciatus*) [67], and previous immune gene annotations from one monotreme, the
162 platypus [41]. These six genomes differ in quality, from scaffold assemblies generated using only 10x
163 Chromium linked-reads (numbat, antechinus), short-read with high-throughput chromosome
164 conformation capture (HiC) (wombat), long and short-read (woylie), to high-quality chromosome-
165 length genomes generated using multiple data types (koala and platypus) (Table 1). We assess the
166 accuracy of automated immune gene annotation by Fgenesh++, MAKER and NCBI pipelines in these
167 non-model species. To account for the impact of species-specific gene expansion/contraction on
168 automated immune gene annotation, we also annotated two versions of the platypus genome from
169 2021 (GCA_004115215.4) and 2018 (GCA_002966995.1) with Fgenesh++. This study provides a guide
170 of the impact of genome quality on immune gene annotation. Here we show that high quality
171 chromosome-length genomes are necessary for accurate immune annotation in the context of wildlife
172 disease.

173 Analyses

174 Immune genes were annotated in the koala, woylie, wombat, antechinus, and numbat genomes and
175 transcriptomes using similarity-based search methods such as BLAST [68] and HMMER [69] with

176 known marsupial immune gene sequences as queries. This resulted in the manual characterisation of
 177 over 2,700 immune genes amongst the five species, from six immune gene families or groups: toll-like
 178 receptors (TLR), T cell receptors (TCR), immunoglobulins (IG), major histocompatibility complex
 179 (MHC), natural killer (NK) cell receptors and cytokines (Table 2). Platypus immune gene families have
 180 previously been annotated [41, 70-81], some of which had already been mapped within the 2021
 181 genome assembly (MHC and TCR) [41] and the remainder were mapped in both the 2018 and 2021
 182 assemblies in this study. Genomic coordinates of all immune genes annotated in this study are
 183 available in Additional file 1. A comprehensive summary of results for each immune gene family are
 184 available in Additional file 2.

185 Table 1. Assembly metrics and genome annotations for the five marsupial and two monotreme
 186 genome assemblies used in this study. The wombat and koala genome assemblies used in this study
 187 are not available on NCBI, hence the accession ID is not provided.

	Platypus		Koala	Woylie	Wombat	Antechinus	Numbat
Genome assembly version	GCA_0041152 15.4 [41] 2021	GCA_00296 6995.1 2018	phaCin_uns w_v4.1 [57, 63, 64]	GCA_023 548195.1 [65]	vu-2k [63, 64]	GCA_01669 6395.1 [66]	GCA_023 553655.1 [67]
Data types	PacBio 10x Chromium BioNano HiC (Phase genomics & Dovetail) RNAseq (19 transcrip- tomes)	PacBio Illumina RNAseq (19 transcrip- tomes)	PacBio RS II Illumina BioNano HiC (DNAzoo) RNAseq (16 transcrip- tomes)	PacBio HiFi Illumina RNAseq (4 transcrip- tomes)	Illumina HiC (DNAzoo)	10x Chromium RNAseq (12 transcrip- tomes)	10x Chromiu m RNAseq (3 transcrip- tomes)
Genome size (Gbp)	2.13	1.99	3.19	3.39	3.34	3.31	3.42
GC (%)	46.23	46.64	39.05	38.64	38.89	36.20	36.3
No. scaffolds	322	4,568	1,318	1,116	633,737	30,876	112,299
No. contigs	834	5,044	1,935	3,016	685,859	106,199	219,447
Scaffold N50 (Mbp)	83.33	18.71	480.11	6.94	576.1	72.7	0.223
Contig N50 (Mbp)	15.1	7.5	11.4	1.995	0.07	0.08	0.038
Gaps (%)	0.81	0.0002	0.01	0.403	0.54	2.75	3.52

Complete mammalian BUSCOv5.3.2	83.0%	81.5%	94.1%	94.1%	89.3%	92.5%	78.7%
Genome annotations used in this study	NCBI Fgenesh++ (this study)	Fgenesh++ (this study)	Fgenesh++ (this study)	Fgenesh+ + [65]	MAKER [63, 64] Fgenesh+ + (this study)	Fgenesh++ [66]	Fgenesh+ + [67]

188

189 Table 2. Number of annotated immune genes in each of the five marsupials and one monotreme in
190 this study. The percentage overlap of genomic coordinates between manual and automated
191 annotations of immune genes is also provided for each family and species.

	Platypus	Koala	Woylie	Wombat	Antechinus	Numbat
Cytokines	49 (48%, 8%)	82 (20%, 22%)	77 (19%, 38%)	76 (33%, 44%)	68 (17%, 21%)	67 (21%, 30%)
TLR	10 (90%, 10%)	10 (0%, 20%)	10 (6%, 37%)	10 (100%, 0%)	10 (10%, 20%)	10 (10%, 20%)
MHC I	6 (14%, 0%)	19 (21%, 21%)	17 (5%, 5%)	5 (60%, 0%)	7 (22%, 10%)	3 (22%, 11%)
MHC II	5 (25%, 25%)	16 (6%, 25%)	23 (12%, 16%)	7 (42%, 0%)	14 (33%, 6%)	6 (33%, 6%)
MHC III	58 (88%, 4%)	39 (11%, 7%)	37 (23%, 2%)	38 (65%, 7%)	36 (11%, 32%)	35 (12%, 12%)
Ext. MHC & framework genes	20 (100%, 0%)	27 (13%, 10%)	31 (32%, 8%)	34 (41%, 11%)	31 (21, 10%)	33 (11%, 42%)
NKC	122 (31%, 63%)	17 (27%, 11%)	17 (27%, 11%)	11 (9%, 36%)	11 (18%, 27%)	17 (33%, 5%)
LRC	4 (0%, 0%)	25 (3%, 18%)	60 (3%, 63%)	33 (28%, 54%)	49 (5%, 38%)	41 (5%, 38%)
Extended LRC	11 (36%, 0%)	6 (0%, 12%)	22 (0%, 60%)	9 (0%, 100%)	15 (37%, 18%)	11 (56%, 31%)
IG constant	14 (5%, 50%)	15 (0%, 66%)	20 (4%, 22%)	10 (16%, 66%)	7 (28%, 14%)	6 (0%, 33%)
IG variable	118 (0.5%, 80%)	289 (0%, 58%)	226 (0%, 58%)	98 (0.9%, 81%)	145 (0.6%, 43%)	121 (0%, 34%)
TCR constant	19 (0%, 88%)	10 (0%, 45%)	12 (0%, 29%)	10 (0%, 81%)	11 (0%, 36%)	9 (0%, 22%)
TCR variable	252 (0%, 78%)	103 (0%, 58%)	122 (0%, 76%)	92 (2%, 86%)	126 (0%, 59%)	104 (0%, 71%)
Total	678 (21%, 57%)	658 (5%, 41%)	674 (6%, 48%)	440 (21%, 57%)	531 (8%, 37%)	463 (9%, 38%)

192 Table 2 legend. Includes complete and partial gene sequences. A more detailed comparison of
193 immune genes annotated in this study, with those identified in other marsupials and humans is
194 available in Supplementary Table 2 within Additional file 2. The first percentage represents $\geq 90\%$
195 overlap and the second represents $\leq 10\%$ overlap between automated and manual annotations of

196 the respective immune genes for each species. Values for the NCBI annotation of the 2021 platypus
197 genome are presented here.

198 Overall, the immune gene repertoire of the koala, woylie, wombat, antechinus, and numbat was
199 similar to other marsupials [58, 82], with marsupial-specific genes and eutherian orthologs identified.
200 Relatively conserved immune genes such as TLRs and constant regions of TCR and IG, as well as
201 polymorphic genes such as MHC and NK receptors, were identified in all five species. Numerous koala
202 immune gene sequences have been characterised previously due to their involvement in chlamydiosis
203 and koala retrovirus which threaten populations [56]. These include MHC [57, 83-85], IG [58], TCR
204 [57], NK receptors [59] and selected cytokines [58, 86-89] (Supplementary Table S2 in Additional file
205 2). We mapped the location of these genes within the current version of the genome, and identified
206 additional new sequences within the LRC, IG and cytokine families (Table 2, Supplementary Table S2
207 in Additional file 2). Immune genes unique to the marsupial lineage were also characterised in the five
208 species studied here. These included MHC class II genes DA, DB and DC, TLR1/6 and TCR μ . Large
209 marsupial-specific gene expansions within the LRC NK receptors were characterised in all five species,
210 as well as reduced gene content within the NKC cluster of NK receptors. Consistent with other
211 marsupials investigated to date Ig δ was not found in any of the five assemblies [90]. A detailed outline
212 of immune genes annotated in this study compared to those of other marsupials and humans is
213 provided in Supplementary Table S2 within Additional file 2.

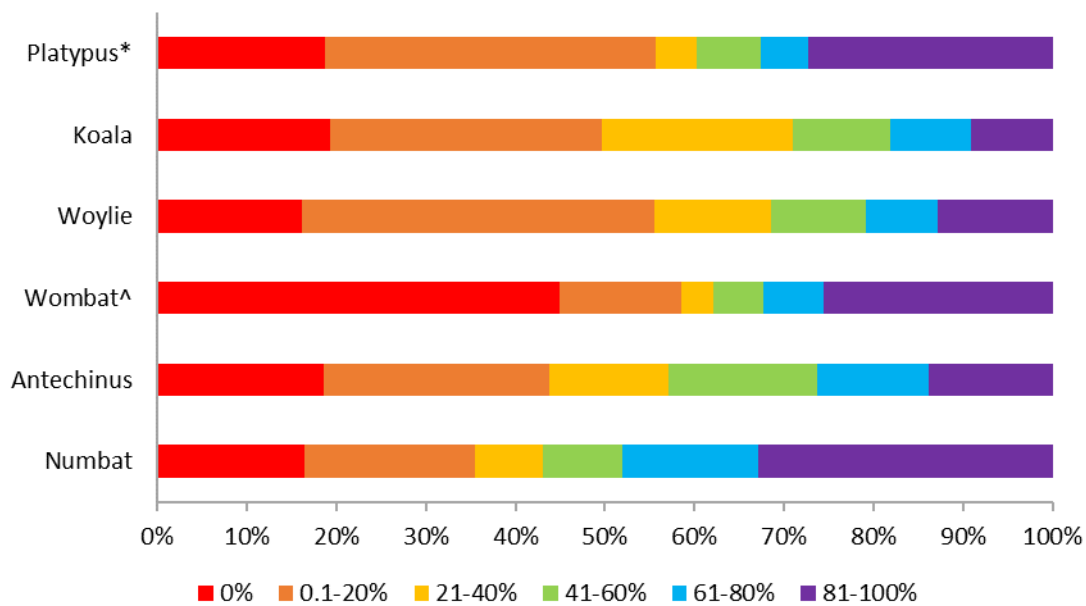
214 **Automated versus manual immune gene annotation**

215 We assessed how well our manual immune gene annotation aligned with automated annotations by
216 Fgenesh++ (2018 platypus, woylie, koala, antechinus, numbat and wombat), MAKER (wombat) and
217 the NCBI pipeline (2021 platypus). Inclusion of the 2021 platypus NCBI and wombat MAKER
218 annotations ensures that any differences in automated and manual immune gene annotation were
219 not due to deficiencies within the Fgenesh++ annotation pipeline, as the woylie, antechinus and
220 numbat genomes were all annotated with Fgenesh++ using the same parameters.

221 Automated annotation pipelines failed to characterise the complete immune repertoire of the
222 platypus or any of the five marsupial species (Figure 1). Only 21.27%, 5.66%, 6.89%, 21.82%, 8.68%,
223 9.07% of immune genes were correctly annotated by the automated pipeline in the 2021 platypus,
224 koala, woylie, wombat, antechinus, and numbat respectively, defined as $\geq 90\%$ overlap in genomic
225 coordinates of immune genes between our manual annotations and the automated annotations
226 (Figure 1). Interestingly, more immune genes were correctly annotated by the automated software in
227 the low-quality wombat, antechinus, and numbat genomes than the high-quality platypus, koala and
228 woylie genomes. This inverse relationship between genome quality and proportion of correctly
229 annotated immune genes is likely related to the characterisation of additional divergent and
230 polymorphic genes such as MHC class I and II in woylie, koala and platypus, which could not be
231 identified by automated or manual annotation in the wombat, antechinus, and numbat due to
232 genome fragmentation (Table 3). All genomes analysed in this study displayed a high proportion of
233 immune genes which were very poorly annotated by automated pipelines ($\leq 10\%$ overlap between
234 immune gene coordinates from manual versus automated annotation); 57.01%, 41.78%, 48.96%,
235 57.01%, 37.05% and 38.22% for 2021 platypus, koala, woylie, wombat, antechinus, and numbat
236 respectively (Figure 1).

237 A breakdown of this analysis by immune family revealed that marsupial- and monotreme-specific
238 immune genes which are not orthologous to those in eutherians were generally poorly annotated,
239 regardless of automated pipeline or genome quality (Supplementary Figure 1). This was particularly
240 the case for TCR and IG gene families, with up to 88% of genes in these families incorrectly annotated
241 by automated pipelines ($\leq 10\%$ overlap) amongst the six species (Table 2). This is likely due to highly
242 duplicated variable gene segments that do not encode conventional exon-intron splice sites which
243 may hinder annotation with automated pipelines. Poor gene annotations of TCR and IG families was
244 somewhat recovered at the exon level, as some TCR and IG variable gene segments were annotated
245 as exons by automated pipelines. Correct annotation ($\geq 90\%$ overlap) of the TCR family increased from
246 0-2% at the gene level to 2-15% at the exon level amongst the six genomes (Supplementary Figure 2).

247 This improvement was even greater for the IG family, with an increase from 0-2% correct annotation
 248 at the gene level to 15-43% at the exon level amongst the six genomes (Supplementary Figure 2).
 249 Despite this, up to 67% of TCR and IG variable segments were still not annotated at the exon level (0%
 250 overlap) amongst the six genomes, highlighting the difficulty in automated annotation of these
 251 regions. Similarly, marsupial-specific gene expansions within the leukocyte receptor complex (LRC)
 252 and monotreme-specific gene expansions within the natural killer complex (NKC) family of NK
 253 receptors were also poorly annotated by automated pipelines (Supplementary Figure 1). As with TCR
 254 and IG families, correct annotation increased from the gene- (0-28% marsupial LRC, 31% platypus NKC)
 255 to exon-level (6-65% marsupial LRC, 79% platypus NKC) (Table 2, Supplementary Figure 2), likely due
 256 to the presence of variable numbers of duplicated immunoglobulin superfamily (IGSF) domains and C-
 257 type lectin (CLEC) domains within each LRC and NKC gene respectively.



258

259 Figure 1. Percentage overlap of genomic coordinates between manual and automated annotations of
 260 immune genes in six genomes. *Denotes automated annotation by NCBI and ^denotes automated
 261 annotation by MAKER. The remaining genomes were annotated using Fgenesh++.

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

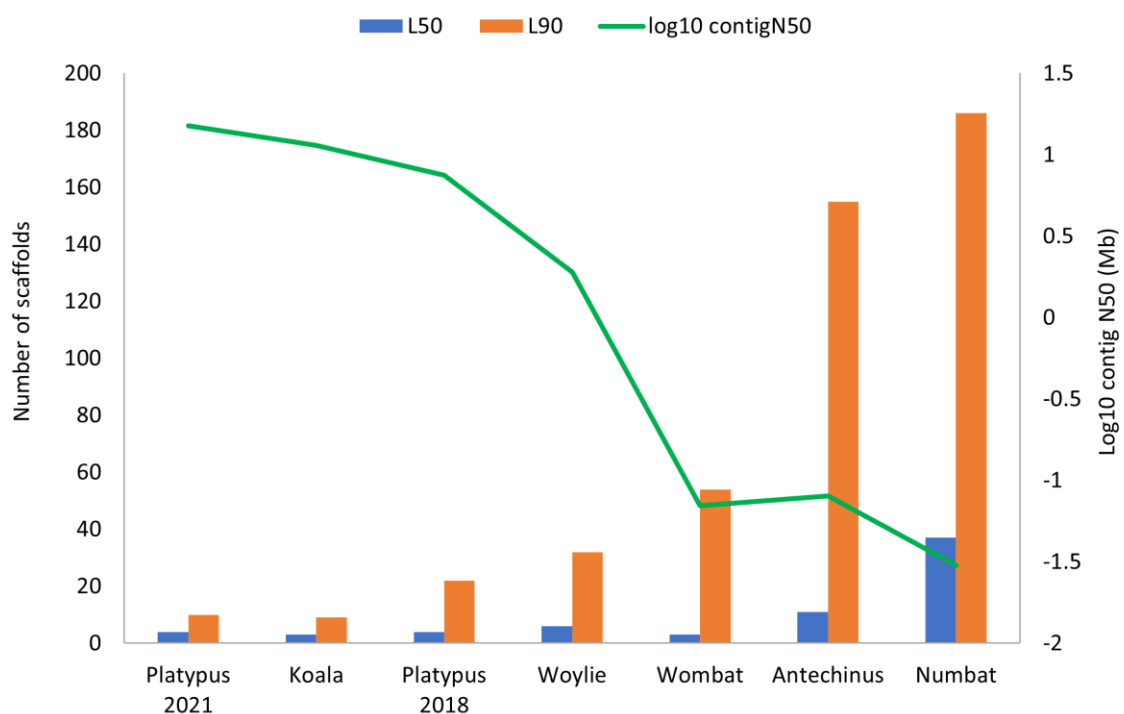
265 This pattern of poor immune gene annotation was not an artefact of inherent differences between
266 automated annotation pipelines amongst the six genomes (NCBI, MAKER and Fgenesh++) nor genome
267 quality, as similar patterns were observed for Fgenesh++ annotations of the 2021 platypus and
268 wombat genome generated as part of this study (Supplementary Figure 3, Supplementary Figure 4).
269 Generally, the Fgenesh++ annotation resulted in fewer correctly annotated immune genes ($\geq 90\%$
270 overlap) compared to NCBI (2021 platypus) or MAKER (wombat) (Supplementary Figure 3). Although,
271 the proportion of missing immune genes (0% overlap) was higher in the NCBI (2021 platypus) and
272 MAKER (wombat) annotation than the Fgenesh++ annotation of both species genomes. As with NCBI
273 and MAKER, Fgenesh++ poorly annotated TCR and IG families at the gene-level (Supplementary Figure
274 4) in the high-quality platypus and low-quality wombat. Correct annotations were somewhat
275 recovered at the exon-level in both genomes (Supplementary Figure 5), although, the number of
276 missing TCR and IG exons in the Fgenesh++ annotation was almost half that of NCBI and MAKER in
277 platypus and wombat respectively.

278 Relationship between genome quality and manual immune gene annotation

279 Manual annotation of immune genes across the koala, woylie, wombat, antechinus and numbat
280 genomes, and mapping of previous annotations to both the 2018 and 2021 versions of the platypus
281 genome, highlighted a clear relationship between immune gene fragmentation and genome quality
282 (Figure 2). Overall, the high-quality koala, 2021 platypus and woylie genomes all contained complete
283 immune gene family clusters, which were highly fragmented in the lower quality wombat, antechinus,
284 and numbat genomes. Fragmentation was particularly evident within families which contain genes
285 that do not share orthology to those in eutherians, such as LRC NK receptors and TCR μ , and highly
286 duplicated families such as MHC (Figure 3).

287 To rule out species-specific differences in our direct assessment of assembly quality on immune gene
 288 annotation, we annotated a previous version of the platypus genome from 2018 (GCA_002966995.1)
 289 with Fgenesh++ to enable comparison with our Fgenesh++ annotation of the 2021 platypus genome
 290 (GCA_004115215.4) also generated as part of this study. Compared to the 2021 assembly, the 2018
 291 platypus assembly was more fragmented given the 6-fold increase in the number of contigs, 14-fold
 292 increase in the number of scaffolds, and associated 2-fold decrease in contig N50 and 4-fold decrease
 293 in scaffold N50 between the two assemblies. Despite these metrics, the 2018 platypus assembly is still
 294 highly contiguous as it was generated using long-read data.

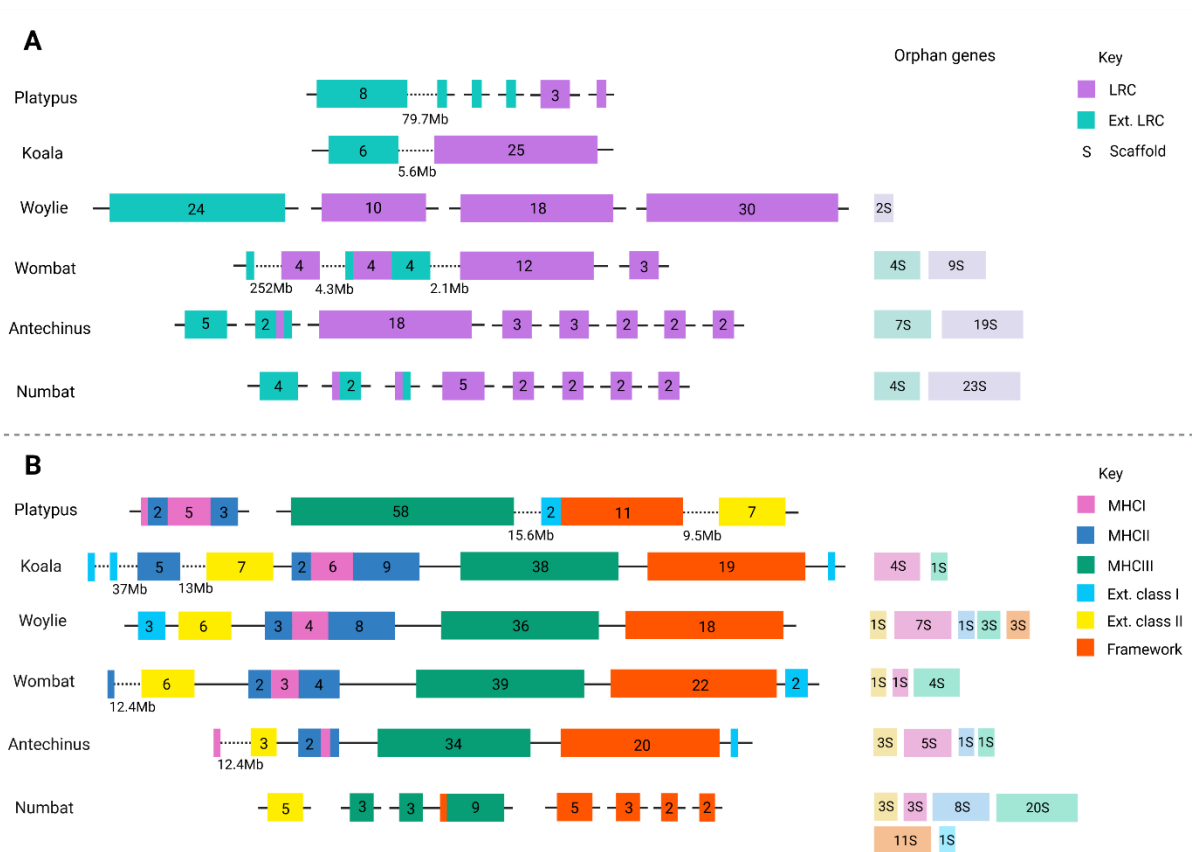
295 To investigate the relationship between immune gene fragmentation and genome quality further, we
 296 calculated the number of scaffolds which encoded 50% (L50) and 90% (L90) of manually annotated
 297 immune genes in each of the seven genomes from six species (Figure 2).



298 Figure 2. L50 and L90 immune gene metric for seven genomes from six species, compared to log₁₀
 299 contig N50.

300 The 2021 platypus, koala and woylie had an L90 of 10, 9 and 36 respectively, which suggests immune
301 gene families were highly contiguous within all three genomes (Figure 2). Complete coding sequences
302 were identified for 98% and 95% of immune genes in koala and woylie respectively. In addition, 90%
303 of annotated immune genes were located on scaffolds greater than 33.3 Mbp, 75 Mbp and 1 Mbp in
304 the 2021 platypus, koala, and woylie respectively. Complex multi-gene immune families such as MHC,
305 NK receptors and TCR were highly intact in all three species. The koala and woylie MHC regions were
306 both primarily located on a single scaffold (Figure 3). Class I and II genes were interspersed, and
307 flanked by class III, framework and extended class I and II gene clusters, which reflected the MHC
308 organisation of other marsupials (Figure 3) [18, 57]. Unlike marsupials, the platypus MHC is encoded
309 within a pseudoautosomal region of two sex chromosomes. MHC class I and II genes were interspersed
310 in a single cluster on chromosome X3, and class III, extended class I and II, and framework genes
311 located in a single cluster on chromosome X5 (Figure 3) in the 2021 assembly [41]. Large gene
312 expansions within the LRC NK receptors were encoded on a single scaffold in koala and six scaffolds in
313 woylie (Figure 3). The number and type of monotreme NK receptor genes differs to marsupials, as
314 they have a large expansion within the NKC gene cluster and reduction within the LRC gene cluster
315 [72]. More than 80% of platypus NKC genes were located in a single cluster on chromosome 17, with
316 LRC genes located on 5 different chromosomes in the 2021 assembly [72]. Fragmentation of the LRC
317 cluster is not a factor of genome quality but reflects the evolutionary history of this immune family
318 [72]. The four TCR loci (α/δ , β , γ and μ) were encoded in single clusters on three chromosomes in
319 platypus 2021 assembly and single scaffolds in koala. The TCR loci were fragmented across up to three
320 scaffolds in woylie. This includes genes known to flank these loci in other marsupials, which enabled
321 resolution of TCR locus organisation in these species, and confirmed gene synteny across marsupials,
322 human and mouse as identified previously [18].

323



324

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

326 Figure 3 legend. The number of genes within each cluster are given, as well as scaffold counts of

327 orphan genes (genes on single scaffolds). In A, LRC genes are purple, extended LRC genes are teal. In

328 B, MHC class I genes are red, class II blue, class III green, extended class I pink, extended class II yellow

329 and framework genes orange. Large distances between genes are given below the scaffold, otherwise

330 the distance between genes and/or clusters was within the expected range for each family. Figure

331 created with BioRender.com.

332 Fragmentation of immune genes in the wombat genome differed between immune families, with an

333 L90 of 56 (Figure 2). 22% of scaffolds encoding immune genes were shorter than 100Kb and partial

334 coding sequences were identified for 7% of annotated immune genes. The MHC region was relatively

335 contiguous in the wombat, with 92% of genes encoded on a single scaffold (Figure 3). Although, a

336 number of MHC genes were encoded as orphan genes to the main MHC cluster, indicating this family

337 is misassembled in the wombat genome. In addition, some MHC genes could not be identified in the

338 wombat genome, while only single copies could be identified for others which are known to be
339 duplicated in all other marsupials studied to date (Additional file 2). While this reduced MHC gene
340 content in the wombat may reflect the true MHC gene repertoire of this species, it is likely MHC genes
341 could not be annotated due to assembly error. The LRC cluster was highly fragmented across 16
342 scaffolds (Figure 3), of which more than 80% encoded a single gene and were less than 10kb in length.
343 Extended LRC and LRC genes were interspersed, likely due to mis-assembly of the region as these
344 genes should be located in separate clusters as observed in koala and woylie (Figure 3). TCR α , β and γ
345 loci were encoded on individual scaffolds, however TCR μ was fragmented across 10 scaffolds, with
346 34% of genes located on individual scaffolds of less than 15Kb. While the TCR β locus was encoded in
347 a single cluster in the wombat, half of the locus was in the reverse orientation. This organisation is
348 unusual amongst mammalian TCR and is likely a result of the HiC scaffolding error and not a true
349 inversion.

350 Immune gene families were highly fragmented in the antechinus and numbat genomes, with an L90
351 of 156 and 218 respectively (Figure 2). 29% and 43% of immune genes were located on scaffolds less
352 than 100Kb, and partial coding sequences were identified for 5.7% and 10.8% of immune genes, in
353 antechinus and numbat respectively. Complex multi-gene families such as MHC, NK receptors and TCR
354 were highly fragmented, with individual genes or exons located on short scaffolds. While 86% of MHC
355 genes were located on a single scaffold in antechinus (Figure 3), genome fragmentation prevented the
356 identification of additional MHC genes, hence the true MHC gene content could not be determined.
357 The numbat MHC region was highly fragmented across 52 scaffolds, 63% of which were less than
358 100Kb in length (Figure 3). Large gene expansions of LRC NK receptors were fragmented across 34
359 scaffolds in antechinus and numbat, of which 67% (antechinus) and 35% (numbat) were less than
360 10Kb, and 76% of scaffolds encoded individual LRC genes in both species (Figure 3). Similar to wombat,
361 extended LRC and LRC genes were interspersed, likely a mis-assembly as these genes should be
362 encoded within separate clusters as observed in koala and woylie. All four TCR loci were fragmented
363 in numbat, and all except TCR α in antechinus, with individual loci encoded across up to 6 scaffolds in

364 numbat and 19 in antechinus. Low contiguity within genomic regions encoding immune gene families
365 in the antechinus and numbat limited investigation of genomic organisation, synteny and evolution in
366 these species.

367 This relationship between genome quality and immune gene fragmentation is not an artefact of
368 species-specific differences in immune gene repertoires. Comparison of manual immune gene
369 annotations in the 2021 and 2018 platypus genome assemblies revealed similar patterns of immune
370 gene fragmentation in the lower-quality 2018 assembly (Supplementary Figure 6 and 7). The 2018
371 platypus assembly had an L90 metric of 22, indicating immune gene clusters were intact within this
372 genome but not to the extent of the 2021 assembly (L90 of 10) (Figure 2). In the 2018 assembly only
373 28% of NKC genes were encoded on a single scaffold (compared to 80% in the 2021 assembly), the
374 MHC was encoded across six scaffolds (compared to two in the 2021 assembly), and only 2 of the 4 TCR
375 clusters were intact (all were intact in the 2021 assembly). Automated annotation of both assemblies
376 with FgenesH++, and comparison with our manual immune gene annotations, yielded the same result
377 as presented for the five marsupial genomes: immune genes are poorly characterised by automated
378 pipelines regardless of genome quality. In the 2021 and 2018 assemblies, a similar proportion of
379 immune genes were correctly annotated (10% and 9% respectively) and not annotated (10% and 15%
380 respectively) by FgenesH++ (Supplementary Figure 6). As observed in the five marsupial genomes, TCR
381 and IG were the most poorly annotated families by FgenesH++ in both platypus assemblies
382 (Supplementary Figure 7).

383 Discussion

384 By manually annotating immune genes in five marsupial genomes and two versions of the platypus
385 genome, all varying qualities, we have confirmed that genome quality is directly linked to our ability
386 to annotate complex immune gene families. Without long reads and scaffolding technologies, immune
387 genes are scattered across many individual scaffolds and gene family organisation and evolution
388 cannot be elucidated. We conclude that long-read data, with or without HiC technology, to generate

389 a high-quality genome assembly with a contig N50 of at least 1MB is required to investigate immunity
390 and disease in wildlife. However, a kitchen sink approach to genome sequencing and assembly will
391 enable complete reconstruction of complex and duplicated families such as MHC, TCR and LRC NK
392 receptors as in the platypus 2021 and koala genomes.

393 The immune gene repertoire of the koala, woylie, wombat, antechinus and numbat was similar to
394 other marsupials such as Tasmanian devil [46, 49, 53], tammar wallaby (*Macropus eugenii*) [74, 91-94]
395 and grey short-tailed opossum (*Monodelphis domestica*) [82]. The platypus immune gene repertoire
396 has been characterised previously [41], and we identified their location within both the 2021 and 2018
397 genome assemblies. Fewer MHC genes were identified in the wombat, antechinus, and numbat,
398 compared to the platypus, koala, and woylie (Table 2, Supplementary Table S2 in Additional file 2).
399 This is likely due to poor read assembly within this highly variable and duplicated region of the
400 genome, rather than a true reduction in MHC gene content within these three species however,
401 further investigation into the MHC gene repertoire of additional marsupial species is required. The
402 assembly of a complete MHC cluster in the platypus, koala and woylie is due to the ability of long reads
403 to span duplicated and variable sequences, which enables assembly algorithms to accurately
404 reconstruct this complex region of the genome.

405 **Automated annotation poorly characterises immune genes in non-model species**

406 Despite mammalian BUSCO scores of up to 94.1% amongst the seven genomes in this study, indicating
407 that the genomes were “functionally complete”, on average 59% of immune genes were not
408 accurately annotated ($\leq 80\%$ overlap) and 21% of genes were not annotated (0% overlap) by the
409 automated software Fgenesh++ and MAKER, nor the NCBI pipeline, compared to our manual
410 annotations (Figure 3). Aside from TCR and IG, the majority of immune genes incorrectly annotated or
411 missing from the automated annotations were divergent genes not orthologous to those in eutherian
412 mammals, such as MHC, marsupial-specific gene expansions within the LRC and monotreme-specific
413 gene expansions within the NKC. Given their divergence, these genes often have low or no BLAST

414 homology to nucleotide or protein databases. Gene models generated by automated annotation
415 software are hypotheses based on supporting evidence such as RNAseq data and homology to
416 nucleotide and protein databases. While immune transcripts were identified in the transcriptomes
417 from these species, RNAseq data only supported gene models for a low proportion of MHC, LRC and
418 NKC genes. RNAseq data only supported 8-16% of LRC gene predictions and 16-37% of MHC gene
419 predictions amongst the four marsupial genome annotations which used RNAseq data as gene model
420 evidence (koala, woylie, antechinus and numbat). Similarly, around 60% of NKC genes within the
421 platypus genomes were supported by RNAseq data. Overall, RNAseq data did not provide enough
422 evidence to support gene models for ~20% of immune genes within the genome. Some immune genes
423 may not have been expressed in the tissue sequenced, were expressed at low levels, or were
424 fragmented. For human and mouse, comprehensive and curated gene sets such as GENCODE and
425 RefSeq are available to guide gene model predictions, comprising data from more than 10,000 RNA
426 experiments and decades of dedicated work in this field [95, 96]. Given time, budget and sample
427 constraints for wildlife, these curated gene sets are not available, hence RNAseq evidence is
428 incomplete resulting in deficient gene models by automated annotation software.

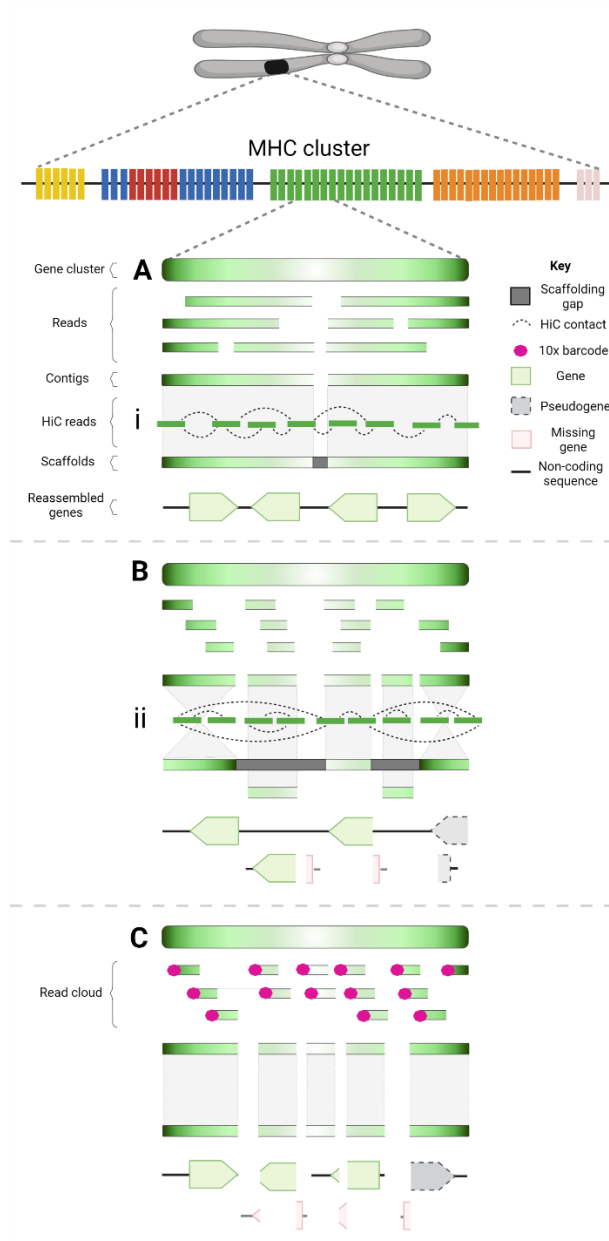
429 It is not surprising that TCR and IG V segments were poorly or not annotated by all automated pipelines
430 used to annotate the genomes in this study. These genes are notoriously difficult to characterise and
431 are manually annotated in the human and mouse genome on Ensembl using the International
432 Immunogenetics Information System (IMGT) database [38, 97]. Alignment of mature IG and TCR
433 sequences from RNAseq data to the genome results in poor automated annotation, as V segments
434 utilize different sequence signal splice sites to introns, which are not recognized by the open reading
435 frame prediction algorithms. Indeed, RNAseq evidence only supported 7% to 18% of TCR V segment
436 and 0% to 6.9% of IG V segment gene predictions by automated pipelines amongst the four
437 marsupial and platypus genomes. V sequences from three marsupials and two monotremes are
438 available in IMGT, however as non-model species, they are not included in the scope for manual
439 annotation by Ensembl or NCBI, so these important functional features are not annotated.

440 Our results highlight the importance of manual annotation and curation of complex and variable
441 immune genes, and caution reliance on BUSCO metrics to assess functional completeness of a
442 genome. If this pattern is observed more widely across non-model species and other complex gene
443 families, functionally important genes may not be accurately represented in genome annotations,
444 which will flow on to downstream applications [36, 98]. While automated annotation is required to
445 keep pace with the rapid sequencing of genome assemblies, manual gene characterisation is still the
446 gold standard for genome annotation [95] and is conducted for the human, mouse, zebrafish and rat
447 genomes on Ensembl [99]. For non-model species, manual annotation is conducted by individual
448 research groups following genome assembly accession with NCBI or Ensembl, who conduct in-house
449 automated annotation for some but not all species [100, 101]. These highly valuable manual gene
450 annotations are not incorporated into the Ensembl annotation release but are often listed in the
451 supplementary materials of multiple individual publications. NCBI does have some capacity to
452 incorporate manual changes to existing annotation records [102]. Changes to multiple annotations,
453 such as adding new genes as is the case in this study, require the genome to be re-annotated, which
454 is not feasible for all research groups. Given NCBI and Ensembl annotations are widely used by the
455 scientific community, these institutions should consider incorporating manual gene annotations into
456 the annotation record or provide scope for permanently storing this valuable data alongside the
457 respective assembly.

458 **Genome quality correlates with immune gene fragmentation**

459 As expected, we found that genome quality directly correlates with likelihood that an immune gene
460 family was assembled and annotated correctly. Immune genes fragment as genome quality declines
461 (Figure 2 and 3). This highlights the importance of long reads and HiC scaffolding to re-assemble
462 complex gene families (platypus, koala, woylie), which are poorly assembled in short read and linked-
463 read assemblies (wombat, antechinus, numbat). Figure 4 provides a graphical representation of the
464 impact of different sequencing technologies on the assembly and fragmentation of immune gene
465 clusters. When the average read or contig length is shorter than the gene length, the assembly

466 algorithm is unable to reconstruct genes, which are fragmented across multiple short contigs [98]. The
467 average immune gene in this study was ~10 kbp in length. Long reads greater than 10 kbp in both
468 platypus, koala and woylie genomes were able to span these genes, whereas the ~150 bp short reads
469 in the wombat, antechinus and numbat genomes were insufficient to re-assemble the entire gene,
470 resulting in gene fragments on short scaffolds. Gene families with copy number variation such as MHC
471 and NK receptors are notoriously difficult to assemble and annotate [26, 29], so it is not surprising
472 these gene families were highly fragmented in the antechinus and numbat genomes. Gene copies
473 within these families can contain almost identical domains, may be pseudogenes and are encoded in
474 clusters within the genome [36]. For example, koala NK LRC genes share up to 96% amino acid
475 sequence identity and are encoded within a single cluster. For these reasons, assembly and annotation
476 of MHC and NK receptors have been used to illustrate improvements in assembly quality. For example,
477 MHC class I genes were located on a single contig in a recent release of the human genome [29],
478 however the highly repetitive MHC class II locus remains unresolved [29].



479

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

482 Figure 4 legend. The impact of long-read (A – platypus, koala and woylie), short-read (B – wombat)
 483 and 10x Chromium linked read (C – antechinus and numbat) sequencing technologies, alone or in
 484 combination with HiC scaffolding (i – koala & platypus, and ii – wombat), on the assembly of complex
 485 and repetitive immune gene clusters such as the MHC. Colour gradient represents gene orientation
 486 (A) Long read sequencing generates reads which span complex and repetitive sequences, resulting in
 487 long contigs and scaffolds which contain multiple immune genes with complete coding sequences. (B)

488 Short-read sequencing generated reads which are unable to span immune genes, hence reads are
489 assembled into multiple short contigs which end when the algorithm is unable to assemble a repetitive
490 and complex immune gene sequence. (C) In linked-read sequencing, individual DNA molecules are
491 partitioned into gel beads and identical barcodes attached, then sequenced using short-read
492 technology resulting in read clouds [103]. As no individual read within the cloud spans the entire
493 length of the DNA molecule, the algorithm is unable to assemble repetitive and complex sequences,
494 resulting in multiple short contigs similar to a short-read assembly. Short contigs in B and C result in
495 fragmentation of immune genes, leading to false pseudogenization and “missing” genes. (i) HiC
496 sequencing provides contact information for DNA sequences located in close proximity within the
497 nucleus, as frequency decreases with increasing linear distance within the genome assembly [104].
498 This contact information can be used to cluster, order and orient contigs into chromosome-size
499 scaffolds [105]. Long contigs scaffolded with HiC result in near-complete reconstruction of immune
500 gene clusters. (ii) Short contigs scaffolded with HiC generates what appears to be long scaffolds,
501 however complex immune gene clusters are incomplete. As multiple HiC contacts can span the length
502 of the contig, the correct contig orientation is not apparent leading to inversions and mis-placed
503 contigs during scaffolding. This leads to incorrect orientation of genes, which can cause
504 pseudogenization and/or gene fragmentation. Manual immune gene annotation reveals that the true
505 gene complement of the immune cluster is not contained within the scaffolded sequence. Figure
506 created with BioRender.com.

507 HiC scaffolding of contigs derived from platypus and koala long reads resulted in complete and
508 accurate reassembly of immune gene clusters in both genomes (Figure 4A). Conversely, HiC scaffolding
509 of contigs from wombat short reads resulted in immune gene fragmentation (Figure 4B), reflected in
510 the high immune gene L90 for the wombat genome (Figure 2). Both the koala and wombat genomes
511 were scaffolded with DNazoo HiC data using the same 3D-DNA pipeline [63, 64, 106]. This result
512 underscores the importance of assessing annotations when determining genome quality, as the
513 wombat genome is classified as chromosome-length yet is highly fragmented within functionally

514 important genomic regions. Input genome assembly contiguity is known to influence HiC scaffolding
515 ordering and orientation errors [107], despite claims that HiC scaffolding with 3D-DNA generates
516 chromosome-length scaffolds from US\$1,000 short read contigs [63]. Problems with HiC scaffolding
517 within repetitive and duplicated regions are well documented [31, 107, 108], which is exacerbated by
518 short contigs [107]. Modelling of human genome scaffolding performance using 3D-DNA revealed
519 scaffold chimeras, ordering and orientation errors increased as contig length decreased [107]. While
520 the koala and platypus genomes used as input to HiC scaffolding benefited from polishing with short
521 read data and optical mapping [57], HiC scaffolding is insufficient to recover the majority of immune
522 clusters from a fragmented genome.

523 The 3D-DNA pipeline orientates contigs within scaffolds by maximizing contact frequency between
524 contig ends [64]. Short contigs, such as those from the wombat, would have multiple contacts that
525 span the length of the contig. This means both true and false contig orientations would have a similar
526 frequency, resulting in errors such as the partial inversion of the TCRB locus which is likely false
527 (Additional file 2). At a gene level, these errors lead to the misplacement of genes on short scaffolds
528 outside the main immune cluster and false pseudogenisation (Figure 4B). Long contigs, such as those
529 from the koala, would have fewer contacts that span the length of the contig, hence the true
530 orientation of the contig would be clear from the higher contact frequency at the correct joining end.
531 The combination of long contigs which span repetitive and highly heterozygous regions with HiC
532 scaffolding maximizes contiguity within immune gene clusters (Figure 4A).

533 10x Chromium linked-read sequencing was insufficient to accurately re-assemble immune gene
534 clusters in our study (Figure 4C). While this technology is no longer available for genome sequencing,
535 acknowledging the limitations of this technology for immune gene annotation remains valid in order
536 to make use of existing 10x genomes. Complete marsupial immune gene clusters can span hundreds
537 of kilobases to megabases, as shown by annotation of the complete MHC, NK receptor and TCR regions
538 in the koala (Additional file 2). DNA molecules input to 10x library preparation were on average 74

539 kbp and 23 kbp in antechinus and numbat respectively. This molecule size only spanned smaller
540 immune clusters in the antechinus, such as the 70 kbp TRG locus, but was insufficient to span any
541 cluster in the numbat. Interestingly, the antechinus MHC cluster appears to be intact (Figure 3),
542 however manual annotation revealed multiple genes were “missing” within the scaffold and instead
543 were located on individual short scaffolds. Regardless of input DNA molecule length, 10x libraries are
544 still subject to the limitations of short-read sequencing regarding assembly of complex sequences.
545 Antechinus and numbat 10x libraries were sequenced as short ~150 bp reads, hence while reads can
546 be assigned back to the corresponding input DNA molecule, no single read spans the molecule length.
547 Gaps between the reads make *de novo* assembly of repetitive and complex immune sequences
548 difficult, often resulting in termination of contig extension and gene fragments scattered across short
549 scaffolds [109-111]. These gene fragments can be misinterpreted as pseudogenes owing to loss of
550 up/downstream coding regions (Figure 4C). For example, antechinus and numbat NK LRC genes share
551 up to 97% and 98% amino acid sequence identity amongst the genes identified in each species
552 respectively. The LRC should be encoded within a single cluster, as in the koala genome (Figure 3).
553 Instead, the antechinus and numbat LRC clusters are fragmented across 33 and 34 scaffolds
554 respectively.

555 As the global biodiversity crisis deepens, the need to sequence eukaryotic life while it remains is
556 imperative [1, 7, 8]. High quality genomes, using a combination of long-read and HiC, have recently
557 been generated for a number of wildlife species [8], which have been used to answer questions
558 involving chromosome evolution [112], comparative genomics [113] and runs of homozygosity [114]
559 amongst others. Our results show that high-quality genomes are also necessary to study immune
560 genes in wildlife.

561 Draft quality *de novo* genomes, in this study the antechinus and numbat (linked reads), have limited
562 capacity for usefully informing immunogenetics studies as only partial sequences will be identified for
563 most immune genes. A scaffold-quality genome, in this study the woylie and 2018 platypus assembly

564 (long-reads) or wombat (short-reads with HiC), would be suitable for immune marker development
565 targeting most immune gene families, and studying TCR and IG diversity. Long-reads will provide
566 contiguity within duplicated MHC and NK families, which should reassemble into complete clusters.
567 HiC data may resolve some immune gene clusters from a short-read assembly, however, may
568 introduce errors as discussed earlier. Finally, the kitchen sink approach, in this study the 2021 platypus
569 and koala genomes (multiple data types), will accurately assemble immune gene clusters, which is
570 essential for investigating genomic organisation, synteny and evolution. In the context of wildlife
571 disease both sample availability and research dollars will dictate the type of data able to be generated
572 for genome assembly, from this study we recommend a minimum of long-read sequencing such as
573 PacBio HiFi to allow for complete annotation of immune gene regions

574 **Potential implications**

575 The biodiversity crisis and increasing impact of wildlife disease on animal and human health provides
576 impetus for studying immune genes in wildlife. Genomes are now available for many wildlife species,
577 however utility of these assemblies for annotating complex immune gene families is unknown. We
578 have provided an assessment of complex immune gene annotation across genomes of varying quality,
579 using immune genes in five marsupials and one monotreme as an example. Genome quality directly
580 influenced the reassembly of immune gene clusters, and ability to investigate evolution, organisation,
581 and true gene content of the immune repertoire. A high-quality genome generated from long-reads,
582 with or without HiC, accurately assembles immune gene clusters. However, draft-quality genomes
583 generated from short-reads with HiC, or the now obsolete 10x Chromium linked-reads, were unable
584 to achieve this. Aside from genome quality, manual annotation of immune genes is required to cover
585 the shortfall in deficient gene models used by automated annotation software. Our results highlight
586 the limitations of different sequencing technologies and established workflows for genome
587 annotation and quality assessment, when applied to non-model species and the investigation of
588 wildlife disease and immunity.

589 Methods

590 Five published marsupial genomes, koala [57, 63, 64], woylie [65], wombat [63], antechinus [66] and
591 numbat [67] (Table 1), and one monotreme genome, platypus [41], were selected for this study based
592 on use of different sequencing technologies (alone and in combination) and variation in assembly
593 quality. These include assemblies generated using multiple data types (koala and platypus), long and
594 short-reads (woylie), short-reads and HiC (wombat) or 10x Chromium linked-reads (antechinus and
595 numbat). BUSCO scores were generated by uploading the six genome assemblies to the Galaxy web
596 platform [115], where the public server at galaxy.org was used to run BUSCOv5.3.2 [35] against the
597 mammalian database.

598 Immune genes were annotated in the koala (phaCin_unsw_v4.1_HiC) [57, 63, 64], antechinus
599 (anrechinusM_pseudohap2.1) [66], woylie (mBetpen1.pri.20210916) [65], wombat (vu-2k) [63, 64]
600 and numbat genome (mMyrfas1.pri.20210917) [67] using multiple search strategies. BLAST was used
601 to search genome assemblies, associated annotation files and/or transcriptomes using published
602 marsupial, monotreme and eutherian immune gene sequences as queries, with default parameters
603 and an e-value threshold of 10 so as not to exclude any potential gene candidates. HMMERv3.2 [116]
604 was also used to identify putative genes within immune families that are known to contain
605 duplications in other marsupials, such as NK receptors. Hidden markov models (HMM) were
606 constructed using ClustalW alignments of published marsupial and eutherian immune gene sequences
607 constructed in BioEditv7.2.5 [117], which were then used to search all genomes and transcriptomes
608 using HMMER v3.2 with an e-value threshold of 10. For variable segments of T cell receptor and
609 Immunoglobulin families, recombination signal sequences (RSS) downloaded from the IMGT database
610 [97] and published koala sequences [57], were aligned using ClustalW in BioEditv7.2.5 [117] and used
611 to construct HMM. These RSS HMM were then used to search each genome using HMMERv3.2 [116],
612 to identify conserved RSS which flank each variable segment. For NK receptors, putative NKC and LRC
613 sequences from BLAST+v2.7.1 [68] and HMMERv3.2 [116] searches were queried against the swissprot

614 nonredundant database, and any sequences with top hits to swissprot NK genes, marsupial-specific
615 NK genes or the protein families database (Pfam) [118] immunoglobulin domain PF00047 or C-type
616 lectin domain PF00059 HMM model were retained. IGSF domains within putative NK sequences from
617 each species were identified using the simple modular architecture research tool (SMART) database
618 [119], and IGSF domains within 5 kbp were considered exons of a single LRC gene. Putative immune
619 genes were named following the appropriate nomenclature for each family, with duplicated genes
620 named according to their genomic location from the 5' to 3' end of the locus. For each immune gene
621 family, amino acid sequences from all five species, in addition to other marsupial, monotreme and
622 eutherian sequences, were aligned using ClustalW in BioEditv7.2.5 [117]. This alignment was then
623 used to construct neighbour-joining phylogenetic trees in MEGAXv10.2.4 [120] using the p-distance
624 method, pairwise deletion and 1000 bootstrap replicates.

625 To investigate the impact of genome assembly quality on immune gene annotation, and discount
626 species differences from our assessment, Fgenesh++ v7.2.2 [33] was used to annotate two different
627 assemblies of the platypus genome; GCA_004115215.4 generated using multiple data types [41], and
628 GCA_002966995.1 generated using only long and short-read data. In addition, Fgenesh++ v7.2.2 [33]
629 was used to annotate the koala and wombat genome assemblies to investigate the influence of
630 automated annotation method on immune gene annotation. To generate mRNA evidence for input to
631 Fgenesh++, RNAseq data from 19 platypus tissues and 16 koala tissues accessioned with the NCBI
632 sequence read archive (SRA) (Supplementary Table S3) was used to generate reference-guided global
633 transcriptomes for each genome assembly (koala, platypus GCA_004115215.4 and
634 GCA_002966995.1). No wombat RNAseq data was available on the SRA, hence a global transcriptome
635 was not generated for this species. Briefly, raw RNAseq reads were quality and length trimmed using
636 Trimmomatic v0.39 [121] with the following parameters: ILLUMINACLIP:TruSeq3-SE.fa:2:30:10
637 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25. Over 90.53% of paired trimmed reads were
638 retained for all 35 datasets (Supplementary Table S3). Trimmed reads were then aligned to the
639 respective species genome, and assembly version, using HISAT2 v2.1.0 [122] with default parameters.

640 Resulting sam files were converted to sorted bam files using SAMTOOLS v1.9 [123], then StringTie
641 v2.1.6 [124] used to generate gtf files for each tissue. Tama merge [125] was then used to merge
642 aligned reads for each tissue into a single global transcriptome for each genome assembly (koala,
643 platypus GCA_004115215.4 and GCA_002966995.1), with a 5' threshold of 3 and a 3' threshold of 500.
644 CPC2 [126] was used to determine the coding potential of each transcript, and Transdecoder v2.0.1
645 [127] to predict open reading frames within each transcript, for each global transcriptome.

646 The wombat, koala and two platypus genome assemblies (GCA_004115215.4 and GCA_002966995.1)
647 were annotated using Fgenesh++ v7.2.2 with general mammalian parameters using a custom machine
648 at the Pawsey Supercomputing Centre with 64 CPUs, 256GB RAM and 1TB of disk storage. An
649 optimised gene-finding matrix from Tasmanian devils was used for koala and wombat genome
650 annotations, while the platypus gene finding matrix was used for both platypus genome assembly
651 annotations. Transcripts with the longest open reading frame for each predicted gene were extracted
652 from the global transcriptomes for platypus and koala as outlined in the previous section and used as
653 mRNA-based gene predictions. The compute wall-time required to complete each annotation was as
654 follows: wombat 8 days, 1 hour and 15 minutes, koala 7 days, 8 hours and 38 minutes, platypus
655 GCA_002966995.1 2 days 2 hours and 37 minutes and platypus GCA_004115215.4 1 day, 16 hours and
656 13 minutes.

657

658 **Additional files**

659 File name: Additional file 1

660 File format: .xls

661 Title of data: Supplementary Table S1

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

665 File name: Additional file 2

666 File format: .doc

667 Title of data: Supplementary results

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

676 **Data availability**

677 The published woylie and numbat genome and global transcriptome assemblies are available through
678 Amazon Web Services Open Datasets Program [128], NCBI under BioProject accession PRJNA763700
679 and GigaDB for woylie [129] and PRJNA786364 and GigaDB [130] for numbat. The published koala
680 genome assembly and annotation (phaCin_unsw_v4.1_HiC.fasta) are available from the DNAAzoo
681 website [131]. The published wombat genome assembly and annotation (vu-2k.fasta) are also
682 available from the DNAAzoo website [131]. The published antechinus genome assembly and annotation
683 (anrechinusM_pseudohap2.1.fasta) are available from NCBI under BioProject accession PRJNA664282
684 and GigaDB [129], and published platypus genome assembly and annotation (mOrnAna1.pri.v4) under
685 BioProject accession PRJNA489114. Genomic coordinates for all immune gene sequences annotated

686 in this study are available in Additional file 1. Supporting information for this study is available in
687 Additional file 2. Data contained in this paper for all species and annotations are available in the
688 GigaScience database GigaDB [132], including BUSCO analyses, .gff files of the annotations, and all
689 data used to create the figures and phylogenetic trees.

690 **Declarations**

691 **List of abbreviations**

692 Bacterial artificial chromosome (BAC), basic local alignment search tool (BLAST), benchmarking single
693 copy gene orthologs (BUSCO), complementary DNA (cDNA), devil facial tumour disease (DFTD), giga-
694 base-pair (Gpb), high-throughput chromosome conformation capture (HiC), hidden markov model
695 (HMM), immunoglobulin (IG), immunoglobulin superfamily (IGSF), interferon (IFN), international
696 immunogenetic information system (IMGIT), kilo-base-pair (kbp), leukocyte receptor complex (LRC),
697 major histocompatibility complex (MHC), mega-base-pair (Mbp), National Center for Biotechnology
698 Information (NCBI), natural killer complex (NKC), natural killer receptor (NK), Pacific Biosciences
699 (PacBio), protein families database (Pfam), recombination signal sequence (RSS), simple modular
700 architecture research tool (SMART), single nucleotide polymorphisms (SNPs), T cell receptor (TCR)
701 and toll-like receptor (TLR).

702 **Consent for publication**

703 Not applicable

704 **Competing interests**

705 The authors declare that they have no competing interests

706 **Funding**

707 This work has been funded by the Australian Research Council Centre of Excellence for Innovations in
708 Peptide and Protein Science (CE200100012) and Discovery Project (DP180102465). LS was supported
709 by LP180100244 and PB was supported by an Australian Postgraduate award. YZ is supported by the
710 China Scholarship Council.

711 Authors' contributions

712 LS assembled and annotated the woylie genome and transcriptomes, PB assembled and annotated
713 the numbat genome and transcriptomes, EP assisted with both. EP, PB, LS, YC and YZ annotated
714 immune genes. KB, CJH and EP designed the study. EP drafted the manuscript, all authors read and
715 commented on drafts of the manuscript and have approved the submission.

716 Acknowledgements

717 The authors would also like to acknowledge the generous contribution of the Presbyterian Ladies'
718 College Sydney and Bioplatforms Australia. This work was supported by the Australian Fgenesh++
719 Service (biocommons.org.au/fgenesh-plus-plus) provided by the Australian BioCommons and the
720 Pawsey Supercomputing Research Centre.

721 References

- 722 1. Diaz S, Settle J, Brondizio ES, Ngo HT, Gueze M, Agard J, Balvanera P, Brauman A, Butchart
723 SHM, Chan KMA, et al. IPBES (2019): Summary for policymakers of the global assessment
724 report on biodiversity and ecosystem services of the Intergovernmental Science-Policy
725 Platform on Biodiversity and Ecosystem Services. pp. 56 pages. Bonn, Germany; 2019:56
726 pages.
- 727 2. Scheele BC, Pasmans F, Skerratt LF, Berger L, Martel A, Beukema W, Acevedo AA, Burrowes
728 PA, Carvalho T, Catenazzi A, et al. Amphibian fungal panzootic causes catastrophic and
729 ongoing loss of biodiversity. *Science* 2019;363:1459. 10.1126/science.aav0379.
- 730 3. Hoyt JR, Kilpatrick AM, Langwig KE. Ecology and impacts of white-nose syndrome on bats.
731 *Nature Reviews Microbiology* 2021;19:196-210. 10.1038/s41579-020-00493-5.
- 732 4. Woods GM, Lyons AB, Bettiol SS. A Devil of a Transmissible Cancer. *Tropical Medicine and*
733 *Infectious Disease* 2020;5:50.
- 734 5. Rohr JR, Civitello DJ, Halliday FW, Hudson PJ, Lafferty KD, Wood CL, Mordecai EA. Towards
735 common ground in the biodiversity–disease debate. *Nature Ecology & Evolution* 2020;4:24-
736 33. 10.1038/s41559-019-1060-6.
- 737 6. Hohenlohe PA, Funk WC, Rajora OP. Population genomics for wildlife conservation and
738 management. *Molecular Ecology* 2021;30:62-82. <https://doi.org/10.1111/mec.15720>.
- 739 7. Lewin H, Robinson G, Kress WJ, Baker W, Coddington J, Crandall K, Durbin R, Edwards S,
740 Forest F, Gilbert MT, et al. Earth BioGenome Project: Sequencing life for the future of life.
741 *Proceedings of the National Academy of Sciences (PNAS)* 2018;115:4325-4333.
742 10.1073/pnas.1720115115.
- 743 8. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
744 Functamman A, Kim J, et al. Towards complete and error-free genome assemblies of all
745 vertebrate species. *Nature* 2021;592:737-746. 10.1038/s41586-021-03451-0.
- 746 9. Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E. Bat Biology, Genomes, and
747 the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species.
748 *Annual Review of Animal Biosciences* 2018;6:23-46. 10.1146/annurev-animal-022516-
749 022811.
- 750 10. Zhang G. Bird sequencing project takes off. *Nature* 2015;522:34-34. 10.1038/522034d.

- 751 11. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An
752 Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model
753 Species. PLOS ONE 2012;7:e37135. 10.1371/journal.pone.0037135.
- 754 12. Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D,
755 Vaillancourt RE, Kilian A. A high-density Diversity Arrays Technology (DArT) microarray for
756 genome-wide genotyping in Eucalyptus. Plant Methods 2010;6:16. 10.1186/1746-4811-6-16.
- 757 13. McLennan EA, Grueber CE, Wise P, Belov K, Hogg CJ. Mixing genetically differentiated
758 populations successfully boosts diversity of an endangered carnivore. Animal Conservation
759 2020;23:700-712. <https://doi.org/10.1111/acv.12589>.
- 760 14. Scally A, Yngvadottir B, Xue Y, Ayub Q, Durbin R, Tyler-Smith C. A Genome-Wide Survey of
761 Genetic Variation in Gorillas Using Reduced Representation Sequencing. PLOS ONE
762 2013;8:e65066. 10.1371/journal.pone.0065066.
- 763 15. Robledo-Ruiz DA, Pavlova A, Clarke RH, Magrath MJL, Quin B, Harrison KA, Gan HM, Low
764 GW, Sunnucks P. A novel framework for evaluating in situ breeding management strategies
765 in endangered populations. Molecular Ecology Resources
766 2021;n/a:<https://doi.org/10.1111/1755-0998.13476>.
- 767 16. Lott MJ, Wright BR, Kemp LF, Johnson RN, Hogg CJ. Genetic Management of Captive and
768 Reintroduced Bilby Populations. The Journal of Wildlife Management 2020;84:20-32.
769 <https://doi.org/10.1002/jwmg.21777>.
- 770 17. Irving AT, Ahn M, Goh G, Anderson DE, Wang L-F. Lessons from the host defences of bats, a
771 unique viral reservoir. Nature 2021;589:363-370. 10.1038/s41586-020-03128-0.
- 772 18. Parra ZE, Baker ML, Hathaway J, Lopez AM, Trujillo J, Sharp A, Miller RD. Comparative
773 genomic analysis and evolution of the T cell receptor loci in the opossum *Monodelphis*
774 *domestica*. BMC Genomics 2008;9:1-19.
- 775 19. Glusman G, Rowen L, Lee I, Boysen C, Roach JC, Smit AFA, Wang K, Koop BF, Hood L.
776 Comparative Genomics of the Human and Mouse T Cell Receptor Loci. Immunity
777 2001;15:337-349. [https://doi.org/10.1016/S1074-7613\(01\)00200-X](https://doi.org/10.1016/S1074-7613(01)00200-X).
- 778 20. Sun Y, Wei Z, Li N, Zhao Y. A comparative overview of immunoglobulin genes and the
779 generation of their diversity in tetrapods. Developmental and Comparative Immunology
780 2013;39:103-109.
- 781 21. Kelley J, Trowsdale J. Features of MHC and NK gene clusters. Transplant Immunology
782 2005;14:129-134. <https://doi.org/10.1016/j.trim.2005.03.001>.
- 783 22. Krasnec K, Sharp AR, Williams TL, Miller RD. The opossum MHC genomic region revisited.
784 Immunogenetics 2015;67:259-264.
- 785 23. Kelley J, Walfer L, Trowsdale J. Comparative genomics of natural killer cell receptor gene
786 clusters. PLOS Genetics 2005;1:129-139.
- 787 24. Takeda K, Kaisho T, Akira S. Toll-like receptors. Annual Review of Immunology 2003;21:
788 25. Wong ESW, Papenfuss AT, Belov K. Genomic identification of chemokines and cytokines in
789 opossum. Journal of Interferon and Cytokine Research 2011;31:317-330.
- 790 26. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot
791 CC, Wright MW, et al. Gene map of the extended human MHC. Nature Reviews Genetics
792 2004;5:889-899. 10.1038/nrg1489.
- 793 27. Robinson J, Waller MJ, Parham P, Groot Nd, Bontrop R, Kennedy LJ, Stoehr P, Marsh SGE.
794 IMGT/HLA and IMGT/MHC: sequence databases for the study of the major
795 histocompatibility complex. Nucleic Acids Research 2003;31:311-314. 10.1093/nar/gkg070.
- 796 28. Trowsdale J, Parham P. Mini-review: Defense strategies and immunity-related genes.
797 European Journal of Immunology 2004;34:7-17. <https://doi.org/10.1002/eji.200324693>.
- 798 29. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes
799 IT, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads.
800 Nature Biotechnology 2018;36:338-345. 10.1038/nbt.4060.

- 801 30. Ming L, Wang Z, Yi L, Batmunkh M, Liu T, Siren D, He J, Juramt N, Jambl T, Li Y, Jirimutu.
802 Chromosome-level assembly of wild Bactrian camel genome reveals organization of immune
803 gene loci. *Molecular Ecology Resources* 2020;20:770-780. [https://doi.org/10.1111/1755-](https://doi.org/10.1111/1755-0998.13141)
804 [0998.13141](https://doi.org/10.1111/1755-0998.13141).
- 805 31. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I,
806 Sullivan ST, et al. Single-molecule sequencing and chromatin conformation capture enable
807 de novo reference assembly of the domestic goat genome. *Nature Genetics* 2017;49:643-
808 650. 10.1038/ng.3802.
- 809 32. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool
810 for second-generation genome projects. *BMC Bioinformatics* 2011;12:491. 10.1186/1471-
811 2105-12-491.
- 812 33. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes,
813 pseudogenes and promoters. *Genome Biology* 2006;7:S10. 10.1186/gb-2006-7-s1-s10.
- 814 34. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
815 Functammasan A, Gedman GL, et al. Towards complete and error-free genome assemblies
816 of all vertebrate species. Cold Spring Harbor Laboratory; 2020.
- 817 35. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation
818 Completeness. In: Edited by Kollmar M *Gene Prediction: Methods and Protocols*. New York,
819 NY: Springer New York; 2019: 227-245
- 820 36. Mudge JM, Harrow J. The state of play in higher eukaryote gene annotation. *Nature Reviews*
821 *Genetics* 2016;17:758-772. 10.1038/nrg.2016.119.
- 822 37. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M,
823 Bajic VB, Birney E, et al. EGASP: the human ENCODE genome annotation assessment project.
824 *Genome Biology* 2006;7:S2. 10.1186/gb-2006-7-s1-s2.
- 825 38. Annotation of immunoglobulin and T cell receptor genes
826 [https://m.ensembl.org/info/genome/genebuild/ig_tcr.html]
- 827 39. Hogg CJ, Ottewell K, Latch P, Rossetto M, Biggs J, Gilbert A, Richmond S, Belov K. Threatened
828 Species Initiative: Empowering conservation action using genomic resources. *Proceedings of*
829 *the National Academy of Sciences* 2022;119:e2115643118. 10.1073/pnas.2115643118.
- 830 40. Gordon D, Huddleston J, Chaisson Mark JP, Hill Christopher M, Kronenberg Zev N, Munson
831 Katherine M, Malig M, Raja A, Fiddes I, Hillier LaDeana W, et al. Long-read sequence
832 assembly of the gorilla genome. *Science* 2016;352:aae0344. 10.1126/science.aae0344.
- 833 41. Zhou Y, Shearwin-Whyatt L, Li J, Song Z, Hayakawa T, Stevens D, Fenelon JC, Peel E, Cheng Y,
834 Pajpach F, et al. Platypus and echidna genomes reveal mammalian biology and evolution.
835 *Nature* 2021;10.1038/s41586-020-03039-0.
- 836 42. Quigley BL, Timms P. The Koala Immune Response to Chlamydial Infection and Vaccine
837 Development—Advancing Our Immunological Understanding. *Animals* 2021;11:380.
- 838 43. Madden D, Whaite A, Jones E, Belov K, Timms P, Polkinghorne A. Koala immunology and
839 infectious diseases: How much can the koala bear? *Developmental & Comparative*
840 *Immunology* 2018;82:177-185. <https://doi.org/10.1016/j.dci.2018.01.017>.
- 841 44. Peel E, Belov K. Lessons learnt from the Tasmanian devil facial tumour regarding immune
842 function in cancer. *Mammalian Genome* 2018;10.1007/s00335-018-9782-3.
- 843 45. Murchison EP, Schulz-Trieglaff OB, Ning Z, Alexandrow LB, Bauer MJ, Fu B, Hims M, Ding Z,
844 Ivakhno S, Stewart C, et al. Genome sequencing and analysis of the Tasmanian devil and its
845 transmissible cancer. *Cell* 2012;148:780-791.
- 846 46. Morris B, Cheng Y, Warren W, Papenfuss AT, Belov K. Identification and analysis of divergent
847 immune gene families within the Tasmanian devil genome. *BMC Genomics* 2015;16:
- 848 47. Murchison EP, Tovar C, Hsu AL, Bender HS, Kheradpour P, Rebbeck CA, Obendorf DL, Conlan
849 C, Bahlo M, Blizzard CA, et al. The Tasmanian devil transcriptome reveals schwann cell
850 origins in a clonally transmissible cancer. *Science* 2010;327:84-87.

- 851 48. Hewavisenti RV, Morris KM, O'Meally D, Cheng Y, Papenfuss AT, Belov K. The identification
852 of immune genes in the milk transcriptome of the Tasmanian devil (*Sarcophilus harrisi*).
853 PeerJ 2016;4:1569.
- 854 49. Cheng Y, Stuart A, Morris K, Taylor R, Siddle HV, Deakin JE, Jones M, Amemiya CT, Belov K.
855 Antigen-presenting genes and genomic copy number variations in the Tasmanian devil MHC.
856 BMC Genomics 2012;13:87.
- 857 50. Morris KM, Wright B, Grueber CE, Hogg C, Belov K. Lack of genetic diversity across diverse
858 immune genes in an endangered mammal, the Tasmanian devil (*Sarcophilus harrisi*).
859 Molecular Ecology 2015;24:3860-3872.
- 860 51. Siddle HV, Kreiss A, Eldridge MDB, Noonan E, Clarke CJ, Pyecroft S, Woods GM, Belov K.
861 Transmission of a fatal clonal tumor by biting occurs due to depleted MHC diversity in a
862 threatened carnivorous marsupial. PNAS 2007;104:16221-16226.
- 863 52. Siddle HV, Kreiss A, Tovar C, Yuen CK, Chen Y, Belov K, Swift K, Pearse AM, Hamede R, Jones
864 ME, et al. Reversible epigenetic down-regulation of MHC molecules by devil facial tumour
865 disease illustrates immune escape by a contagious cancer. PNAS 2013;110:5103-5108.
- 866 53. Siddle HV, Sanderson CE, Belov K. Characterization of major histocompatibility complex
867 class I and II genes from the Tasmanian devil (*Sarcophilus harrisi*). Immunogenetics
868 2007;59:753-760.
- 869 54. Cheng Y, Sanderson CE, Jones M, Belov K. Low MHC class II diversity in the Tasmanian devil.
870 Immunogenetics 2012;64:525-533.
- 871 55. Cheng Y, Grueber C, Hogg CJ, Belov K. Improved high-throughput MHC typing for non-model
872 species using long-read sequencing. Molecular Ecology Resources
873 2021;n/a:<https://doi.org/10.1111/1755-0998.13511>.
- 874 56. Quigley BL, Timms P. Helping koalas battle disease – Recent advances in Chlamydia and
875 koala retrovirus (KoRV) disease understanding and treatment in koalas. FEMS Microbiology
876 Reviews 2020;10.1093/femsre/fuaa024.
- 877 57. Johnson RN, O'Meally D, Chen Z, Etherington GJ, Ho SYW, Nash WJ, Grueber CE, Cheng Y,
878 Whittington CM, Dennison S, et al. Adaptation and conservation insights from the koala
879 genome. Nature Genetics 2018;50:1102-1111. 10.1038/s41588-018-0153-5.
- 880 58. Morris K, Prentis PJ, O'Meally D, Pavasovic A, Brown AT, Timms P, Belov K, Polkinghorne A.
881 The koala immunological toolkit: sequence identification and comparison of key markers of
882 the koala (*Phascolarctos cinereus*) immune response. Australian Journal of Zoology
883 2014;62:195-199.
- 884 59. Morris KM, Matthew M, Waugh C, Ujvari B, Timms P, Polkinghorne A, Belov K. Identification,
885 characterisation and expression analysis of natural killer receptor genes in *Chlamydia*
886 *pecorum* infected koalas (*Phascolarctos cinereus*). BMC Genomics 2015;16:
- 887 60. Morris KM, O'Meally D, Zaw T, Song X, Gillett A, Molloy MP, Polkinghorne A, Belov K.
888 Characterisation of the immune compounds in koala milk using a combined transcriptomic
889 and proteomic approach. Scientific Reports 2016;6:e35011.
- 890 61. Lau Q, Griffith JE, Higgins DP. Identification of MHCII variants associated with chlamydial
891 disease in the koala (*Phascolarctos cinereus*). PeerJ 2014;2:443.
- 892 62. Robbins A, Hanger J, Jelocnik M, Quigley BL, Timms P. Koala immunogenetics and chlamydial
893 strain type are more directly involved in chlamydial disease progression in koalas from two
894 south east Queensland koala populations than koala retrovirus subtypes. Scientific Reports
895 2020;10:15013. 10.1038/s41598-020-72050-2.
- 896 63. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, Pham M, Glenn St
897 Hilaire B, Yao W, Stamenova E, et al. The Juicebox Assembly Tools module facilitates *de novo*
898 assembly of mammalian genomes with chromosome-length scaffolds for under \$1000.
899 bioRxiv 2018;254797. 10.1101/254797.

- 900 64. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MD, Machol I,
901 Lander ES, Aiden AP, Aiden EL. De novo assembly of the *Aedes aegypti* genome using Hi-C
902 yields chromosome-length scaffolds. *Science advances* 2017;356:92-95.
- 903 65. Peel E, Silver L, Brandies P, Hogg CJ, Belov K. A reference genome for the critically
904 endangered woylie, *Bettongia penicillata ogilbyi*. *Gigabyte*
905 2021;1:<https://doi.org/10.46471/gigabyte.35>.
- 906 66. Brandies PA, Tang S, Johnson RSP, Hogg CJ, Belov K. The first *Antechinus* reference genome
907 provides a resource for investigating the genetic basis of semelparity and age-related
908 neuropathologies. *Gigabyte* 2020;2020:0. 10.46471/gigabyte.7.
- 909 67. Peel E, Silver L, Brandies PA, Hayakawa T, Belov K, Hogg CJ. Genome assembly of the numbat
910 (*Myrmecobius fasciatus*), the only termitivorous marsupial. *Gigabyte*
911 2022;<https://doi.org/10.46471/gigabyte.47>.
- 912 68. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+:
913 architecture and applications. *BMC bioinformatics* 2009;10:
- 914 69. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical
915 significance estimation. *PLOS Computational Biology* 2008;4:e1000069.
- 916 70. Belov K, Lam MKP, Hellman L, Colgan DJ. Evolution of the major histocompatibility complex:
917 isolation of the class II beta cDNAs from two monotremes, the platypus and the short-
918 beaked echidna. *Immunogenetics* 2003;55:402-411.
- 919 71. Belov K, Hellman L. Immunoglobulin genetics of *Ornithorhynchus anatinus* (platypus) and
920 *Tachyglossus aculeatus* (short-beaked echidna). *Comparative Biochemistry and Physiology*
921 2003;136:811-819.
- 922 72. Wong ESW, Sanderson CE, Deakin JE, Whittington CM, Papenfuss AT, Belov K. Identification
923 of natural killer cell receptor clusters in the platypus genome reveals an expansion of C-type
924 lectin genes. *Immunogenetics* 2009;61:565-579.
- 925 73. Dohm JC, Tsend-Ayush E, Reinhardt R, Grutzner F, Himmelbauer H. Distribution and
926 pseudoautosomal localization of the major histocompatibility complex in monotremes.
927 *Genome Biology* 2007;8:R175-R175.116.
- 928 74. Papenfuss AT, Feng Z, Krasnec K, Deakin JE, Baker ML, Miller RD. Marsupials and
929 monotremes possess a novel family of MHC class I genes that is lost from the eutherian
930 lineage. *BMC Genomics* 2015;16:535.
- 931 75. Nowak MA, Parra ZE, Hellman L, Miller RD. The complexity of expressed kappa light chains in
932 egg-laying mammals. *Immunogenetics* 2004;56:555-563.
- 933 76. Johansson J, Aveskogh M, Munday BL, Hellman L. Heavy chain V region diversity in the duck-
934 billed platypus (*Ornithorhynchus anatinus*): long and highly variable complementary-
935 determining region 3 compensates for limited germline diversity. *The Journal of Immunology*
936 2002;168:5155-5162.
- 937 77. Johansson J, Salazar JN, Aveskogh M, Munday B, Miller RD, Hellman L. High variability in
938 complementarity-determining regions compensates for a low number of V gene families in
939 the λ light chain locus of the platypus. *European Journal of Immunology* 2005;35:3008-3019.
940 <https://doi.org/10.1002/eji.200425574>.
- 941 78. Parra ZE, Arnold T, Nowak MA, Hellman L, Miller RD. TCR gamma chain diversity in the spleen
942 of the duckbill platypus (*Ornithorhynchus anatinus*). *Developmental and Comparative*
943 *Immunology* 2006;30:699-671.
- 944 79. Parra ZE, Lillie M, Miller RD. A model for the evolution of the mammalian T-cell receptor
945 alpha/delta and mu loci based on evidence from the duckbill platypus. *Molecular Biology*
946 *and Evolution* 2012;29:3205-3214.
- 947 80. Wang X, Parra ZE, Miller RD. Platypus TCRmu provides insight into the origins and evolution
948 of a uniquely mammalian TCR locus. *The Journal of Immunology* 2011;187:5246-5254.
- 949 81. Wong ESW, Papenfuss AT, Miller RD, Belov K. Hatching time for monotreme immunology.
950 *Australian Journal of Zoology* 2006;57:185-198.

- 951 82. Belov K, Sanderson CE, Deakin JE, Wong ESW, Assange D, McColl KA, Gout A, Bono B, Barrow
952 AD, Speed TP, et al. Characterization of the opossum immune genome provides insight into
953 the evolution of the mammalian immune system. *Genome Research* 2007;17:982-991.
- 954 83. Lau Q, Jobbins SE, Belov K, Higgins DP. Characterisation of four major histocompatibility
955 complex class II genes of the koala (*Phascolarctos cinereus*). *Immunogenetics* 2013;65:37-46.
- 956 84. Jobbins SE, Sanderson CE, Griffith JE, Krockenberger MB, Belov K, Higgins DP. Diversity of
957 MHC class II *DAB1* in the koala (*Phascolarctos cinereus*). *Australian Journal of Zoology*
958 2012;60:1-9. <https://doi.org/10.1071/ZO12013>.
- 959 85. Cheng Y, Polkinghorne A, Gillett A, Jones EA, O'Meally D, Timms P, Belov K. Characterisation
960 of MHC class I genes in the koala. *Immunogenetics* 2018;70:125-133.
- 961 86. Matthew M, Beagley KW, Timms P, Polkinghorne A. preliminary characterisation of tumour
962 necrosis factor alpha and interleukin-10 responses to *Chlamydia pecorum* infection in the
963 koala (*Phascolarctos cinereus*). *PLOS one* 2013;8:e59958.
- 964 87. Matthew M, Pavasovic A, Prentis PJ, Beagley KW, Timms P, Polkinghorne A. Molecular
965 characterisation and expression analysis of Interferon gamma in response to natural
966 *Chlamydia* infection in the koala, *Phascolarctos cinereus*. *Gene* 2013;527:570-577.
- 967 88. Mathew M, Waugh C, Beagley KW, Timms P, Polkinghorne A. Interleukin 17A is an immune
968 marker for chlamydial disease severity and pathogenesis in the koala (*Phascolarctos*
969 *cinereus*). *Developmental & Comparative Immunology* 2014;46:423-429.
970 <http://dx.doi.org/10.1016/j.dci.2014.05.015>.
- 971 89. Maher IE, Griffith JE, Lau Q, Reeves T, Higgins DP. Expression profiles of the immune genes
972 CD4, CD8 beta, IFN gamma, IL-4, IL-6 and IL-10 in mitogen-stimulated koala lymphocytes
973 (*Phascolarctos cinereus*) by qRT-PCR. *PeerJ* 2014;2:
- 974 90. Miller RD. Those other mammals: The immunoglobulins and T cell receptors of marsupials
975 and monotremes. *Seminars in Immunology* 2010;22:3-9.
- 976 91. Siddle HV, Deakin JE, Coggill P, Whilming LG, Harrow J, Kaufman J, Beck S, Belov K. The
977 tammar wallaby major histocompatibility complex shows evidence of past genomic
978 instability. *BMC Genomics* 2011;12:421.
- 979 92. Cheng Y, Siddle HV, Beck S, Eldridge MDB, Belov K. High levels of genetic variation at MHC
980 class II DBB loci in the tammar wallaby (*Macropus eugenii*). *Immunogenetics* 2009;61:111-
981 118.
- 982 93. Zuccolotto P, Harrison GA, Deane EM. Cloning of marsupial T cell receptor alpha and beta
983 constant region cDNAs. *Immunology and Cell Biology* 2000;78:103-109.
- 984 94. Daly KA, Digby M, Lefevre C, Mailer S, Thomson P, Nicholas KR, Williamson P. Analysis of the
985 expression of immunoglobulins throughout lactation suggests two periods of immune
986 transfer in the tammar wallaby (*Macropus eugenii*). *Veterinary Immunology and*
987 *Immunopathology* 2007;120:187-200.
- 988 95. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C,
989 Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse
990 genomes. *Nucleic Acids Research* 2018;47:D766-D773. 10.1093/nar/gky955.
- 991 96. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang Y-C, Madugundu AK,
992 Pandey A, Salzberg SL. CHES: a new human gene catalog curated from thousands of large-
993 scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology*
994 2018;19:208. 10.1186/s13059-018-1590-2.
- 995 97. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E,
996 Duvergey H, Houles A, Paysan-Lafosse T, et al. IMGT®, the international ImmunoGeneTics
997 information system® 25 years on. *Nucleic Acids Research* 2015;43:D413-D422.
998 10.1093/nar/gku1056.
- 999 98. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. *Genome*
1000 *Biology* 2019;20:92. 10.1186/s13059-019-1715-2.

1001 99. Manual gene annotation by Havana
1002 [\[https://m.ensembl.org/info/genome/genebuild/manual_havana.html\]](https://m.ensembl.org/info/genome/genebuild/manual_havana.html)
1003 100. The NCBI Eukaryotic Genome Annotation Pipeline
1004 [\[https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/\]](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/)
1005 101. Gene annotation in Ensembl [\[https://m.ensembl.org/info/genome/genebuild/index.html\]](https://m.ensembl.org/info/genome/genebuild/index.html)
1006 102. Updating information on GenBank genome records
1007 [\[https://www.ncbi.nlm.nih.gov/genbank/wgs_update/\]](https://www.ncbi.nlm.nih.gov/genbank/wgs_update/)
1008 103. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid
1009 genome sequences. *Genome Research* 2017;27:757-767. 10.1101/gr.214874.116.
1010 104. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I,
1011 Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive Mapping of Long-Range Interactions
1012 Reveals Folding Principles of the Human Genome. *Science* 2009;326:289.
1013 10.1126/science.1181369.
1014 105. Luo J, Wei Y, Lyu M, Wu Z, Liu X, Luo H, Yan C. A comprehensive review of scaffolding
1015 methods in genome assembly. *Briefings in Bioinformatics* 2021;10.1093/bib/bbab033.
1016 106. Durand NC, Robinson JT, Shamim MD, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox
1017 provides a visualisation system for Hi-C contact maps with unlimited zoom. *Cell Systems*
1018 2016;3:99-101.
1019 107. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating
1020 Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational*
1021 *Biology* 2019;15:e1007273. 10.1371/journal.pcbi.1007273.
1022 108. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale
1023 scaffolding of de novo genome assemblies based on chromatin interactions. *Nature*
1024 *Biotechnology* 2013;31:1119-1125. 10.1038/nbt.2727.
1025 109. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C,
1026 Delaney J, Fehr A, et al. Resolving the full spectrum of human genome variation using
1027 Linked-Reads. *Genome Research* 2019;29:635-645. 10.1101/gr.234443.118.
1028 110. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nature Reviews*
1029 *Genetics* 2020;21:171-189. 10.1038/s41576-019-0180-9.
1030 111. Ott A, Schnable JC, Cheng-Ting Y, Wu L, Liu C, Heng-Cheng H, Dalgard CL, Sarkar S, Schnable
1031 PS. Linked read technology for assembling large complex and polyploid genomes. *BMC*
1032 *Genomics* 2018;19:<http://dx.doi.org/10.1186/s12864-018-5040-z>.
1033 112. Damas J, Corbo M, Lewin HA. Vertebrate Chromosome Evolution. *Annual Review of Animal*
1034 *Biosciences* 2021;9:1-27. 10.1146/annurev-animal-020518-114924.
1035 113. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and
1036 conservation. *Nature* 2020;587:240-245. 10.1038/s41586-020-2876-6.
1037 114. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into
1038 population history and trait architecture. *Nature Reviews Genetics* 2018;19:220-234.
1039 10.1038/nrg.2017.109.
1040 115. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor
1041 N, Grüning BA, et al. The Galaxy platform for accessible, reproducible and collaborative
1042 biomedical analyses: 2018 update. *Nucleic Acids Research* 2018;46:W537-W544.
1043 10.1093/nar/gky379.
1044 116. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3
1045 and convergent evolution of coiled-coil regions. *Nucleic Acids Research* 2013;41:e121-e121.
1046 10.1093/nar/gkt263.
1047 117. Hall T. BioEdit vol. v7.2.2, v7.2.2 edition: Ibis Biosciences; 2013.
1048 118. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar Gustavo A, Sonnhammer ELL, Tosatto
1049 SCE, Paladin L, Raj S, Richardson LJ, et al. Pfam: The protein families database in 2021.
1050 *Nucleic Acids Research* 2020;49:D412-D419. 10.1093/nar/gkaa913.

1051 119. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020.
1052 Nucleic Acids Research 2020;49:D458-D460. 10.1093/nar/gkaa937.

1053 120. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics
1054 Analysis across Computing Platforms. Molecular biology and evolution 2018;35:1547-1549.
1055 10.1093/molbev/msy096.

1056 121. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
1057 Bioinformatics 2014;30:2114-2120.

1058 122. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and
1059 genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 2019;37:907-915.
1060 10.1038/s41587-019-0201-4.

1061 123. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
1062 McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. GigaScience
1063 2021;10:giab008. 10.1093/gigascience/giab008.

1064 124. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly
1065 from long-read RNA-seq alignments with StringTie2. Genome Biology 2019;20:278.
1066 10.1186/s13059-019-1910-1.

1067 125. Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. Illuminating the dark
1068 side of the human transcriptome with long read transcript sequencing. BMC Genomics
1069 2020;21:751. 10.1186/s12864-020-07123-7.

1070 126. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, Gao G. CPC2: a fast and accurate
1071 coding potential calculator based on sequence intrinsic features. Nucleic Acids Research
1072 2017;45:W12-W16. 10.1093/nar/gkx428.

1073 127. Bryant DM, Johnson KM, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, Lee TJ, Kuo T,
1074 Davis FG, Baterna J, et al. A tissue-mapped axolotl de novo transcriptome enables
1075 identification of limb regeneration factors. Cell Reports 2017;18:762-776.

1076 128. Amazon Web Services Open Datasets Program - Australasian Genomes
1077 [\[https://registry.opendata.aws/australasian-genomics/\]](https://registry.opendata.aws/australasian-genomics/)

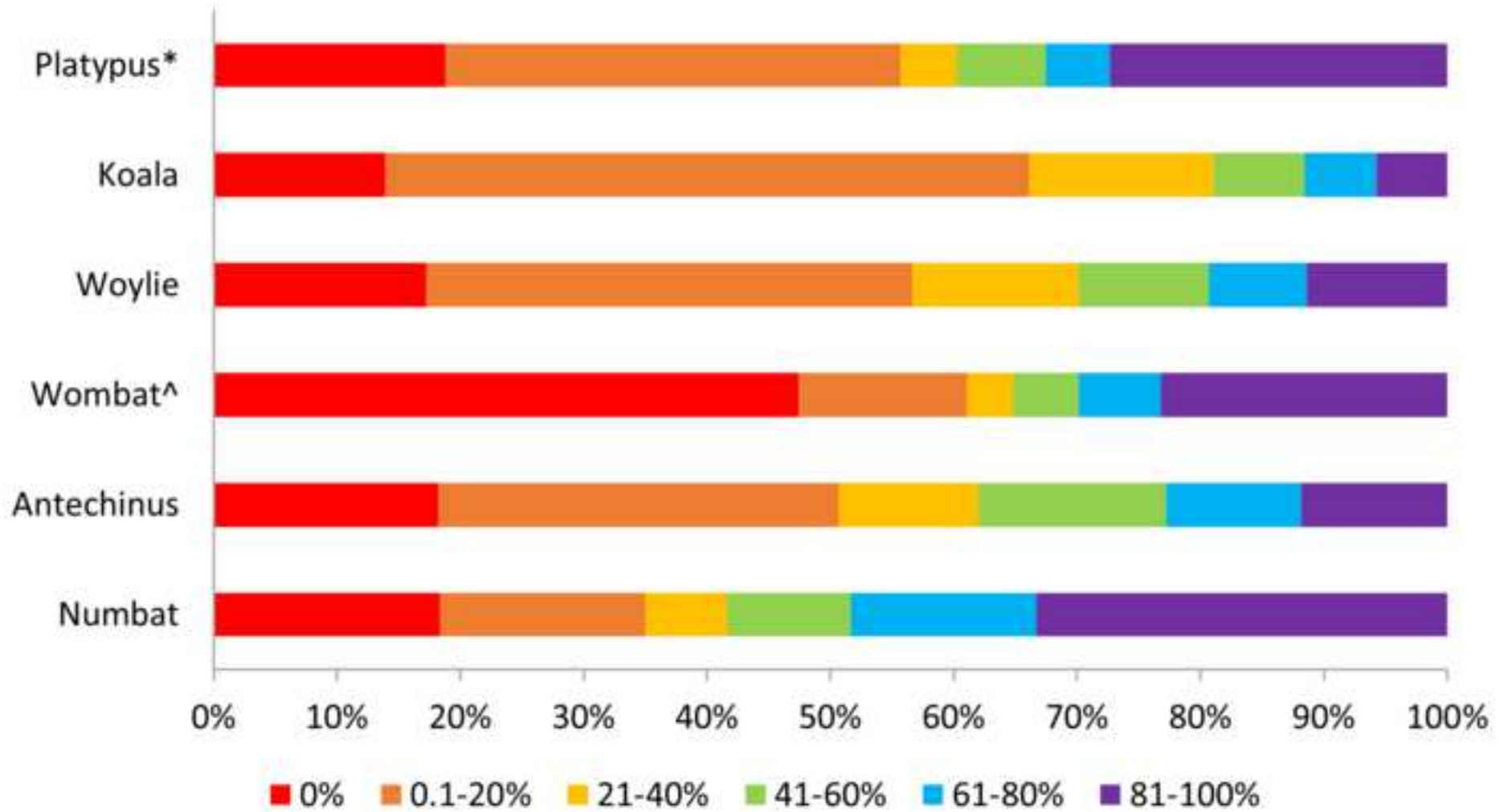
1078 129. Peel E, Silver L, Brandies PA, Hogg CJ, Belov K. Supporting data for "A reference genome for
1079 the critically endangered woylie, *Bettongia penicillata ogilbyi*". GigaDB
1080 2021;<http://dx.doi.org/10.5524/100951>.

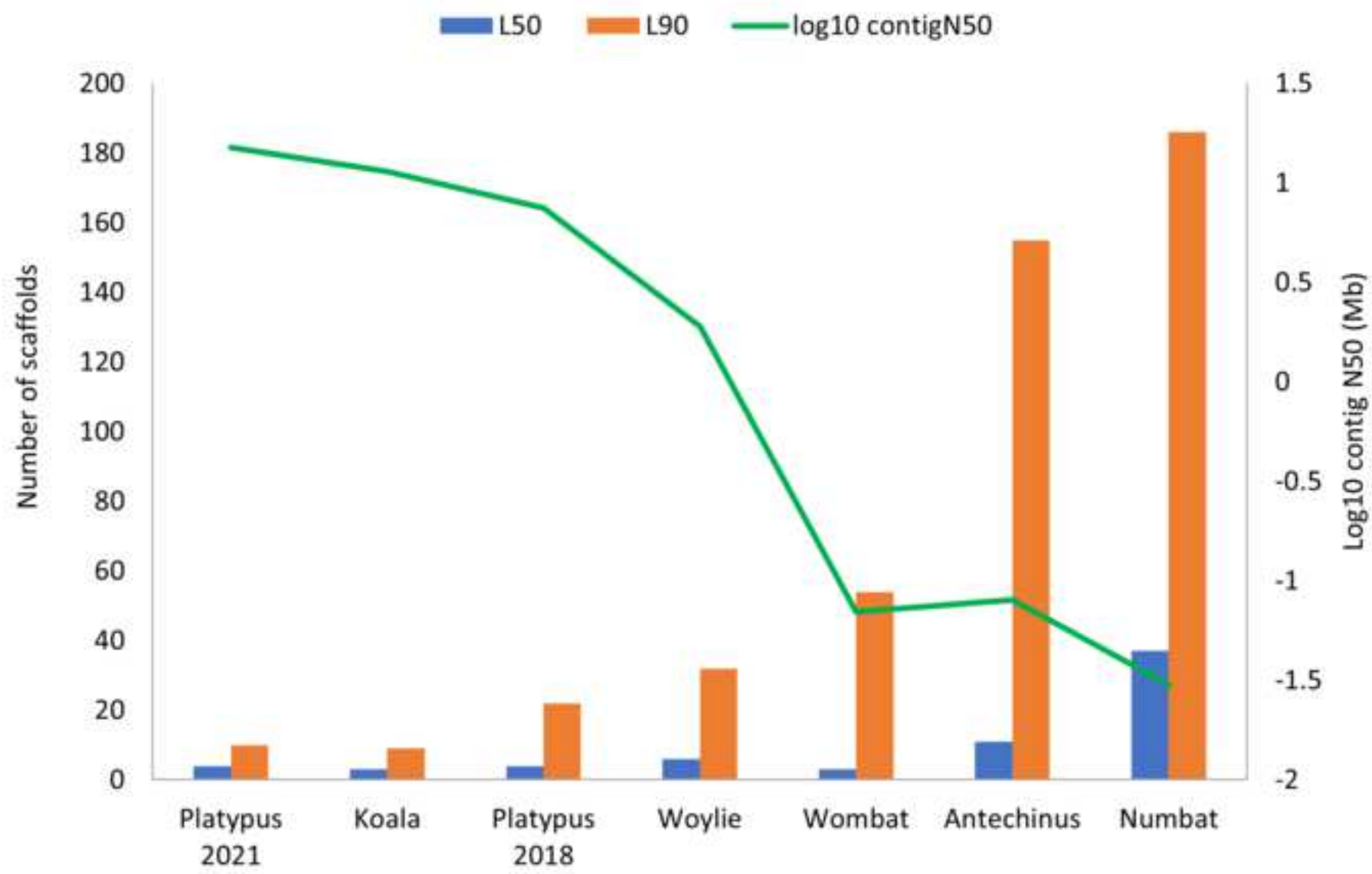
1081 130. Peel E, Silver L, Brandies PA, Hayakawa T, Belov K, Hogg CJ. Supporting data for "Genome
1082 assembly of the numbat (*Myrmecobius fasciatus*), the only termitivorous marsupial. GigaDB
1083 2022;<http://dx.doi.org/10.5524/100999>.

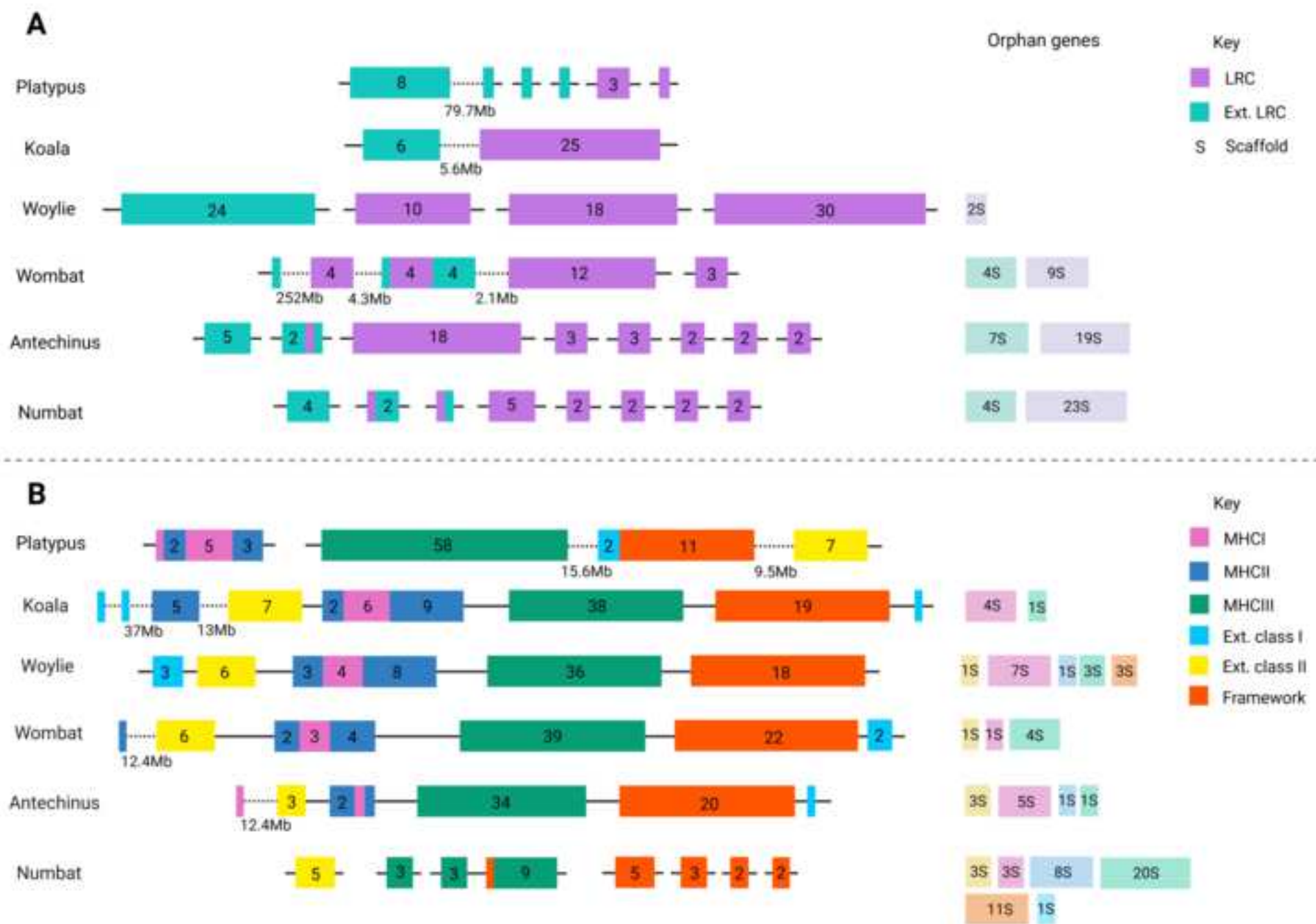
1084 131. DNAzoo [<https://www.dnazoo.org/assemblies>]

1085 132. Peel E, Silver L, Brandies PA, Zhu Y, Cheng Y, Hogg CJ, Belov K. Supporting data for "Best
1086 genome sequencing strategies for annotation of complex immune gene families in wildlife"
1087 GigaDB. 2022. <http://dx.doi.org/10.5524/102263>

1088









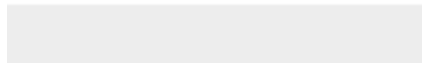


Click here to access/download
Supplementary Material
Additional file 1_amended.xlsx





Click here to access/download
Supplementary Material
Additional file 2_amended.docx



Response to reviewers

Reviewer #1:

Knowledge about immune genes is critical for species conservation programs. However, immune genes occur in large gene clusters that are difficult to assemble and annotate. This important and timely study uses a number of marsupial genomes and the platypus to assess which sequencing technologies enable complete reconstructions of immune gene clusters and which methods enable annotations of these immune genes.

I have the following comments.

Since Fgenesh++ and Maker produce automatic annotations, I wonder why not all 6 genomes were annotated with these two methods? This would allow a comparison between Fgenesh++ against Maker. Maybe it is possible to annotate at least a few genomes with both methods.

All six genomes were not annotated using Fgenesh++ and Maker as the authors wanted to utilise existing annotations available for 5 of the 6 genomes in our study (all except koala). The authors agree that re-annotating all genomes with both Fgenesh++ and Maker would enable a direct comparison between the two methods. However, determining the best automated annotation software for immune gene annotation was not the focus of this study, but rather the impact of assembly quality on immune gene annotation. A secondary aim of the paper was to investigate whether automated annotation software was able to accurately identify immune genes, compared to our manual annotations. While it is widely known within the field of wildlife immunogenetics that automated genome annotations fail to correctly characterise immune genes, to date there are no publications which quantitatively assess this observation.

The computation required to annotate all six genomes using both Fgenesh++ and MAKER was not feasible within the given three-month timeframe provided for changes to the manuscript. As such, the koala, wombat and 2021 platypus genomes have been annotated with Fgenesh++ which will enable investigation of how this popular annotation software performs for immune gene annotation within all genome assemblies of varying quality included in this study (woylie, antechinus and numbat were already annotated with Fgenesh++). The methods, results and figure 1 have been modified to reflect this. See lines 213-218, 261-273 of the results and below. Additional supplementary figures have been generated in response to the reviewer's comment. See supplementary figure 3, 4 and 5 in Additional file 2.

Table 1 has been modified to also include all genome annotations used in this study. This includes existing published annotations by NCBI, MAKER and Fgenesh++, as well as Fgenesh++ annotations conducted as part of this study.

Lines 213-218

"We assessed how well our manual immune gene annotation aligned with automated annotations by Fgenesh++ (2018 platypus, woylie, koala, antechinus, numbat and wombat), MAKER (wombat) and the NCBI pipeline (2021 platypus). Inclusion of the 2021 platypus NCBI and wombat MAKER annotations ensures that any differences in automated and manual immune gene annotation were not due to deficiencies within the Fgenesh++ annotation pipeline, as the woylie, antechinus and numbat genomes were all annotated with Fgenesh++ using the same parameters."

Lines 263-275

“This pattern of poor immune gene annotation was not an artefact of inherent differences between automated annotation pipelines amongst the six genomes (NCBI, MAKER and Fgenesh++) nor genome quality, as similar patterns were observed for Fgenesh++ annotations of the 2021 platypus and wombat genome generated as part of this study (Supplementary Figure 3, Supplementary Figure 4). Generally, the Fgenesh++ annotation resulted in fewer correctly annotated immune genes ($\geq 90\%$ overlap) compared to NCBI (2021 platypus) or MAKER (wombat) (Supplementary Figure 3). Although, the proportion of missing immune genes (0% overlap) was higher in the NCBI (2021 platypus) and MAKER (wombat) annotation than the Fgenesh++ annotation of both species genomes. As with NCBI and MAKER, Fgenesh++ poorly annotated TCR and IG families at the gene-level (Supplementary Figure 4) in the high-quality platypus and low-quality wombat. Correct annotations were somewhat recovered at the exon-level in both genomes (Supplementary Figure 5), although, the number of missing TCR and IG exons in the Fgenesh++ annotation was almost half that of NCBI and MAKER in platypus and wombat respectively.”

Direct assessments of assembly quality should ideally be done on different assemblies of the same species to rule out real differences between species. Would it be possible to include previous koala or platypus genome that was much more fragmented?

The authors agree that multiple versions of the same genome assembly would enable direct assessment of assembly quality on immune gene annotation. As such, the authors have annotated the latest 2021 version of the platypus genome assembly published by Zhou et al 2021 (NCBI ID GCA_004115215.4) and the previous 2018 version (GCA_002966995.1) with Fgenesh++. Platypus was selected as the species for this comparison over koala (the only other species in our study with multiple genome assemblies available) as the improvement in assembly metrics between the 2021 and 2018 platypus genome assemblies is more significant than the 2018 and 2020 koala genome assemblies. This is due to the addition of numerous data types to the 2021 platypus assembly since the 2018 version. Genome assembly metrics for the 2018 platypus genome have been added to Table 1. The results section “Relationship between genome quality and manual immune gene annotation” has been modified to include a comparison between the 2018 and 2021 platypus assemblies. Specifically, see lines 283-290 and below. Figure 2 has also been updated to include the 2018 platypus genome assembly.

Fgenesh++ was selected for automated annotation of the two platypus assemblies over other methods such as MAKER as this would enable direct comparison of Fgenesh++ performance across all genomes in this study.

Lines 285-292

“To rule out species-specific differences in our direct assessment of assembly quality on immune gene annotation, we annotated a previous version of the platypus genome from 2018 (GCA_002966995.1) with Fgenesh++ to enable comparison with our Fgenesh++ annotation of the 2021 platypus genome (GCA_004115215.4) also generated as part of this study. Compared to the 2021 assembly, the 2018 platypus assembly was more fragmented given the 6-fold increase in the number of contigs, 14-fold increase in the number of scaffolds, and associated 2-fold decrease in contig N50 and 4-fold decrease in scaffold N50 between the two assemblies. Despite these metrics, the 2018 platypus assembly is still highly contiguous as it was generated using long-read data.”

Figure 1 shows a useful of all immune genes. However, some genes like TLRs are actually easy to annotate as they are have a standard gene structure. Therefore, it would be informative to provide in this figure a breakdown of how well the different immune gene families are annotates, as the authors nicely did in table 2. This would inform on which immune genes are particularly difficult to annotate.

A breakdown of Figure 1 by immune family is now presented in Supplementary Figure 1 of Additional File 2. A breakdown of annotation at the exon-level by immune family has been added as Supplementary Figure 2. See lines 235-255 and below. Similar breakdown of this analysis by immune family have also been added for Fgenesh++ versus MAKER (wombat) or NCBI (2021 platypus) annotations of the platypus and wombat genome assemblies at the gene level for all seven families (Supplementary Figure 4), in addition to exon-level for TCR and IG families (Supplementary figure 5).

Lines 235-255

“A breakdown of this analysis by immune family revealed that marsupial- and monotreme-specific immune genes which are not orthologous to those in eutherians were generally poorly annotated, regardless of automated pipeline or genome quality (Supplementary Figure 1). This was particularly the case for TCR and IG gene families, with up to 88% of genes in these families incorrectly annotated by automated pipelines ($\leq 10\%$ overlap) amongst the six species (Table 2). This is likely due to highly duplicated variable gene segments that do not encode conventional exon-intron splice sites which may hinder annotation with automated pipelines. Poor gene annotations of TCR and IG families was somewhat recovered at the exon level, as some TCR and IG variable gene segments were annotated as exons by automated pipelines. Correct annotation ($\geq 90\%$ overlap) of the TCR family increased from 0-2% at the gene level to 2-15% at the exon level amongst the six genomes (Supplementary Figure 2). This improvement was even greater for the IG family, with an increase from 0-2% correct annotation at the gene level to 15-43% at the exon level amongst the six genomes (Supplementary Figure 2). Despite this, up to 67% of TCR and IG variable segments were still not annotated at the exon level (0% overlap) amongst the six genomes, highlighting the difficulty in automated annotation of these regions. Similarly, marsupial-specific gene expansions within the leukocyte receptor complex (LRC) and monotreme-specific gene expansions within the natural killer complex (NKC) family of NK receptors were also poorly annotated by automated pipelines (Supplementary Figure 1). As with TCR and IG families, correct annotation increased from the gene- (0-28% marsupial LRC, 31% platypus NKC) to exon-level (6-65% marsupial LRC, 79% platypus NKC) (Table 2, Supplementary Figure 2), likely due to the presence of variable numbers of duplicated immunoglobulin superfamily (IGSF) domains and C-type lectin (CLEC) domains within each LRC and NKC gene respectively.”

Figure 3B is not colorblind friendly.

Colours in Figure 3B have been amended according to the colourblind friendly palette outlined in Wong, B. Points of view: Color blindness. *Nat Methods* **8**, 441 (2011). <https://doi.org/10.1038/nmeth.1618>

Line 275: The discussion makes it clear that this is a scaffolding error and not a real inversion. This should be clarified here as well.

This has been clarified in the text, see lines 345-347 and below.

“This organisation is unusual amongst mammalian TCR and is likely a result of the HiC scaffolding error and not a true inversion.”

I fully agree with the value of the manual annotations. Therefore, it would be helpful to provide the manual annotations also as a gff3 or gtf file that provide the full exon structure. Additional file 2 only lists the start and end coordinates of genes with multiple exons. The assembly accession should also be listed.

Additional file 1 (previously Additional file 2) has been amended to include both the gene and exon coordinates for all immune genes across the 7 genome assemblies.

As a suggestion: A haplotype-resolved assembly of a marsupial is likely not yet available, but such an assembly would provide an opportunity to further investigate the influence of assembly quality and haplotype variation in immune genes.

The authors agree with the reviewer's comment. However, a haplotype-resolved assembly for marsupials will be challenging to generate given current recommendations include the use of trios to completely resolve paternal and maternal haplotypes. Samples from trios are incredibly difficult to obtain for wildlife such as marsupials given the opportunistic nature of most sample collection. This would be especially difficult for marsupials which are threatened or endangered, or are not currently housed in captivity.

Reviewer #2:

In this work, Peel and collaborators assess the accuracy of immune gene annotation in marsupial species by comparing the outcome of manually and automated annotation approaches. This allowed them to conclude that sequence data type and assembly quality determine the accuracy of gene annotation. I find the study interesting, although I have some general comments. I find that both the introduction and discussion sections would benefit from some re-structuring. Both sections are a bit long, with some repetitions. Also, the discussion section contains material from results. I would also appreciate more detailed figure legends.

The authors thank reviewer 2 for their comments. In light of no specific changes provided by reviewer 2, and changes already made to both the discussion and introduction for reviewer 1 and 3, we took no further action.

Reviewer #3:

In this manuscript, Peel et al examine the impact of assembly quality and sequencing/assembly method on the ability to annotate complex genes of the immune system, using a case study the five marsupial genomes and one monotreme genome of varying quality. While the conclusions the authors present are not particularly surprising given what we know about genome assembly, this manuscript does a nice job outlining the reasons why higher quality (in particular, long-read) assemblies are important to facilitate annotation of these critical genes, and exploring in depth the impact of various aspects of assembly quality. The authors present their results in a convincing and clear way, and this work provides a useful summary for the genomics community.

I do have some minor comments that I hope will help improve this work, listed below.

1. The conditional "in wildlife" is perhaps a little confusing in the title, as I believe the issues the authors raise should be widely relevant to vertebrate, or at least mammalian, genomes, and

"wildlife" is a term with varying colloquial definitions among the readership of Gigascience. Relatedly the discussion in the background section of the abstract, as well as the intro of the manuscript and some parts of the discussion, could probably focus on mammals generally, or even vertebrates, not wildlife specifically. It would also make sense to make the implicit vertebrate focus explicit.

The authors agree that the issues raised in our manuscript would be applicable to many mammalian or vertebrate genomes. However, genomics projects for non-model species such as wildlife generally work within constraints that are not always applicable to mammalian or vertebrate genomes more broadly. These include budget considerations, access to samples (remote locations, permits, CITES listing, threat status) and sample quantity (volume and tissue types available, sample quality (opportunistic sampling, non-invasive samples, sub-optimal preservation method, no access to liquid nitrogen or -80 freezer), amongst many others. All these factors influence the type of genome sequencing available to wildlife genomics projects, and hence resulting assembly quality. Mammals and many vertebrates more broadly, do not generally face these multitude of challenges when generating reference genomes. While the link between input sample, assembly quality and curation to generate a high-quality assembly has been established in wildlife (Rhie et al 2021), what has not been assessed is the impact of assembly quality on functionally important regions of the genome, such as immune genes. Our aim was to provide guidance for the wildlife genomics community, particularly those working on species impacted by disease, on how different genome sequencing strategies impact quality of immune gene annotations.

2. The introduction goes into extensive detail about the case study systems presented here - perhaps more detail than is really needed (e.g., lines 130 - 136 on DFTD and chlamydial vaccines). However, there is little background information about the specific immune gene families that are the focus of this work. The authors present a compelling argument for why studying these gene families is important, but some additional information to help guide readers who may not be expert in the specific immune families under discussion would be valuable. In particular, reminding readers why these genes in particular are such a challenge to annotate, with perhaps a brief overview of the six immune gene families that are the focus of the work.

Additional detail regarding the six immune gene families that are the focus of the manuscript, and why immune genes are challenging to annotate has been provided in the introduction at lines 66-101 and below for easy reference.

"The COVID-19 pandemic is one of many examples which highlight the ever-increasing importance of understanding wildlife immunity and disease to better understand and manage disease spill over [17]. In the case of wildlife threatened by disease, conservation questions are more challenging to answer and typically involve immunogenetic diversity which relies on accurate immune gene annotations. Immune genes in mammals can be classified into six major families based on their evolutionary history and function: T cell receptors (TCR), immunoglobulins (IG), major histocompatibility complex (MHC), natural killer (NK) receptors, toll-like receptors (TLR) and cytokines. Mammals utilise two antigen recognition systems: TCR and IG expressed by T lymphocytes and B lymphocytes respectively. TCR and IG are encoded in large clusters within the genome, each of which contain few constant sequences that define the receptor sub-type, and multiple highly duplicated variable segments that recognise and bind antigens. The number and sequence polymorphism of IG and TCR V segments varies significantly between mammalian species [18-20]. Another major family of immune genes is the major histocompatibility complex which contains three classes of genes (class I, II and III). MHC class I and II genes encode cell-surface receptors which bind and present self- and pathogen-derived antigens to T

lymphocytes, activating the adaptive immune response. Class I and II genes evolve via duplication and can be highly polymorphic, hence gene number differs between species [21, 22]. Natural killer (NK) cells directly kill virus-infected and cancerous cells and are an important component of innate immunity. Their activity is mediated via cell-surface receptors encoded by genes classified into two functionally similar but structurally dissimilar families; the leukocyte receptor complex (LRC) and natural killer complex (NKC). These families are encoded in separate clusters within the genome, and as they evolve via gene duplication, gene number varies significantly between species [23]. TLRs are membrane-spanning receptors expressed by immune and non-immune cells which bind pathogen-associated molecular patterns (PAMP), activating the innate and adaptive immune response. Compared to other immune genes, TLRs gene number and sequence is relatively conserved across mammals [24]. Lastly, cytokines are small proteins secreted by numerous cell types which direct the immune response. Cytokines can be classified into multiple families including interferons (IFN), tumour necrosis factors (TNF) and interleukins (IL), and gene content within each family varies between mammals [25].

Immune genes are some of the most polymorphic regions of the genome, owing to the need to generate diversity in response to ever-changing pathogenic pressures [26, 27]. Diversity within these 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 [26]. Generally, immune genes are encoded within repetitive clusters in the genome, especially highly duplicated families such as the MHC and NK receptors [28]. Given these factors, accurate assembly and annotation of genomic regions encoding immune genes can be challenging [29-31], especially in wildlife."

3. I would recommend ordering the species in Table 1, Table 2, Figure 1, Figure 2, and Figure 3 in a consistent order, perhaps from highest to lowest contig N50. This will help readers keep track of the key patterns.

Ordering of species and immune families in figures and tables (except for table 1) in the main manuscript and Additional file 2 is now consistent with the reviewer's suggestion. Species are presented in the order of platypus, koala, woylie, wombat, antechinus then numbat, and immune families are presented in the order of cytokines, TLR, MHC, NKC, LRC, IG and TCR.

4. The authors present a qualitative assessment of the kinds of genes where automated annotation fails in lines 202-212 and Fig 3. However a quantitative breakdown here would also I think be useful to the community, and should be easy to generate. One could simply list the fraction of manually annotated genes correctly recovered (and completely missed with <10% overlap) for each class in Table 2 for each species. This would also allow the authors to put some numbers alongside statements in this paragraph like "Most of these genes comprised... [line 210]"

≥90% and ≤10% overlap in genomic coordinates between manual and automated annotation of immune genes has been added for each species and immune family in table 2. A quantitative breakdown has been added to this section of the results. See lines 235-255 and below. The authors have also added additional detail regarding automated versus manual immune annotation at the exon-level for the TCR, IG and LRC families which were poorly annotated by automated pipelines at the gene-level.

Lines 235-255

“A breakdown of this analysis by immune family revealed that marsupial- and monotreme-specific immune genes which are not orthologous to those in eutherians were generally poorly annotated, regardless of automated pipeline or genome quality (Supplementary Figure 1). This was particularly the case for TCR and IG gene families, with up to 88% of genes in these families incorrectly annotated by automated pipelines ($\leq 10\%$ overlap) amongst the six species (Table 2). This is likely due to highly duplicated variable gene segments that don't encode conventional exon-intron splice sites which may hinder annotation with automated pipelines. Poor gene annotations of TCR and IG families was somewhat recovered at the exon level, as some TCR and IG variable gene segments were annotated as exons by automated pipelines. Correct annotation ($\geq 90\%$ overlap) of the TCR family increased from 0-2% at the gene level to 2-15% at the exon level amongst the six genomes (Supplementary Figure 2). This improvement was even greater for the IG family, with an increase from 0-2% correct annotation at the gene level to 15-43% at the exon level amongst the six genomes (Supplementary Figure 2). Despite this, up to 67% of TCR and IG variable segments were still not annotated at the exon level (0% overlap) amongst the six genomes, highlighting the difficulty in automated annotation of these regions. Similarly, marsupial-specific gene expansions within the leukocyte receptor complex (LRC) and monotreme-specific gene expansions within the natural killer complex (NKC) family of NK receptors were also poorly annotated by automated pipelines (Supplementary Figure 1). As with TCR and IG families, correct annotation increased from the gene- (0-28% marsupial LRC, 31% platypus NKC) to exon-level (6-65% marsupial LRC, 79% platypus NKC) (Table 2, Supplementary Figure 2), likely due to the presence of variable numbers of duplicated immunoglobulin superfamily (IGSF) domains and C-type lectin (CLEC) domains within each LRC and NKC gene respectively.”

5. I am not sure the statement (298-299): "that a kitchen sink approach, that uses long-read data combined with HiC technology, to generate a high-quality genome assembly is required to investigate immunity and disease in wildlife" is fully supported by the results the authors present. The annotation of the woylie genome, which as I understand it does not include any HiC scaffolding, seems to be as good or nearly as good as the two kitchen sink genomes. I would propose that the key conclusion is that long-read data specifically (with or without HiC) and high contig N50 (probably at least 1 Mb) is what is required for a successful manual annotation of these complex immune genes. This issue resurfaces in the discussion section, where again the point that HiC + Illumina is not sufficient is quite clear, but the converse does not seem well supported: long-read data in the absence of HiC does just fine.

The authors agree that this statement could be improved. Our results do support the reviewer's suggestion that assemblies based on long-read data, with or without scaffolding technology, are required for successful immune gene annotation. However, as outlined in the results section lines 298-320, immune gene families in the kitchen sink genomes represented by the 2021 platypus and koala assemblies were more intact than the woylie or 2018 platypus assembly (results presented in lines 368-380), both of which are based on long-read data. This was especially true for highly duplicated families such as the MHC, LRC NK receptors and TCR. The opening statement of the discussion has been modified to reflect the reviewer's suggestion, see lines 382-390 and below.

“By manually annotating immune genes in five marsupial genomes and two versions of the platypus genome, all varying qualities, we have confirmed that genome quality is directly linked to our ability to annotate complex immune gene families. Without long reads and scaffolding technologies, immune genes are scattered across many individual scaffolds and gene family organisation and evolution cannot be elucidated. We conclude that long-read data, with or without HiC technology, to generate a high-quality genome assembly with a contig N50 of at least 1MB is required to investigate immunity

and disease in wildlife. However, a kitchen sink approach to genome sequencing and assembly will enable complete reconstruction of complex and duplicated families such as MHC, TCR and LRC NK receptors as in the platypus 2021 and koala genomes.”

6. The discussion of the limits of automated annotation is very important, but I found this section (starting on line 311) a little muddled. One key clarification is that it would probably be useful to separately discuss TCR and IG variable segments from all other immune genes. As the authors mention, automated analysis is not expected to successfully recover these variable regions, and it would probably be more useful to readers to get a sense of how automated analysis and RNA-seq alignment performs excluding these elements, in addition to the discussion on lines 329-337 of the specific challenges of variable regions.

This section of the discussion has been revised in response to the reviewer’s comment and additional detail added. Automated annotation and RNAseq support for immune genes other than TCR and IG is now discussed in lines 408-437, while TCR and IG are solely discussed in lines 424-434. See amended text below.

“Aside from TCR and IG, the majority of immune genes incorrectly annotated or missing from the automated annotations were divergent genes not orthologous to those in eutherian mammals, such as MHC, marsupial-specific gene expansions within the LRC and monotreme-specific gene expansions within the NKC. Given their divergence, these genes often have low or no BLAST homology to nucleotide or protein databases. Gene models generated by automated annotation software are hypotheses based on supporting evidence such as RNAseq data and homology to nucleotide and protein databases. While immune transcripts were identified in the transcriptomes from these species, RNAseq data only supported gene models for a low proportion of MHC, LRC and NKC genes. RNAseq data only supported 8-16% of LRC gene predictions and 16-37% of MHC gene predictions amongst the four marsupial genome annotations which used RNAseq data as gene model evidence (koala, woylie, antechinus and numbat). Similarly, around 60% of NKC genes within the platypus genomes were supported by RNAseq data. Overall, RNAseq data did not provide enough evidence to support gene models for ~20% of immune genes within the genome. Some immune genes may not have been expressed in the tissue sequenced, were expressed at low levels, or were fragmented. For human and mouse, comprehensive and curated gene sets such as GENCODE and RefSeq are available to guide gene model predictions, comprising data from more than 10,000 RNA experiments and decades of dedicated work in this field [95, 96]. Given time, budget and sample constraints for wildlife, these curated gene sets are not available, hence RNAseq evidence is incomplete resulting in deficient gene models by automated annotation software.

It is not surprising that TCR and IG V segments were poorly or not annotated by all automated pipelines used to annotate the genomes in this study. These genes are notoriously difficult to characterise and are manually annotated in the human and mouse genome on Ensembl using the International Immunogenetics Information System (IMGT) database [38, 97]. Alignment of mature IG and TCR sequences from RNAseq data to the genome results in poor automated annotation, as V segments utilize different sequence signal splice sites to introns, which are not recognized by the open reading frame prediction algorithms. Indeed, RNAseq evidence only supported 7% to 18% of TCR V segment and 0% to 6.9% of IG V segment gene predictions by automated pipelines amongst the four marsupial and platypus genomes. V sequences from three marsupials and two monotremes are available in IMGT, however as non-model species, they are not included in the scope for manual annotation by Ensembl or NCBI, so these important functional features are not annotated.”

7. Regarding "it is not a requirement for manual changes to annotations to be tracked between genome versions" on line 353, I am not sure this is so simple. Even lifting over the old manual curation to new assembly coordinates probably needs itself to be manually verified before one can be confident that the new model is correct. But I do not think this would mean the information is lost, as I believe NCBI and Ensembl both maintain old annotations and assembly versions.

The authors agree that this statement was vague and so has been removed from the manuscript. While NCBI and Ensembl maintain old annotations and assembly versions, our argument still stands as there is currently limited scope to include manual gene annotations of the scale presented in our manuscript alongside existing automated annotations from these databases.

8. Given that 10x linked reads are no longer available for genome assembly, the extensive discussion of their uses and limitations on lines 431-457 could probably be condensed considerably. This section of the discussion has been condensed, see lines 531—552 and text below. However, the authors feel discussing the limitations of 10x genomes for immune gene annotation is still warranted to make use of existing 10x assemblies, particularly for species where additional genome sequencing is unlikely due to sample or budget constraints.

"10x Chromium linked-read sequencing was insufficient to accurately re-assemble immune gene clusters in our study (Figure 4C). While this technology is no longer available for genome sequencing, acknowledging the limitations of this technology for immune gene annotation remains valid in order to make use of existing 10x genomes. Complete marsupial immune gene clusters can span hundreds of kilobases to megabases, as shown by annotation of the complete MHC, NK receptor and TCR regions in the koala (Additional file 2). DNA molecules input to 10x library preparation were on average 74 kbp and 23 kbp in antechinus and numbat respectively. This molecule size only spanned smaller immune clusters in the antechinus, such as the 70 kbp TRG locus, but was insufficient to span any cluster in the numbat. Interestingly, the antechinus MHC cluster appears to be intact (Figure 3), however manual annotation revealed multiple genes were "missing" within the scaffold and instead were located on individual short scaffolds. Regardless of input DNA molecule length, 10x libraries are still subject to the limitations of short-read sequencing regarding assembly of complex sequences. Antechinus and numbat 10x libraries were sequenced as short ~150 bp reads, hence while reads can be assigned back to the corresponding input DNA molecule, no single read spans the molecule length. Gaps between the reads make de novo assembly of repetitive and complex immune sequences difficult, often resulting in termination of contig extension and gene fragments scattered across short scaffolds [109-111]. These gene fragments can be misinterpreted as pseudogenes owing to loss of up/downstream coding regions (Figure 4C). For example, antechinus and numbat NK LRC genes share up to 97% and 98% amino acid sequence identity amongst the genes identified in each species respectively. The LRC should be encoded within a single cluster, as in the koala genome (Figure 3). Instead, the antechinus and numbat LRC clusters are fragmented across 33 and 34 scaffolds respectively."



THE UNIVERSITY OF
SYDNEY

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

11th August 2022

Dr Scott Edmunds
Chief Editor
GigaScience

Dear Dr Edmunds,

Please find attached our revised manuscript "*Best genome sequencing strategies for annotation of complex immune gene families in wildlife*" which we are re-submitting as a research article for publication in GigaScience. The text of the manuscript totals 7882 words, with four figures, two tables and two additional files. We would like to thank the reviewers for their valuable contributions. We have undertaken a re-analysis of the genomes using the same automated software, Fgenesh++, as recommended by reviewer 1, in addition to our previous work.

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. Due to our decade of research in wildlife immunogenetics we are increasingly asked the minimum genome quality required to effectively annotate immune genes which underpin wildlife disease investigations. In this manuscript we aimed 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 now obsolete 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.

We feel the manuscript is now improved and will appeal to researchers involved in sequencing, assembly, annotation and translation of genomics data. 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