

RESEARCH

Open Access



# The swan genome and transcriptome, it is not all black and white

Anjana C. Karawita<sup>1,2</sup>, Yuanyuan Cheng<sup>3</sup>, Keng Yih Chew<sup>1</sup>, Arjun Challagulla<sup>2</sup>, Robert Kraus<sup>4,5</sup>, Ralf C. Mueller<sup>4,5</sup>, Marcus Z. W. Tong<sup>1</sup>, Katina D. Hulme<sup>1</sup>, Helle Bielefeldt-Ohmann<sup>1</sup>, Lauren E. Steele<sup>1</sup>, Melanie Wu<sup>1</sup>, Julian Sng<sup>1</sup>, Ellesandra Noye<sup>1</sup>, Timothy J. Bruxner<sup>6</sup>, Gough G. Au<sup>2</sup>, Suzanne Lowther<sup>2</sup>, Julie Blommaert<sup>7,8</sup>, Alexander Suh<sup>7,9</sup>, Alexander J. McCauley<sup>2</sup>, Parwinder Kaur<sup>10</sup>, Olga Dudchenko<sup>11,12</sup>, Erez Aiden<sup>10,11,12,13,14</sup>, Olivier Fedrigo<sup>15</sup>, Giulio Formenti<sup>15</sup>, Jacquelyn Mountcastle<sup>15</sup>, William Chow<sup>16</sup>, Fergal J. Martin<sup>17</sup>, Denye N. Ogeh<sup>17</sup>, Françoise Thiaud-Nissen<sup>18</sup>, Kerstin Howe<sup>16</sup>, Alan Tracey<sup>16</sup>, Jacqueline Smith<sup>19</sup>, Richard I. Kuo<sup>19</sup>, Marilyn B. Renfree<sup>20</sup>, Takashi Kimura<sup>21</sup>, Yoshihiro Sakoda<sup>21</sup>, Mathew McDougall<sup>22</sup>, Hamish G. Spencer<sup>23</sup>, Michael Pyne<sup>24</sup>, Conny Tol<sup>25</sup>, Jonas Waldenström<sup>25</sup>, Erich D. Jarvis<sup>15</sup>, Michelle L. Baker<sup>2</sup>, David W. Burt<sup>1</sup> and Kirsty R. Short<sup>1\*</sup> 

\*Correspondence:  
k.short@uq.edu.au

<sup>1</sup> School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia  
Full list of author information is available at the end of the article

## Abstract

**Background:** The Australian black swan (*Cygnus atratus*) is an iconic species with contrasting plumage to that of the closely related northern hemisphere white swans. The relative geographic isolation of the black swan may have resulted in a limited immune repertoire and increased susceptibility to infectious diseases, notably infectious diseases from which Australia has been largely shielded. Unlike mallard ducks and the mute swan (*Cygnus olor*), the black swan is extremely sensitive to highly pathogenic avian influenza. Understanding this susceptibility has been impaired by the absence of any available swan genome and transcriptome information.

**Results:** Here, we generate the first chromosome-length black and mute swan genomes annotated with transcriptome data, all using long-read based pipelines generated for vertebrate species. We use these genomes and transcriptomes to show that unlike other wild waterfowl, black swans lack an expanded immune gene repertoire, lack a key viral pattern-recognition receptor in endothelial cells and mount a poorly controlled inflammatory response to highly pathogenic avian influenza. We also implicate genetic differences in *SLC45A2* gene in the iconic plumage of the black swan.

**Conclusion:** Together, these data suggest that the immune system of the black swan is such that should any avian viral infection become established in its native habitat, the black swan would be in a significant peril.

**Keywords:** Genomes, Virology, Black swan



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

The distinctive black plumage of the native Australian black swan (*Cygnus atratus*) is in stark contrast to the white swans that are native to Europe and North America. This unique feature has resulted in the black swan playing an important role in Australian heraldry and culture. The limited native geographic range (Australia) and relative isolation of the black swan has direct consequences for its immune repertoire and susceptibility to infectious disease common to other parts of the world. Specifically, geographic isolation can result in founder effects and reduced immune diversity as a result of limited pathogen challenge [1].

The native Australian black swan has a remarkably distinct response to infection by highly pathogenic avian influenza (HPAI) virus compared to the closely related white swans (e.g., the mute swan; *Cygnus olor*) and other waterfowl [2, 3]. Unlike mallard ducks and mute swans, the black swan is extremely sensitive to HPAI, succumbing to the disease within 2 to 3 days post-infection. This disease pathogenesis mirrors that of infected chickens, viewed as the most susceptible species to HPAI [3]. One of the striking features common to both black swans and chickens is that HPAI viruses preferentially infect endothelial cells, which may contribute to the disease severity in these two species [3]. These experimental studies are consistent with reports of natural infections, which suggest that captive black swans quickly succumb to HPAI while co-housed mute swans survive the infection [4].

Comparative genomics has played an important role in understanding species-dependent differences in HPAI pathogenesis [5] while also revealing the unique immune systems of many native Australian fauna [6]. However, comparative genomics is contingent upon the availability of high-quality species-specific genomes and transcriptomes.

Here, we generate the first black and mute swan reference genomes and transcriptomes, including the transcriptional response of primary black swan endothelial cells to HPAI. These data show that the black swan has numerous unique characteristics including (i) lack of an expanded immune gene repertoire, (ii) undetectable *Toll-like Receptor (TLR) 7* gene expression in infected endothelial cells, and (iii) a dysregulated pro-inflammatory response to viral infection that is likely to leave the species highly susceptible to viral infections such as HPAI. It is also likely that genetic differences in melanin production contribute to the distinctive black plumage of the black swan.

## Results

### Genome landscape

The chromosome-length reference genomes for black and mute swan were constructed according to a Pacbio continuous long-read-based (CLR) DNA Zoo pipeline and Vertebrate Genomes Pipeline 1.5 [7], respectively (Additional file 1: Figures S1 and S2). This included scaffolding contigs with Hi-C (Hi-C contact maps; Additional file 1: Figure S3) and Bionano maps. Curations of the assemblies identified scaffolds representing 34 autosomes plus the Z sex chromosome and were named according to the descending order of the physical size. Others were nominated as unplaced scaffolds due to their lack of resolution to be placed as chromosomes. We assigned final scaffolds to 35 chromosomes and 36 chromosomes for black swan and mute swans

respectively (including the sex chromosome), based on the physical size. The W chromosome was not assigned a scaffold in black swan as the DNA was derived from a male. The expected diploid number of chromosomes in the mute swan is 80 [8]. The black swan genome was sequenced using 90x PacBio CLR coverage, generating a 1.12 Gb reference assembly. The mute swan genome was sequenced using 60x PacBio CLR coverage, generating a 1.13 Gb reference assembly. The details of both genomes, including a comparison to the latest VGP (Vertebrate Genome Project) chicken (*Gallus gallus*) and mallard duck (*Anas platyrhynchos*) genomes, are shown in Table 1. For each genome, we estimated the heterozygosity, QV, and completeness values using Merqury based on the highly accurate Illumina® short reads (Additional file 2: Table S1 and Additional file 3: Online Methods).

The total classified repeat content of the genome was 10.56% for the black swan, with 1.71% unclassified repeats, and 10.76% for the mute swan genome, with 1.55% unclassified. This is lower than the repeat content recorded in the chicken (15%) and the mallard duck (17%) [9, 10].

The black and mute swan genomes were annotated with RNASeq and IsoSeq transcriptome data, homology-based alignments with other species and with bioinformatically inferred gene predictions, according to the methods listed in Additional file 1: Figure S4.

The completeness of the black and mute swan genomes was assessed using the Core Eukaryotic Genes Mapping (CEGMA) and Benchmark Universal Single Copy Orthologues (BUSCO) analyses and compared to the chicken and mallard duck genome (Additional file 4: Table S2, Additional file 5: Table S3). Notably, the black swan genome had the highest complete BUSCOs (8093), followed by the chicken (8054) and the mute swan (8010) genomes. While the chicken genome had the highest complete core-eukaryotic gene content (224), this was only marginally higher than that of the mute (221) and black swan (219) genomes.

One-to-one alignment between the black and mute swan genomes showed 98.35% average nucleotide identity between the 34 autosomes of the black and mute swans (Additional file 1: Figure S5). The Z chromosome of the black and mute swan had several large (>1kb) structural variants (Additional file 1: Figure S6), but were otherwise largely consistent. These structural differences in the Z chromosome may be associated with the speciation of the black and mute swan from their last common ancestor (approximately 6.1 million years ago) [11].

Inversions have been associated with the differential susceptibility of chickens and ducks to avian influenza [12]. We found no substantial inversions between the black and mute swan. However, given that both ducks and swans are *Anseriformes*, we compared structural genomic differences between the susceptible black swan and mute swan and relatively resistant duck, using the ostrich as an outgroup. We investigated genes in the inverted genomic regions present in the duck but absent in the swans on chromosome 1 (Additional file 1: Figure S7). Strikingly, we found that 53 inverted genes (out of 1758 total genes) in chromosome 1 of the duck genome were mapped to immune system processes (Additional file 6: Table S4). However, given the absence of substantial structural variants between the black and mute swan, it is likely that any immunological consequences of these structural variants would be present in multiple swan species.

**Table 1** Comparison of genomes between black and mute swans (generated herein) and published high-quality avian genomes generated by the vertebrate genome project (VGP).

Species	Assembly size (estimated size) (Gb)	G + C content (%) Macro-chromosome <sup>a</sup>	G + C content (%) Micro-chromosome <sup>b</sup>	Contig N50 (Mb)	Contig L50	Number of contigs	Scaffold N50 (Mb)	Scaffold L50	No. scaffolds	ISO-seq coverage <sup>c</sup>	Coding genes	Non-coding genes	Pseudogenes	Ref
Black swan	1.12	47	52	29	12	697	84.6	4	567	68.87% of the genome	15478	1792	104	This study
Mute swan	1.13	47	53	8.3	39	621	81.5	4	170	54.36% of the genome	14791	2720	126	This study
Chicken	1.09	40	48	18.8	18	676	90.9	4	213	NA	16878	7166	312	bGalGal1.mat. broiler;GRCg7b (GCF_016699485.2)
Duck	1.18	40.72	46.51	5.7	57	1660	76.3	5	755	NA	16618	9285	49	ZJU1.0 (GCF_015476345.1)

<sup>a</sup> Macro-chromosomes > 10 Mbps

<sup>b</sup> Micro-chromosomes < 10 Mbps

<sup>c</sup> Of the largest 34 chromosomes

The black and mute swan genomes were then annotated according to the methods listed in Additional file 1: Figure S4. Sixteen thousand two hundred four (16,204) and 15,789 gene models were obtained through Evidence Modeler as the final gene models in the black swan and the mute swan, respectively. Protein alignment against the UniProt/Swiss-Prot database was used to infer 15,478 gene model names for the black swan and 14,791 gene model names for the mute swan.

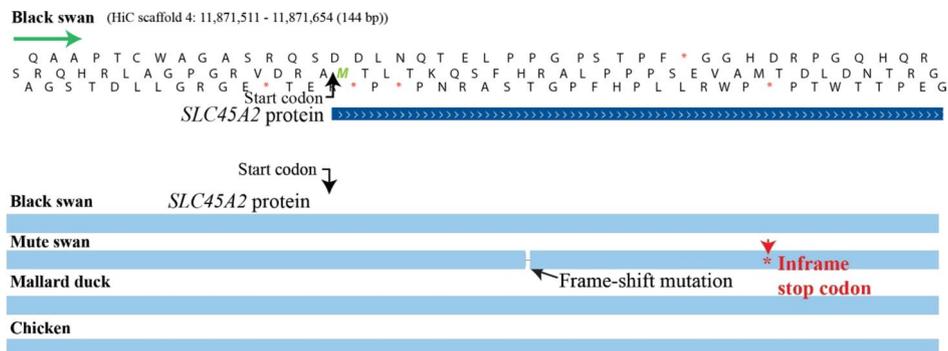
Importantly, it is impossible to infer if these differences represent true biological differences in the number of annotated genes or simply reflect the higher quality of the black swan genome.

**Mutations in *SLC45A2* may account for differences in swan plumage**

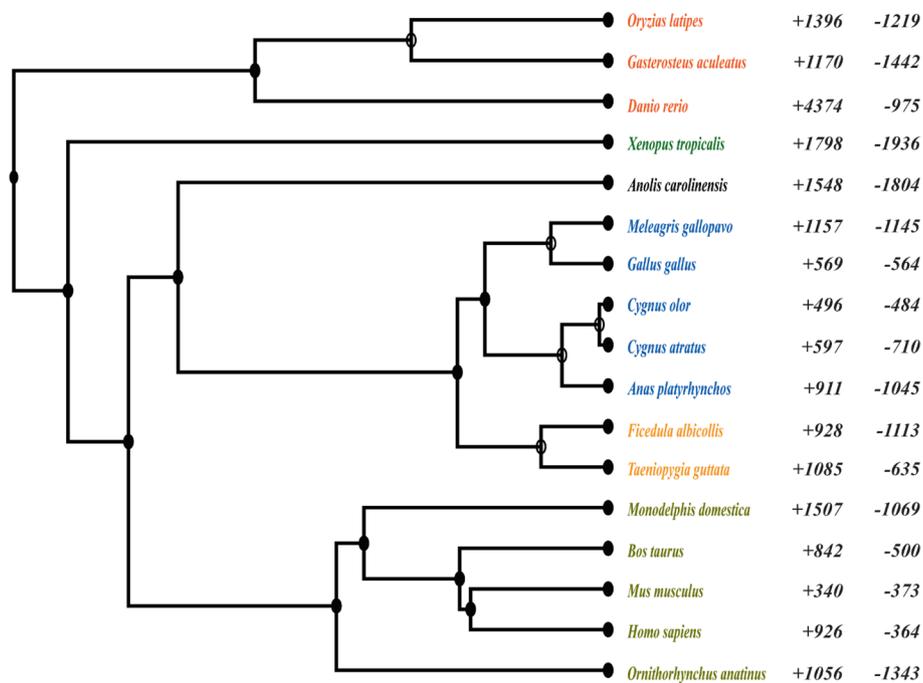
One of the most remarkable features of the black swan is its distinct plumage colour. To determine if the highly annotated genomes presented herein could offer insight into the iconic plumage, we examined genes (Additional file 7: Table S5) known to be associated in avian plumage color [13, 14]. We observed that *SLC45A2* in the mute swan had a nucleotide deletion in the open reading frame instigating a frame-shift mutation and an in-frame early stop codon (Fig. 1). Multiple non-homologous nucleotides were detected in the chicken and the duck *SLC45A2* relative to that of the black swan (Fig. 1). *SLC45A2* encodes a membrane-associated transport protein which regulates the tyrosinase activity and the melanin content of melanosomes [15]. The knockdown of this gene causes reduced tyrosinase activity and low melanin content in human melanoma cell lines [16]. These results suggest that this deletion in *SLC45A2* is a candidate genetic change that could be responsible for the white plumage in white swans in the genus *Cygnus*.

**Immune gene families are expanded in the mallard duck and mute swan genomes**

To understand whether the black swan (which has been isolated in Australia) has an altered immune gene repertoire compared to its close relatives, evolutionary gene gain and loss were determined (Fig. 2) (*p-value* < 0.05). The black swan genome was estimated to be contractive, indicating that the total gene gain was less than the gene loss



**Fig. 1** Position of the frame-shift mutation in the mute swan *SLC45A2* compared to the black swan genome. The figure was created with whole-genome HAL alignment produced from CACTUS. The alignment was visualized with the University of California, Santa Cruz (UCSC) Genome Assembly in a Box (GBiB). The predicted mute swan gene *SLC45A2* was aligned using the BLAT alignment algorithm.



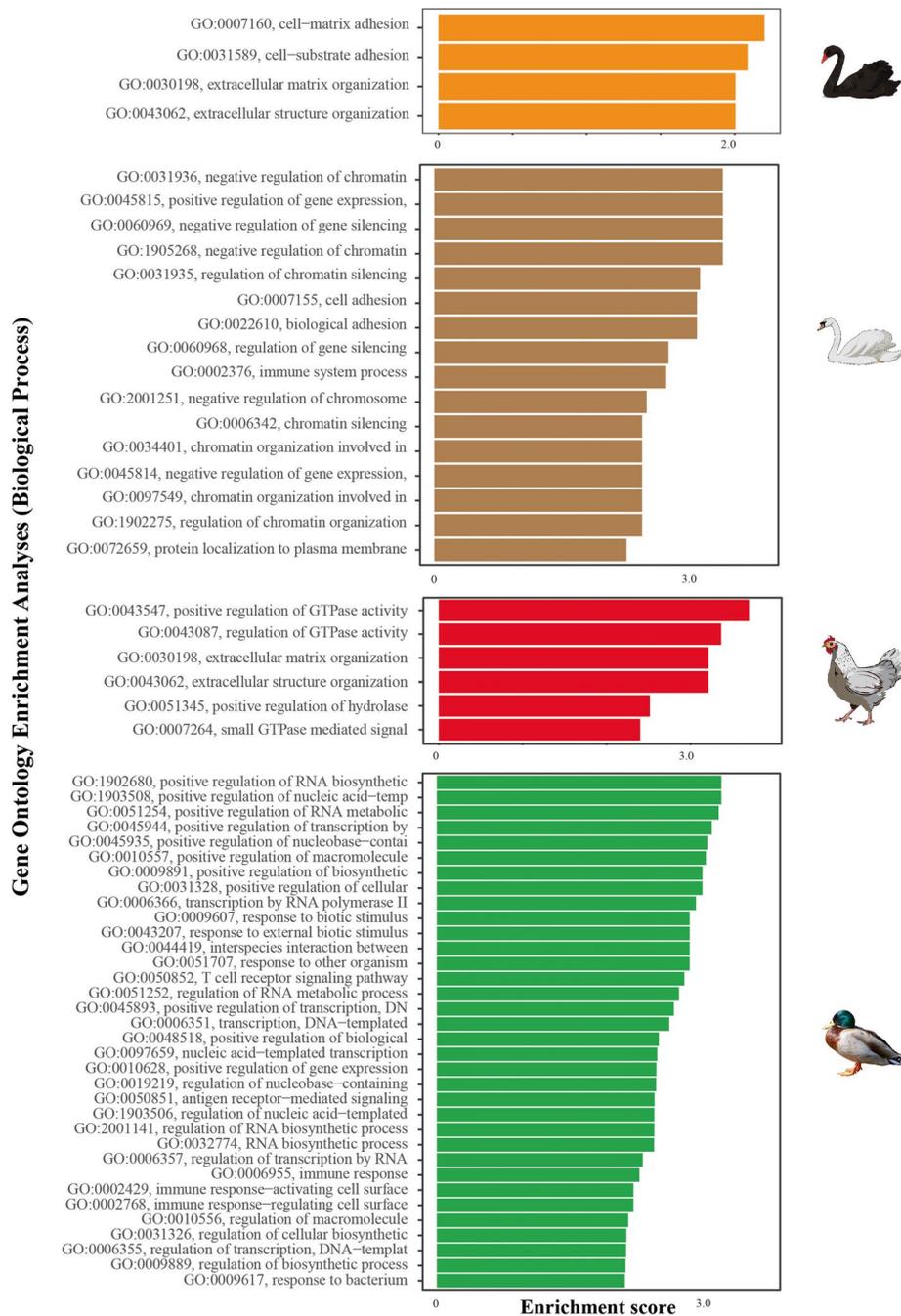
**Fig. 2** Gene losses and gains across vertebrate species tree. Data are shown for 17 vertebrate species, three teleost (red), an amphibian (green), six reptilians (including birds) (black, blue, and orange), and five mammals (yellow). The numbers of gene family gains (+) and losses (−) are given to the right of the taxa ( $p$ -value < 0.05) compared to the taxon’s last common ancestor. The rate of gene birth and death for clades derived from the most recent common ancestor (MRCA) for *Gallianseriformes* is 0.0016 (per gene per million years)

from the last common ancestor. The biological function of expanded genes in the black swan, chicken, mute swan, and duck was then investigated.

Strikingly, gene families related to immune system processes (e.g., GO0002376) were only expanded in the mallard duck and the mute swan genomes (Fig. 3). In contrast, in chickens, expanded gene families were associated with regulation of GTPase activity, extracellular matrix and structure organization, while the over-represented functional terms for expanded black swan gene families included cell-matrix adhesion, cell-substrate adhesion, extracellular matrix organization and extracellular structure organization. To specifically compare the immune gene repertoire of black and mute swans, we used human and mouse immune genes to identify immune gene families in *Cygnus* species. Thirty-nine immune-related gene families of the black swan were contractive compared to the mute swan (Additional file 8: Table S6). The PANTHER pathways related to these genes included apoptosis signaling, cadherin signaling, general transcription by RNA polymerase, gonadotropin-releasing hormone receptor, inflammation mediated by chemokine and cytokine signaling, interleukin signaling, TGF-beta signaling, and Wnt signaling pathways.

### Black and mute swan classical Major Histocompatibility Complex loci are homologous

Major Histocompatibility Complex (MHC) diversity is altered in some avian species [17], which may affect susceptibility to disease [18]. We therefore compared MHC loci between black and mute swans. Two MHC class I and MHC class II loci were identified



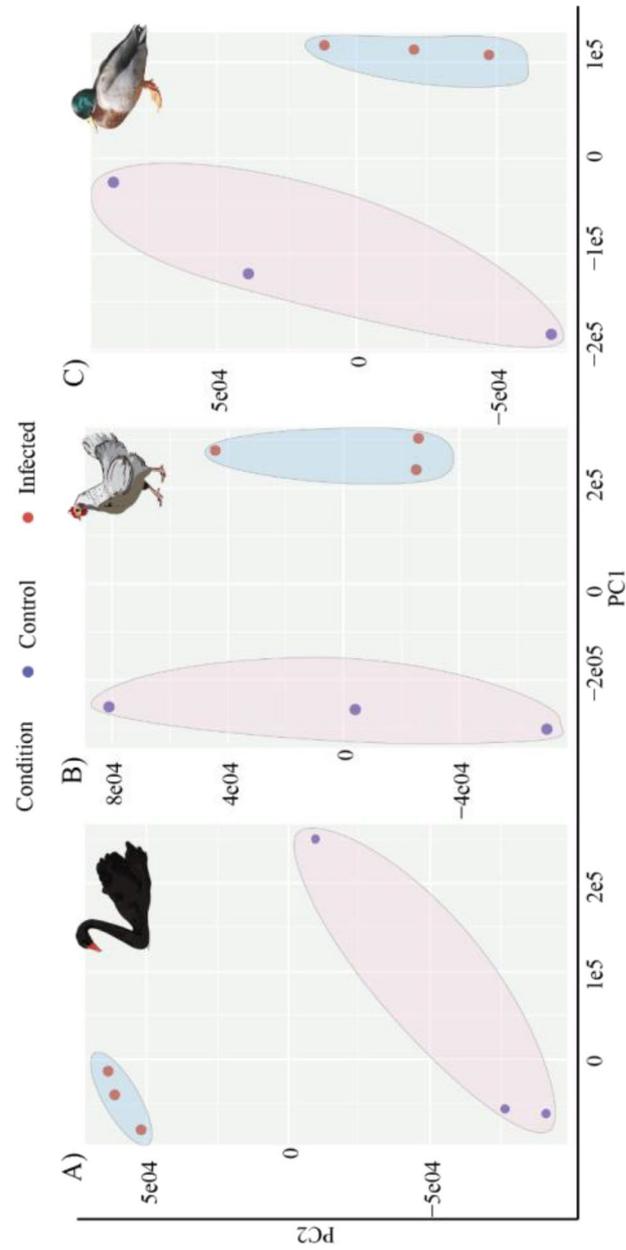
**Fig. 3** GO enrichment analysis for over-represented GO terms associated with significantly expanded gene families in black swan, mute swan, chicken, and mallard duck

in the black and mute swan (Additional file 1: Figure S8). These were located on chromosome 33 in the swan genomes. A similar number of MHC complex associated genes were identified in each species. None of these genes appeared to be pseudogenes. Compared to mammals, both mute and the black swans have a compact, relatively simple MHC B locus (Additional file 1: Figure S9), with two class IIb (*BLB*) genes followed by a

pair of class I (*BF*) genes that flank the *TAP* genes. The *TAPBP* gene in both birds, unlike chickens, does not flank the two-class-IIb genes [19]. Overall, the MHC region of both the black and mute swan share a similar genome landscape and represent a "minimal essential MHC" similar to chicken and mallard duck [20]. It is therefore unlikely that differences in the MHC complex contribute to species-specific differences in the response to HPAI virus infection.

#### **Toll-like receptor 7 expression cannot be detected in black swan endothelial cells**

Toll-like receptor 7 (TLR7) signaling has been implicated in influenza A virus recognition in mammals and birds where it functions as a pathogen recognition receptor that recognizes single-stranded viral-RNA [21]. *TLR7* has been duplicated independently in several avian species [22] and differences in *TLR7* tropism and function have been associated with the increased resistance of ducks to HPAI [23]. There was no notable structural difference in the *TLR7* gene between the black and mute swan genomes. However, strikingly, *TLR7* expression signals were detected in ISO-Seq analysis of mute swans but not in the ISO-seq analysis of black swan (Additional file 1: Figure S10). To independently confirm these data, we investigated the expression of *TLR7* using qRT-PCR in black swan tissues collected post-mortem. *TLR7* mRNA could not be detected in any of the collected black swan tissues (Additional file 9: Table S7). As *TLR7* expression can be induced by interferon, we reasoned that gene expression in the black swan may only be detected in the presence of virus infection. Accordingly, we sought to establish an in vitro model of HPAI infection in black swans. In black swans experimentally infected with HPAI virus, endothelial cells are the primary target of infection [3]. We observed similar infection of endothelial cells in swans naturally infected with HPAI (Additional file 1: Figure S11). We therefore cultured primary black swan endothelial cells according to our previously described protocol for avian species [24] and endothelial cell identity was confirmed by tube formation, uptake of acetylated low density lipoprotein, von Willebrand factor expression and the absence of *CD45* expression (Additional file 1: Figure S12). Chicken, duck, and black swan endothelial cells were infected with A/Chicken/Vietnam/008/2004/H5N1 (VN04) and 6 h later gene expression was examined by RNASeq. PCA plots showed that mock and virus-infected samples clustered separately for all three species (Fig. 4). Viral-RNA was detected in the infected endothelial cells of all three species (data not shown). Importantly, while *TLR7* transcription was upregulated (although not statistically significant) in infected duck and chicken endothelial cells, *TLR7* transcription could not be detected in infected or naïve black swan endothelial cells (Table 2). Moreover, while *myeloid differentiation primary response 88* (*MyD88*), the downstream adaptor of *TLR7* was upregulated in infected duck and chicken endothelial cells, it was downregulated in infected black swan endothelial cells (Table 2). These data are consistent with the absence of *TLR7* expression in black swan endothelial cells, despite an apparently intact *TLR7* gene in the genome. No other notable expression differences in ISOSeq/RNASeq transcripts were recorded in immune genes noted in the influenza A KEGG pathway ([https://www.kegg.jp/kegg-bin/show\\_pathway?ko05164](https://www.kegg.jp/kegg-bin/show_pathway?ko05164)).



**Fig. 4** Principal component analysis (PCA) of infected and uninfected **A** black swan, **B** chicken, and **C** duck endothelial cells. The control and the infected groups showed intergroup clustering, indicating differences in whole transcriptome profiles between the control and the infected group in each species

**Table 2** Quantified *TLR7* and *MyD88* expression following VN04 infection of avian endothelial cells

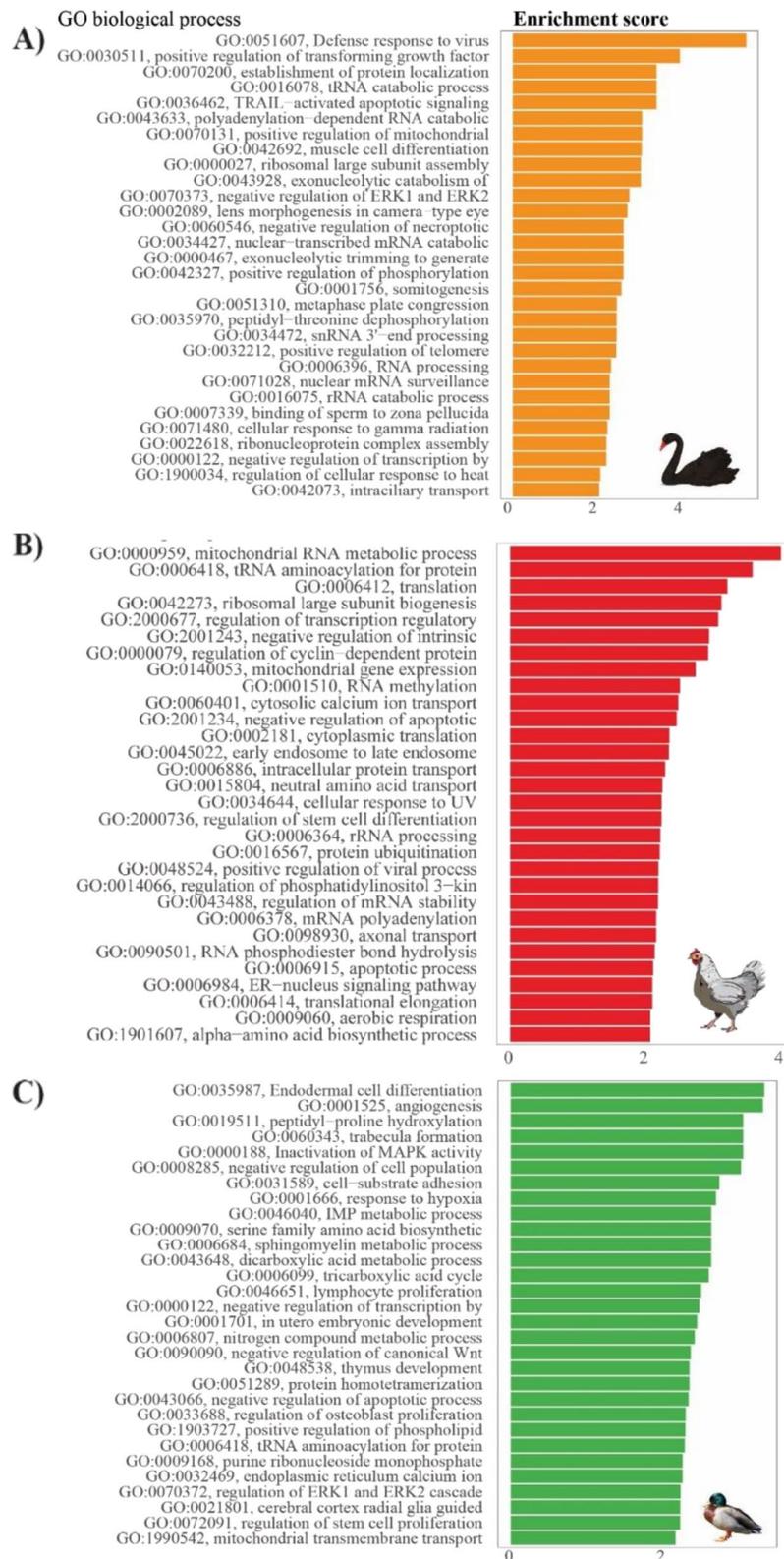
Species	Gene	Fold-change	Adjusted <i>p</i> value	Ensemble gene ID
Duck	<i>TLR7</i>	3.13	>0.1 (0.57)	ENSAPLG00000026818
Duck	<i>MyD88</i>	1.37	<0.1 (9.87e-8)	ENSAPLG00000026818
Chicken	<i>TLR7</i>	2.37	>0.1 (0.23)	ENSGALG00000016590
Chicken	<i>MyD88</i>	1.31	<0.1 (1.9e-7)	ENSGALG00000005947
Black swan	<i>TLR7</i>	NE	NA	<sup>a</sup> ENSACYG00000003732
Black swan	<i>MyD88</i>	-1.22	>0.1 (0.15857)	<sup>a</sup> ENSACYG00000016507

<sup>a</sup> Corresponding gene ID Ensembl annotation for black swan

NA not applicable, NE not expressed/detected

### Black swan endothelial cells display a dysregulated pro-inflammatory response to HPAI virus infection

The transcriptomic data generated herein offer a unique insight into the transcriptomic response of black swans to HPAI virus infection. To explore this matter further, we performed Gene Ontology (GO) enrichment analysis using significantly differentially expressed genes in infected chicken, duck, and black swan endothelial cells (Additional file 10: Table S8, Additional file 11: Table S9, Additional file 12: Table S10). The most significantly upregulated gene in black swans was *IFIT5*, *IL8* in chickens, and *BCOR* in ducks. *IL6* was significantly upregulated in the black swan ( $\log_2$ fold change = 1.89) and chicken ( $\log_2$ fold change = 1.21), indicating a strong pro-inflammatory response while not differentially expressed in ducks. Black swan, chicken, and duck endothelial cells differentially expressed other cytokines/chemokines and their receptors in response to HPAI VN04 infection (Additional file 13: Table S11). Typically, black swan and chicken endothelial cells upregulated more cytokines and cytokine receptors than duck endothelial cells in response to HPAI VN04 infection. Indeed, the highest number of cytokines and cytokine receptors were upregulated by infected chicken endothelial cells (Additional file 13: Table S11). In infected black swan endothelial cells, 113 GO terms were significantly enriched (Additional file 14: Table S12; Fig. 5A). Many of these GO terms were associated with innate immunity, the cytokine signaling response and chemokine signaling. Several innate immunity pathways were increased in response to viral infection ( $z$ -score > 0) while GO terms such as negative regulation of Mitogen-Activated Protein Kinase (MAPK) activity and negative regulation of c-JUN N-terminal Kinase (JNK) cascade were decreased ( $z$ -score < 0). Similarly, 123 enriched GO terms in infected chicken endothelial cells included positive regulation of viral response and regulation of leukocyte chemotaxis (Additional file 15: Table S13; Fig. 5B). Terms such as leukocyte mediated cytotoxicity were increased after infection ( $z$ -score > 0) while negative regulation of apoptotic signaling and the positive regulation of innate immune responses were decreased. GO biological process terms enriched in infected duck endothelial cells were not primarily associated with the innate immune response (Additional file 16: Table S14; Fig. 5C). Rather, most genes were linked to cellular biological signaling and activity. This finding is consistent with our previous study of HPAI viruses in duck endothelial cells [25]. Interestingly, in direct contrast to black swans, the inactivation of MAPK activity was significantly increased in ducks ( $z$ -score < 0). Due to the wide-ranging



**Fig. 5** Top 30 GO biological process terms that were significantly enriched in infected **A** black swan, **B** chicken, and **C** duck endothelial cells. The bars represent the enrichment score for the corresponding GO biological term with a  $p$ -value  $< 0.05$

role of the MAPK cascade, including pro-inflammatory responses, we further investigated the expression profiles of the genes and identified ten genes involved in the “inactivation of MAPK pathway,” five of which were significantly downregulated genes (i.e., *DUSP1*, *DUSP4*, *DUSP7*, *DUSP10*, and *RGS3*) in black swans (Additional file 1: Figure S13). Dual-specificity phosphatases (DUSPs) are negative regulators of MAPKs and their associated pro-inflammatory effects [26]. Accordingly, we specifically examined the differential expression of DUSPs across the three avian species. In the black swan, all DUSPs were either not differentially expressed or downregulated in response to infection. In contrast, in the duck, all DUSPs (except for *DUSP15*) were upregulated. Similar, in the chicken, *DUSP1*, *DUSP5*, *DUSP7*, *DUSP10*, *DUSP15*, and *DUSP16* were significantly downregulated in response to infection (Additional file 17: Table S15). In contrast, in the duck, all *DUSPs* (except for *DUSP15*) were upregulated. This transcriptional profile is consistent with poor regulation of a pro-inflammatory response to HPAI virus in black swans.

## Discussion

The black and mute swan reference genomes provided herein represent the first publicly available swan genomes. The analyses of these genomes, together with the first black swan transcriptome in response to HPAI virus infection, has provided a unique insight into the plumage and immune system of the black swan.

The genomic insights provided by the present study were only possible due to the growing availability and accessibility of third generation sequencing. Specifically, older technologies that generate short-read sequences can result in incorrect assembly, annotation errors, and a large amount of manual effort to correct individual genes [7]. In contrast, and consistent with the broader goals of the Vertebrate Genomes Project [7], the use of longer read sequences herein allowed us to generate black swan and mute swan genomes that were scaffolded to near chromosomal length and that were of comparable quality to the well-annotated chicken reference genome.

Genomic analysis of genes known to be associated with plumage color in other birds [27] identified a potential frame-shift mutation in the first exon of *SLC45A2* in the mute swan, which may have led to pseudogenization of this gene. *SLC45A2* encodes a transporter protein involved in melanin synthesis and is considered one of the most important proteins affecting human pigmentation [28]. Mutations in the *SLC45A2* gene have been reported in albinism in humans [29]. Furthermore, mutations in the gene have been associated with plumage color variation in Japanese quails [13], indicating the importance of the *SLC45A2* in avian plumage. Interestingly, should a mutation of *SLC45A2* have resulted in the differential plumage of the black and mute swan, it would suggest that the last common ancestor of these birds was, in fact, black. This is direct contrast to the metaphor of “black swan events” that are so defined because of their unprecedented and unexpected nature. Instead, it would appear that at one point in history black plumage for the swan was the norm rather than the exception.

Compared to the last common ancestor, mute swan and mallard duck gene families involved in immune system processes were expansive. In contrast, no expansion in immune gene families was noted in the chicken or the black swan. This differential immune gene expansion, and its implications for susceptibility to HPAIV, is likely

compounded by the observed impaired expression of *TLR7* in the endothelial cells of black swans. Interestingly, other genes that have been observed to be differentially expressed between chickens and ducks, and implicated in susceptibility to HPAIV, were not differentially expressed between infected black swan and duck endothelial cells (e.g., *RIG-I* and *IFITM3*) [30–33]. It is interesting to speculate as to whether mute swan endothelial cells would express *TLR7*. However, the presence or absence of *TLR7* in the endothelial cells of mute swans is perhaps irrelevant to the pathogenesis of HPAIV, as the virus is not heavily endothelial tropic in this species [3]. In the black swan, the observed differences in *TLR7* expression in endothelial cells speak to the value of combining genomics with both primary cell culture and transcriptomics, as has recently been suggested as the new standard for comparative genomics by Stephan and colleagues [34]. Either as a result of, or in addition to, these observed immune differences, black swan endothelial cells also mounted a markedly pro-inflammatory response to HPAIV infection. We have previously reported a similar pro-inflammatory response in infected chicken endothelial cells (compared to those of ducks) and speculated that this inflammatory response leads to immunopathology in chickens *in vivo* [25]. Whether disease severity in black swans is driven by immunopathology remains to be determined, although it is consistent with the observed pathology in infected birds [2, 3]. In sum, it is likely that this combination of species-specific differences in the immune response contribute to the marked susceptibility of both the black swan and chicken to HPAIVs.

## Conclusions

The observed species-dependent differences in the immune responses of swans raise the intriguing question as to why the black swan continues to thrive in its native Australia as well as in New Zealand (where it was introduced in nineteenth century). This may be due to the fact that HPAIV is not endemic in Australia and New Zealand. Indeed, captive populations of black swans located in parts of the world frequently exposed to HPAIV are highly susceptible to severe disease [4]. The data presented in this study would therefore suggest that should HPAI become more prevalent in the Oceania region, the ongoing survival of the black swan would be at significant risk. Moreover, many of the immune limitations described herein are not specific to avian influenza viruses. For example, *TLR7* is essential in the immune recognition of a wide number of viral pathogens including avian coronaviruses [35]. These data suggest that should any avian endothelial specific viral infection become established in the native habitat of the black swan, the survival of this iconic species would be in significant peril.

## Methods

A full description of the methods can be found in Additional file 3: Online Methods. No statistical methods were used to predetermine sample size.

### Black swan genome

#### Sequencing

For genomic DNA and RNA extraction, liver, spleen, kidney, skeletal muscles, lung, heart, stomach, intestines, and gonads were obtained from an adult male black swan that had to be euthanized at Currumbin Wildlife Sanctuary hospital (<https://curru>

[mbinsanctuary.com.au/wildlife-hospital](https://mbinsanctuary.com.au/wildlife-hospital)) due to injury. Genomic DNA from the stomach tissue was extracted using a commercially available kit (MagAttract HMW DNA kit – Qiagen®). A DNA library was prepared using Single Molecular Real-Time (SMRT) bell template Prep Kit 1.0 (Pacific Biosciences) with unsheared gDNA with a size cut-off of 20kb using a Blue Pippin instrument (Sage Science ©) according to the manufacturer's protocol. SMRT sequencing was performed using PacBio Sequel instrument at the Institute of Molecular Biosciences, the University of Queensland using 1-6 pica Molar (pM) loading concentration, version 3.0 sequencing chemistry, diffusion loading, and 10-h movies over 10 SMRT cells (1M V.3).

### **Genome assembly**

Raw SMRT sequencing reads were used for de novo black swan genome assembly using open-source FLACON (v2018. 31-08-03.0)/FALCON-UNZIP (6.0.0.47841) diploid aware genome assembly algorithms to produce primary and alternative haplotigs as the primary assembly. For FALCON assembly, the estimated genome size (haplotype) used for the black swan was  $1.4181 \times 10^9$  base pairs. Additional parameters used for FALCON assembly are listed in Additional file 18: Table S16. Raw FASTA sequences were extracted from the subread BAM files produced by SMRT sequencing using PacBio® BAM2fasta tool. Raw FASTA sequences from 10 SMRT cells were given as the input for the FALCON assembler, and the FALCON assembly parameters were provided using a configuration file. Once the FALCON run was completed, FALCON-UNZIP was executed in the same directory as the initial FALCON run. For the FALCON-UNZIP run, the subread BAM data was also used. Once the FALCON UNZIP run was complete, two genome polishing steps were undertaken through PacBio® SMRT Link version 7.0.0 resequencing pipeline using PacBio CLR reads with default settings (-x 30) in addition to FALCON\_UNZIP integrated arrow polishing to produce the primary PacBio assembly. This PacBio assembly was screened for the presence of mitochondrial/non-chromosomal DNA using NCBI Blast+ (v2.9.0+) against the black swan mitochondrial genome [36]. The mitochondrial DNA sequences were removed from the primary PacBio assembly. PURGE\_DUPS (v1.2.5) was next employed to remove false haplotypic duplications from the concatenated primary PacBio assembly and alternate haplotigs. The PacBio assembly completeness was assessed via BUSCO V.5.1.2 against avian orthoDB10. The presence of highly conserved core vertebrate genes was evaluated using the CEGMA pipeline, available online in gVolante (<https://gvolante.riken.jp/>).

The FALCON draft was scaffolded to chromosome-length by the DNA Zoo Consortium following the methodology described here: [www.dnazoo.org/methods](http://www.dnazoo.org/methods). Briefly, the Hi-C data was processed using Juicer [37] and used as input into the 3D-DNA pipeline [38] to produce a candidate chromosome-length genome assembly. We performed additional finishing on the scaffolds using Juicebox Assembly Tools [37, 38]. The contact matrices generated by aligning the Hi-C data to the genome assembly before and after the Hi-C scaffolding are available for browsing at multiple resolutions on [https://www.dnazoo.org/assemblies/Cygnus\\_atratus](https://www.dnazoo.org/assemblies/Cygnus_atratus) visualized using Juicebox.js, a cloud-based visualization system for Hi-C data [39]. Finally, the Hi-C assembly was subjected to gap

closing using the TGS-GapCloser (v1.0.1) algorithm [40] with default settings. PacBio continuous long reads (CLRs) were used as input for the software tool for gap closing.

The *Anatidae* family's established terminology for the chromosome-length scaffolds was based on the homology between black swan scaffolds/chromosomes and the closely related mute swan. MUMMER (v3) was used to establish a 1:1 homology between the black swan and the mute swan before assigning the macro and micro-chromosomes terminology to black swan chromosomes.

### **Mute swan genome**

#### **Sequencing**

The mute swan genome was sequenced by the vertebrate genome project (VGP) according to the "vertebrate genome project assembly phase I" pipeline. The genomic DNA was extracted from the stomach tissue of a male adult mute swan. The sequencing technology and data used to build the mute swan haploid chromosome length hybrid assembly included PacBio Sequel I CLR, Illumina Novaseq<sup>®</sup>, Arima<sup>®</sup> Hi-C linkage data and BioNano optical maps.

#### **Assembly**

FALCON (v2018.31-08-03.0)/FALCON-UNZIP (6.0.0.47841) was used to generate the phased contigs. To remove the false duplications in the phased haplotigs, Purge\_Haplotigs (v1.0.3+) was used. After that, two rounds of scaffolding were performed using 10x Genomics link data with scaff10x (v4.1.0) (<https://github.com/wtsi-hpag/Scaff10X>) for the first step of scaffolding with 10xGenomics<sup>®</sup> link data to build the S1 scaffold set. Next, the second step of scaffolding was performed on S1 using BioNano data.

BioNano solve (v3.2.1\_04\_122018) software was used to produce the S2 scaffold. The last scaffolding was performed to scaffold the S2 set using Arima<sup>®</sup> Hi-C linkage data with SALSA (v2.2) [41, 42]. The Hi-C contact matrices were generated and visualized as described in the vgp assembly pipeline (<https://github.com/VGP/vgp-assembly>). Primary and alternate haplotigs were concatenated, and bases were polished with Arrow (SMRTanalysis v6.0.0.47841) using PacBio CLR reads. Two more rounds of polishing were performed with linked reads by aligning with Longranger (v2.2.2) and variant calling with FreeBayes (v1.3.1) followed by consensus calling with bcftools consensus. Finally, manual curation of the assembly was performed using gEVAL (v2019-09-26).

The chromosome assignment was performed using evidence from Hi-C. A scaffold was considered a chromosome with Hi-C evidence when there is a clear unbroken diagonal in the Hi-C maps in the Juicebox. Every Hi-C box from the largest to the smallest for evidence-validated scaffolds was considered as a chromosome. Subsequently, established terminology for the chromosomes was given for the *Anatidae* family.

#### **Genome annotation**

The final assembly's structural and functional annotation of genes was conducted using de novo homology-based and evidence-based methods (for details, see Additional file 3: Online Methods). Firstly, the primary PacBio assembly was annotated using NCBI eukaryotic and Ensembl gene annotation pipelines. A set of transcriptomic data (of both short-reads and long reads) derived from the liver, spleen, peripheral blood mononuclear

cells and primary bone marrow-derived endothelial cells was used for black swan genome annotation, and the ISOseq data for mute swan annotation was derived from the peripheral blood.

#### **Structural variant assessment of the genes associated with plumage color in swans**

Structural variants in genes associated with plumage color were examined using the University of California, Santa Cruz (UCSC) genome browser with multiple whole-genome alignments. We first aligned four Gallianseriformes' whole genomes (black swan, mute swan, chicken and mallard duck) with CACTUS to produce a HAL alignment file [43]. The HAL file alignment was then visualized in the Genome assembly In a Box (GBiB) using the HAL tools package (V2.1). We then used predicted protein and mRNA sequences plumage color genes (from all four species) as queries to perform protein coding region alignment with the BLAT built in the GBiB and compare with each species in the HAL alignment. We validated putative open reading frames (ORF) for each gene from each species by visual inspection. Validity was assessed based on the presence or absence of putative donor-recipient splice sites, start codon, and in-frame stop codons. We also calculated the percentage identify and coverage for each gene between mute and black swans using Blast pairwise alignment.

#### **Gene family evolution**

The longest peptide sequence for a given protein-coding gene for 17 species was retrieved from Ensembl (release 104) and NCBI Annotation release 103, including zebrafish (*Danio rerio*) (GCA\_000002035.4), Japanese rice fish (*Oryzias latipes*) (GCF\_002234675.1), three-pinned stickleback fish (*Gasterosteus aculeatus*) (BROAD S1), tropical clawed frog (*Xenopus tropicalis*) (GCA\_000004195.3), platypus (*Ornithorhynchus anatinus*) (GCA\_004115215.2), gray short-tailed opossum (*Monodelphis domestica*) (GCA\_000002295.1), mouse (*Mus musculus*) (GCA\_000001635.9), human (*Homo sapiens*) (GCA\_000001405.28), cattle (*Bos taurus*) (GCA\_002263795.2), anole lizard (*Anolis carolinensis*) (GCA\_000090745.2), mallard duck (*Anas platyrhynchos*) (GCA\_002743455.1), turkey (*Meleagris gallopavo*) (GCA\_000146605.4), chicken (*Gallus gallus*) (GCA\_000002315.5), zebra finch (*Taeniopygia guttata*) (GCA\_003957565.2), and collared flycatcher (*Ficedula albicollis*) (GCA\_000247815.2). The longest amino acid sequences from coding genes in the final reference annotations were used for the black and mute swan. The phylogenetic tree for the above 17 species was constructed using the Orthofinder pipeline with – M msa -S blast -I 1.3 settings. The final rooted species tree was inferred using STAG from the Orthofinder, which was then made ultra-metric with the root-age of 435 million years (obtained from [TimeTree.org](https://www.timetree.org/) database). Peptide sequences in 17 species were scored using PANTHER (v15.0) database and clustered into PANTHER families and subfamilies and then identified by a family-specific PANTHER identifier. The maximum likelihood of gene family evolution for a given taxon was estimated by running CAFÉ(v5) with the PANTHER assigned protein families of the 17 species and the Orthofinder-based ultrametric species tree (significant at  $p$ -value < 0.05).

### **Functional profiling of expanded and contracted gene families in swans**

Each PANTHER sub-family identification with a significant gene family evolution was assigned to the gene ontology category (GOSlim identification number) to examine the biological role of significantly evolving (either by expansion or contraction) (Vertibri  $p$ -value < 0.05) gene families through the R-package PANTHER.db. The significant gene families that appeared to have expanded or contracted were separated. Subsequently, GO enrichment analyses were conducted using topGO [44] to test for over-represented functional terms associated with expanded and contracted gene families (a given gene family is considered expanded if the number of genes are increased compared to last common ancestor). The significant gene sub-families annotated with GOSlim identifications were used as the foreground, and the GOSlim annotated gene subfamilies of the last common ancestor with at least one gene present in the family were used as the background. Fisher's exact test was used for statistical testing, and the terms with corrected  $p$ -value < 0.01 were chosen as overrepresented GO terms.

### **Comparison of the immune gene repertoire**

Selected immune databases were used to obtain the list of known immune genes from Ensembl release and were compared to that of the black swan and mute swan genome (see Additional file 3: Online Methods for details)

### **Major histocompatibility complex (MHC) class annotation in black and mute swan genomes**

The major histocompatibility complex (MHC) loci (MHC class I and MHC class IIb) for both mute and the black swan genomes were identified as previously described [45]. Briefly, the avian order-level consensus sequences for MHC I and MHC II loci were built using several different bird species covering at least three exons (exon 2, exon 3, and exon 4). The blast algorithm aligned the order-level consensus sequences to the black swan and mute swan genomes. MHC loci were estimated as the number of blast hits that contained all three exons within 2kb of each other. Each exon was examined for in-frame stop codons, and it was eliminated as a locus if present. Genes present in regions 100kb upstream and downstream of each locus were manually annotated and the gene location plotted through genes to identify the additional genes of each predicted locus. Each predicted MHC locus was manually inspected for premature stop codons, and if present, was eliminated. Identified MHC molecules and MHC B loci associated genes were visualized for black and the mute swan for comparison using Circlize (R-package).

### **TLR7 expression in black swan tissues**

Total RNA was extracted from tissues of an adult male black swan (stomach, kidney, and liver) using an RNA plus mini kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Genomic DNA was extracted from the stomach muscles using the MagAttract HMW DNA kit (Qiagen®, Hilden, Germany) for the positive control, following the manufacturer's recommendation. The black swan-specific *TLR7* (FW: *TTG CACTTCCACACTCCAAG*; RV: *TTCAGTCCAATTGCACCTCTG*; Probe: *CTCCGAAAC AATCGCATTC AACGG*) and *18S* (FW: *CCTGCGGCTTAATTTGACTC*; RV: *AGACAA ATCGCTCCACCAAC*; Probe: *TTGAGAGCTCTTTCTCGATTCCGTGG*) primers and probes were designed using the primer3plus web tool

### **Immunohistochemistry**

Immunohistochemistry for influenza A virus nucleoprotein (NP) was performed according to Additional file 3: Online Methods.

### **Isolation of primary black swan, chicken, and duck endothelial cells**

We used previously isolated and characterized chicken and duck primary aortic endothelial cells as the controls for black swan primary endothelial cell characterization [24, 25]. Briefly, 17-day-old chicken and 21-day-old Pekin duck embryonated eggs were purchased from Darling Downs Hatchery (Queensland, Australia). Primary aortic endothelial cells were cultured from the aortic arches of chicken and duck embryos as described previously using EGM-2MV medium (Lonza, Basel, Switzerland) with 10% FBS (Gibco, Waltham, MA, USA) [305, 307]. All avian cells were cultured at 40 °C 5% CO<sub>2</sub> unless otherwise stated.

The black swans used for cell isolation were sourced from Currumbin Wildlife Animal Hospital (−28.14, 153.48), Queensland, Australia, and Tauranga Harbor, Bay of Plenty, New Zealand (−37.57, 175.97). The animals were either euthanized due to a terminal illness or as part of a government-approved cull. Tibiotarsi and femurs were obtained from all birds. The bones were then partially opened in the middle to expose the bone marrow and then submerged in pre-chilled (4°C) Dulbecco's modified Eagle medium (DMEM) in T175 flasks. Flasks were then transferred to a personal containment level 2 (PC2) laboratory for further processing within 4 to 3 h post-mortem.

Bone marrow was subsequently removed and resuspended in pre-chilled DMEM. A 15-ml suspension of DMEM containing bone marrow was then layered after filtering through a 40-µm sterile filter over 15ml of Lymphoprep<sup>®</sup> in a 50-ml SepMate<sup>®</sup> tube. The layer of cells at the DMEM-Lymphoprep interface was separated according to the SepMate manufacturer's instructions. The separated cell layer was then resuspended in 1ml EGM-2MV media and enumerated. The cell concentration was adjusted to approximately  $1 \times 10^6$  cells per ml before transferring to a Petri dish with 10ml EGM-2MV media. The Petri dish was incubated at 40°C with 5% CO<sub>2</sub>. The media was replaced every other day until cells became confluent. Bone marrow-derived primary cells (EPC) of passage 4 were cultured to confluency and incubated with 1.75µg/ml Alexa Fluor<sup>™</sup> 488 Acetylated low-density lipoprotein (ac-LDL) (Invitrogen, USA) for 4 h. The cells were detached by washing twice with PBS and incubating with 0.05% Trypsin-EDTA (Thermo Fisher<sup>®</sup>, USA). The cells were mixed in 2% (v/v) FBS in PBS and subsequently sorted with the Moflo Astrios<sup>®</sup> High-Speed cell sorter (Beckman Coulter<sup>®</sup>, CA, USA) under sterile conditions. Ac-LDL-stained cells were sorted based on prior knowledge of high and low ac-LDL intake [303] using BD FACSDiva<sup>™</sup> (version 8.0.1 – BD<sup>©</sup>). The sorted cells were counted and transferred to gelatine-coated plates and grown in EGM-2MV (Endothelial Basal Media with EGM – 2MV bullet kit (Lonza<sup>©</sup>). Cells at post-sorted passages were known as black swan primary endothelial cells (bsEPCs). The purity of bsEPCs was confirmed as described by Additional file 3: Online Methods.

### **Modeling HPAI in avian endothelial cells**

The experiment was designed to include two groups viz. challenge and control, from each of the species, i.e., black swan, chicken and duck. Each group consisted of three biological replicates of endothelial cells.

### **Viral infection**

A/Chicken/Vietnam/0008/2004(H5N1) (VN04) virus was amplified in MDCK cells. All experiments using HPAIVs were performed under physical containment level 3 (PC3) settings at the Australian Centre for Disease Preparedness (Geelong, Australia). Primary endothelial cells were grown to at least 60% confluency and were inoculated with  $5 \times 10^5$  PFU/mL of VN04 for 60 min at 40°C under physical containment 3 (PC3) settings. The inoculum was removed, and the cells were washed once with PBS, and EGM-2MV media without FBS was added. The cells were incubated for an additional 5 h.

### **RNA extraction**

According to the manufacturer's instructions, cellular RNA was extracted from cultured cells using the RNeasy Plus Kit (Qiagen, Hilden, Germany). Subsequently, 3M of sodium acetate (NaOAc, pH 5.5) and 100% ethanol were added for RNA precipitation.

### **RNA sequencing**

Paired-end, 150bp long, RNAseq was performed by Macrogen (Macrogen, Seoul, South Korea) approximately 40 million reads per sample. Library preparation was performed using the SMARTer<sup>®</sup> ultra-low kit (Takara Bio<sup>®</sup> USA, Inc.). All the libraries were barcoded and subsequently sequenced using the Novaseq<sup>®</sup> 6000 (Illumina, San Diego, CA, USA) with Novaseq<sup>®</sup> 6000 S4 Reagent Kit (Illumina, San Diego, CA, USA) as per the manufacturer's instructions using within a single lane in a single cell to avoid potential technical variations. RNAseq reads were separated into original samples based on the corresponding bar code.

### **Differential gene expression analysis**

RNAseq-based differential gene expression analysis was performed after assessing the quality of the raw reads using the FastQC tool before and after adaptor and quality trimming with TrimGalore version 0.6.2 (RRID: SCR\_011847). RNAseq reads were then quantified at the transcript level using Salmon v1.5.2. First, the indexes were created for the black swan reference transcriptome and the Ensembl version 104 chicken (GRCg6a) and duck (CAU\_Duck1.0) transcriptomes with following parameters (salmon quant -i (params.ind) -l (params.libtype) -g (params.genemap) -r input\_r1 -f input\_r2 --validateMappings -o (output.dir)). The transcript abundance data of each species were imported through the R package "tximport" by transforming data from transcripts to gene-level quantification. The expression data from each species (black swan, chicken, and duck) were independently analyzed for differential gene expression. Once RNAseq read libraries were normalized for sequencing depth and RNA composition, differential gene expression analysis was performed using DESeq2 version 1.30.1 [46]. A Bonferroni adjusted *p-value* (*adj.p.val*) cut-off of 0.05 and log2fold enrichment of 0.58 (fold-change 1.5) were used as the significance threshold (differential gene expression: mock vs. infected).

**Table 3** The script usage and availability

Usage	DOI	Ref
Genome assembly Black swan	<a href="https://doi.org/10.5281/zenodo.7272708">https://doi.org/10.5281/zenodo.7272708</a>	[50]
Genome assembly mute swan (forked repository from original)	<a href="https://doi.org/10.5281/zenodo.7272713">https://doi.org/10.5281/zenodo.7272713</a>	[7]
MHC class annotation	<a href="https://doi.org/10.5281/zenodo.7272715">https://doi.org/10.5281/zenodo.7272715</a>	[51]
Genome annotation	<a href="https://doi.org/10.5281/zenodo.7272717">https://doi.org/10.5281/zenodo.7272717</a>	[52]
RNAseq data analysis	<a href="https://doi.org/10.5281/zenodo.7272720">https://doi.org/10.5281/zenodo.7272720</a>	[53]
Gene family evolution	<a href="https://doi.org/10.5281/zenodo.7272722">https://doi.org/10.5281/zenodo.7272722</a>	[54]
Comparison of immune gene repertoire	<a href="https://doi.org/10.5281/zenodo.7272722">https://doi.org/10.5281/zenodo.7272722</a>	[54]

**Table 4** Data availability

Title/description	Species	Ref
Swan assembly—primary	Black swan	[55]
Swan assembly—Final	Black swan	[56]
Swan assembly—Final	Mute swan	[57]
Swan ISOseq	Black swan	[58]
Swan ISOseq	Mute swan	[59]
PacBio raw reads	Black swan	[60]
PacBio raw reads	Mute swan	[61]
RNA-seq data—Black swan endothelial cells	Black swan	[62]
RNA-seq data—Black swan endothelial cells	Black swan	[63]
RNA seq data—Black swan endothelial cells	Black swan	[63]
Scaffolding	Black swan	[64]
Short-read DNA sequences—Black swan	Black swan	[65]
Scaffolding	Mute swan	[61]

### Gene Ontology enrichment analysis

Gene Ontology (GO) terms were assigned to all quantified genes in chicken and duck using the BioConductor package BioMart [47], and the GO terms were assigned to all expressed and quantified genes using InterProScan5 [48]. For black swan genes, we used InterProScan 5 [48] for GO annotation. Using the GO annotation, we then performed GO enrichment analysis with topGO [44]. For the statistical enrichment, Fisher's exact test was used with the "weight01" algorithm. GO enrichment analyses were performed for all three GO categories, including "biological process," "cellular component," and "molecular function." Results of the GO enrichment were visualized with "ggplot2" and "GOplot" [49] R packages. We calculated a score (named as "z-score" in the GOplot package) to evaluate the trend (increasing or decreasing biological process) using GOplot in-built mathematical formula (Formula 1) for enriched terms of the GO biological process category in each species.

### Availability of data and materials

The codes and data used for the bioinformatic analysis are detailed in Tables 3 and 4, respectively.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-022-02838-0>.

Additional file 1: Supplementary Figure S1. Schematic overview of the process used to construct the black swan genome. Supplementary Figure S2. Schematic overview of the assembly pipeline version 1.5 of the vertebrate genome project used to construct the mute swan genome. Supplementary Figure S3. Hi-C contact maps for the black swan (a) and mute swan (b). Supplementary Figure S4. Schematic overview of the gene annotation pipeline. Supplementary Figure S5. The largest 34 chromosomes aligned between the black swan (left) and the mute swan (right). Supplementary Figure S6. Collinearity analysis of the Z chromosome between the mute swan (left) and black swan (right). Supplementary Figure S7. Black swan and duck synteny plot between the first chromosome. Supplementary Figure S8. MHC Class I (A) and MHC Class II (B) aligned regions of order-level consensus exons of the mute and black swan. Supplementary Figure S9. Relative locations of MHC complex and associated genes in black and mute swan chromosome 33. Supplementary Figure S10. TLR7 expression can be detected in ISO Seq analysis of the mute swan (A) but not the black swan (B). Supplementary Figure S11. IAV NP antigen distribution in tissues from a black swan naturally infected with A/Black swan/Akita/2/2016 (H5N6) in 2016 in Akita prefecture, Japan. Supplementary Figure S12. The successful culture of primary black swan endothelial cells confirmed by qRT-PCR, immunofluorescence and tube formation. Supplementary Figure S13. Differential regulation of DEGs linked to the “inactivation of MAPK” pathway in black swan endothelial cells in response to VN04 infection along with randomly selected GO terms.

Additional file 2: Supplementary Table S1. Heterozygosity, QV and completeness values for swan genomes.

Additional file 3: Online Methods.

Additional file 4: Supplementary Table S2. BUSCO analysis of the final genomes.

Additional file 5: Supplementary Table S3. CEGMA analysis of the final genomes.

Additional file 6: Supplementary Table S4. GO biological process annotation of inverted genes in the Mallard duck (relative to the black swan).

Additional file 7: Supplementary Table S5. List of genes known to affect plumage colour in birds.

Additional file 8: Supplementary Table S6. Immune gene families are contractive in black swans compared to mute swans. The number of genes in each immune gene sub-family identified is given in each column for the corresponding species.

Additional file 9: Supplementary Table S7. TLR7 expression could not be detected in black swan spleen, liver and kidney by qRT-PCR. \*ND: Not detected.

Additional file 10: Supplementary Table S8. Differentially expressed genes in infected black swan endothelial cells.

Additional file 11: Supplementary Table S9. Differentially expressed genes in infected chicken endothelial cells.

Additional file 12: Supplementary Table S10. Differentially expressed genes in infected duck endothelial cells.

Additional file 13: Supplementary Table S11. Differentially expressed cytokine or cytokine related genes in infected avian endothelial cells.

Additional file 14: Supplementary Table S12. 113 GO terms were significantly enriched in infected black swan endothelial cells.

Additional file 15: Supplementary Table S13. 123 GO terms were significantly enriched in infected chicken endothelial cells.

Additional file 16: Supplementary Table S14. 150 GO terms were significantly enriched in infected duck endothelial cells.

Additional file 17: Supplementary Table S15. Dual-specificity phosphatases in black swan, chicken and duck endothelial cells infected with VN04\*.

Additional file 18: Supplementary Table S16. Parameters used for the FALCON run. Length cut-off was detected automatically for seed-read length.

Additional file 19: Review history.

## Acknowledgements

We would like to acknowledge colleagues from the Friedrich-Loeffler-Institut (Jens Peter Teifke, Robert Klopffleisch, and Angele Breithaupt) for providing FFPE material. Special thanks to Christoph Schulze from the state diagnostic laboratory who was involved in the autopsy of most of the samples use for FFPE material collection and Ashling Charles from DNA Zoo Australia team for routine data processing support. We also would like to thank Leo Joseph from Australian National Wildlife Collection at CSIRO and Tiggy Grillo from Wildlife Health Australia for assisting/obtaining samples for genome sequencing. We are also thankful for research computing center at The University of Queensland for providing the high performance computing facility for black swan genome assembly and annotation.

## Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Review history**

The review history is available as Additional file 19.

**Authors' contributions**

Data synthesis and analysis: ACK, YC, KYC, AC, RK, RCM, MZWT, KDH, HBO, LES, MW, JS, EN, TJB, GGA, SL, JB, AS, AJM, PK, OD, EA, OF, GF, JM, WC, FJM, DNO, FTN, KH, JC, AT, **JS, RIK**, CT, JW, EDJ, KRS. Funding: KRS. Supervision: YC, MLB, DWB, KRS. Study design: ACK, MLB, DWB, KRS. Sample collection: ACK, KYC, MM, HGS, MP, CT, JW, MBR, TK, YS. The authors read and approved the final manuscript.

**Funding**

This study was funded by the Australian Department of Agriculture and Water Resources (CSIRO Science and Innovation Award; KRS), the Australian Department of Agriculture and Water Resources (Australian Eggs Innovation Award; KRS), and an Australian Research Council DECRA (DE180100512; KRS). This research was funded in part by the Wellcome Trust [Grant numbers 108749/Z/15/Z, 222155/Z/20/Z]. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. ACK was supported by the Australian Government's research training program (RTP) scholarship. PK is supported by the University of Western Australia with additional computational resources and support from the Pawsey Supercomputing Centre with funding from the Australian Government and Government of Western Australia. Hi-C data for the Black Swan was created by the DNA Zoo Consortium ([www.dnazoo.org](http://www.dnazoo.org)). DNA Zoo sequencing effort is supported by Illumina, Inc. E.L.A. was supported by the Welch Foundation (Q-1866-20210327), an NIH Encyclopedia of DNA Elements Mapping Center Award (UM1HG009375), a US-Israel Binational Science Foundation Award (2019276), the Behavioral Plasticity Research Institute (NSF DBI-2021795), NSF Physics Frontiers Center Award (NSF PHY-2019745), and an NIH CEGS (RM1HG011016-01A1).

**Declarations****Ethics approval and consent to participate**

Animal ethics approval was obtained from the animal ethics unit, the University of Queensland's research ethics office (ANFRA/311/18; SCMB/002/18). A wildlife research permit was obtained from the Queensland Department of Environment, Land, and Water Planning under the Wildlife Act 1975 (permit number: 10008904). For bone marrow tissue collection in New Zealand, we obtained permission from the Department of Conservation under Wildlife Act Authority on non-public conservation land (70705-FAU & 70709-DOA). Ethical approval to sample mute swans was obtained by the Linköping Animal Ethics board (permit 01125-2020).

**Competing interests**

KRS is a consultant for Sanofi, Roche, and NovoNordisk. The opinions and data presented in this manuscript are of the authors and are independent of these relationships. The other authors declare no competing interests.

**Author details**

<sup>1</sup>School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia. <sup>2</sup>Commonwealth Scientific and Industrial Research Organisation, Australian Centre for Disease Preparedness, 5 Portarlington Road, Geelong, VIC 3220, Australia. <sup>3</sup>School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia. <sup>4</sup>Department of Migration, Max Planck Institute of Animal Behavior, Radolfzell 78315, Germany. <sup>5</sup>Department of Biology, University of Konstanz, Konstanz 78457, Germany. <sup>6</sup>Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia. <sup>7</sup>Department of Organismal Biology – Systematic Biology, Evolutionary Biology Centre, Uppsala University, Science for Life Laboratory, Uppsala 752 36, Sweden. <sup>8</sup>The New Zealand Institute for Plant & Food Research Ltd, Nelson 7010, New Zealand. <sup>9</sup>School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TU, UK. <sup>10</sup>School of Agriculture and Environment, The University of Western Australia, Perth, WA 6009, Australia. <sup>11</sup>The Centre for Genome Architecture, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. <sup>12</sup>Centre for Theoretical Biological Physics and Department of Computer Science, Rice University, Houston, TX 77030, USA. <sup>13</sup>Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. <sup>14</sup>Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech, Pudong 201210, China. <sup>15</sup>The Vertebrate Genome Laboratory, The Rockefeller University, NY 10065, USA. <sup>16</sup>Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK. <sup>17</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. <sup>18</sup>National Centre for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. <sup>19</sup>The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, UK. <sup>20</sup>School of Biosciences, The University of Melbourne, Melbourne, VIC 3052, Australia. <sup>21</sup>Faculty of Veterinary Medicine, Hokkaido University, Sapporo, Hokkaido 060-0818, Japan. <sup>22</sup>New Zealand Fish & Game – Eastern Region, Rotorua 3046, New Zealand. <sup>23</sup>Department of Zoology, University of Otago, Dunedin 9054, New Zealand. <sup>24</sup>Currumbin Wildlife Sanctuary, Currumbin, QLD 4223, Australia. <sup>25</sup>Centre for Ecology and Evolution in Microbial Model Systems (EEMIS), Linnaeus University, Kalmar SE-391 82, Sweden.

Received: 27 May 2022 Accepted: 12 December 2022

Published online: 23 January 2023

**References**

- O'Brien SJ, Evermann JF. Interactive influence of infectious disease and genetic diversity in natural populations. *Trends Ecol Evol.* 1988;3:254–9.
- Brown JD, Stallknecht DE, Swayne DE. Experimental infection of swans and geese with highly pathogenic avian influenza virus (H5N1) of Asian lineage. *Emerg Infect Dis.* 2008;14:136.

3. Short KR, Veldhuis Kroeze EJ, Reperant LA, Richard M, Kuiken T. Influenza virus and endothelial cells: a species specific relationship. *Front Microbiol.* 2014;5:653.
4. van der Hoek, G. Israel: Avian Flu H5N8 in a Zoo and in Poultry - Media - OIE reports. Israel: Avian Flu H5N8 in a Zoo and in Poultry - Media - OIE reports. 2020. <https://flutrackers.com/forum/animal-diseases-of-concern-excludes-h5n1/influenza-in-animals-excl-h5n1/896191-israel-avian-flu-h5n8-in-a-zoo-and-in-poultry-media-oie-reports>.
5. Huang Y, Li Y, Burt DW, Chen H, Zhang Y, Qian W, et al. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat Genet.* 2013;45:776–83.
6. Johnson RN, O'Meally D, Chen Z, Etherington GJ, Ho SY, Nash WJ, et al. Adaptation and conservation insights from the koala genome. *Nat Genet.* 2018;50:1102–11.
7. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature.* 2021;592:737–46.
8. Takagi N, Sasaki M. A phylogenetic study of bird karyotypes. *Chromosoma.* 1974;46:91–120.
9. Zhu F, Yin Z-T, Wang Z, Smith J, Zhang F, Martin F, et al. Three chromosome-level duck genome assemblies provide insights into genomic variation during domestication. *Nat Commun.* 2021;12:1–11.
10. Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, et al. The repetitive landscape of the chicken genome. *Genome Res.* 2005;15:126–36.
11. Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. The global diversity of birds in space and time. *Nature.* 2012;491:444–8.
12. Li J, Zhang J, Liu J, Zhou Y, Cai C, Xu L, et al. A new duck genome reveals conserved and convergently evolved chromosome architectures of birds and mammals. *GigaScience.* 2021;10:giaa142.
13. Gunnarsson U, Hellström AR, Tixier-Boichard M, Minvielle F, Bed'Hom B, Si I, et al. Mutations in SLC45A2 cause plumage color variation in chicken and Japanese quail. *Genetics.* 2007;175:867–77.
14. Li D, Sun G, Zhang M, Cao Y, Zhang C, Fu Y, et al. Breeding history and candidate genes responsible for black skin of Xichuan black-bone chicken. *BMC Genomics.* 2020;21:1–15.
15. Fukamachi S, Shimada A, Shima A. Mutations in the gene encoding B, a novel transporter protein, reduce melanin content in medaka. *Nat Genet.* 2001;28:381–5.
16. Bin B-H, Bhin J, Yang SH, Shin M, Nam Y-J, Choi D-H, et al. Membrane-associated transporter protein (MATP) regulates melanosomal pH and influences tyrosinase activity. *PLoS One.* 2015;10:e0129273.
17. Westerdahl H, Mellinger S, Sigeman H, Kutschera VE, Proux-Wéra E, Lundberg M, et al. The genomic architecture of the passerine MHC region: High repeat content and contrasting evolutionary histories of single copy and tandemly duplicated MHC genes. *Mol Ecol Resour.* 2022;22(6):2379–95.
18. Westerdahl H. Passerine MHC: genetic variation and disease resistance in the wild. *J Ornithol.* 2007;148:469–77.
19. Kaufman J, Milne S, Göbel TW, Walker BA, Jacob JP, Auffray C, et al. The chicken B locus is a minimal essential major histocompatibility complex. *Nature.* 1999;401:923–5.
20. Kaufman J, Völk H, Wallny HJ. A "minimal essential Mhc" and an "unrecognized Mhc": two extremes in selection for polymorphism. *Immunol Rev.* 1995;143:63–88.
21. Wei L, Jiao P, Yuan R, Song Y, Cui P, Guo X, et al. Goose Toll-like receptor 7 (TLR7), myeloid differentiation factor 88 (MyD88) and antiviral molecules involved in anti-H5N1 highly pathogenic avian influenza virus response. *Vet Immunol Immunopathol.* 2013;153:99–106.
22. Velová H, Gutowska-Ding MW, Burt DW, Vinkler M. Toll-like receptor evolution in birds: gene duplication, pseudogenization, and diversifying selection. *Mol Biol Evol.* 2018;35:2170–84.
23. Chen S, Cheng A, Wang M. Innate sensing of viruses by pattern recognition receptors in birds. *Vet Res.* 2013;44:1–12.
24. Davis RL, Choi G, Kuiken T, Quéré P, Trapp S, Short KR, et al. The culture of primary duck endothelial cells for the study of avian influenza. *BMC Microbiol.* 2018;18:1–9.
25. Tong ZWM, Karawita AC, Kern C, Zhou H, Sinclair JE, Yan L, et al. Primary chicken and duck endothelial cells display a differential response to infection with highly pathogenic avian influenza virus. *Genes.* 2021;12:901.
26. Arthur JSC, Ley SC. Mitogen-activated protein kinases in innate immunity. *Nat Rev Immunol.* 2013;13:679–92.
27. Li D, Sun G, Zhang M, Cao Y, Zhang C, Fu Y, et al. Breeding history and candidate genes responsible for black skin of Xichuan black-bone chicken. *BMC Genomics.* 2020;21:511.
28. Branicki W, Brudnik U, Draus-Barini J, Kupiec T, Wojas-Pelc A. Association of the SLC45A2 gene with physiological human hair colour variation. *J Hum Genet.* 2008;53:966–71.
29. Tóth L, Fábos B, Farkas K, Sulák A, Tripolszki K, Széll M, et al. Identification of two novel mutations in the SLC45A2 gene in a Hungarian pedigree affected by unusual OCA type 4. *BMC Med Genet.* 2017;18:27.
30. Barber MRW, Aldridge JR, Fleming-Canepa X, Wang Y-D, Webster RG, Magor KE. Identification of avian RIG-I responsive genes during influenza infection. *Mol Immunol.* 2013;54:89–97.
31. Barber MRW, Aldridge JR, Webster RG, Magor KE. Association of RIG-I with innate immunity of ducks to influenza. *Proc Natl Acad Sci.* 2010;107:5913–8.
32. Smith J, Smith N, Yu L, Paton IR, Gutowska MW, Forrest HL, et al. A comparative analysis of host responses to avian influenza infection in ducks and chickens highlights a role for the interferon-induced transmembrane proteins in viral resistance. *BMC Genomics.* 2015;16:574.
33. Blyth GAD, Chan WF, Webster RG, Magor KE. Duck Interferon-Inducible Transmembrane Protein 3 Mediates Restriction of Influenza Viruses. *J Virol.* 2016;90:103–16.
34. Stephan T, Burgess SM, Cheng H, Danko CG, Gill CA, Jarvis ED, et al. Darwinian genomics and diversity in the tree of life. *Proc Natl Acad Sci.* 2022;119:e2115644119.
35. Wille M, Holmes EC. Wild birds as reservoirs for diverse and abundant gamma-and deltacoronaviruses. *FEMS Microbiol Rev.* 2020;44:631–44.
36. Jiang F, Miao Y, Liang W, Ye H, Liu H, Liu B. The complete mitochondrial genomes of the whistling duck (*Dendrocygna javanica*) and black swan (*Cygnus atratus*): dating evolutionary divergence in Galloanserae. *Mol Biol Rep.* 2010;37:3001–15.
37. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 2016;3:95–8.

38. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017;356:92–5.
39. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst*. 2018;6:256–258.e251.
40. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience*. 2020;9:giaa094.
41. Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics*. 2017;18:1–11.
42. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol*. 2019;15:e1007273.
43. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*. 2013;29:1341–2.
44. Alexa A, Rahnenfuhrer J. topGO: enrichment analysis for gene ontology. R Package Version. 2010;2:2010.
45. He K, Minias P, Dunn PO. Long-read genome assemblies reveal extraordinary variation in the number and structure of MHC loci in birds. *Genome Biol Evol*. 2021;13:evaa270.
46. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:1–21.
47. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 2009;4:1184–91.
48. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30:1236–40.
49. Walter W, Sánchez-Cabo F, Ricote M. GOrplot: an R package for visually combining expression data with functional analysis. *Bioinformatics*. 2015;31:2912–4.
50. Karawita, Anjana. Black swan genome assembly pipeline. Zenodo. <https://doi.org/10.5281/zenodo.7272708>.
51. Karawita, Anjana. MHC class annotation pipeline. Zenodo. <https://doi.org/10.5281/zenodo.7272715>.
52. Karawita, Anjana. Genome annotation pipeline. Zenodo. <https://doi.org/10.5281/zenodo.7272717>.
53. Karawita, Anjana. RNAseq data analysis. Zenodo. <https://doi.org/10.5281/zenodo.7272720>.
54. Karawita, Anjana. Gene family evolution. Zenodo. <https://doi.org/10.5281/zenodo.7272722>.
55. Karawita, A.C., Bruxner, T.J.C., Cheng, Y., Burt, D.W. and Short, K.R., Australian Black swan genome PacBio Assembly Version 1.0, Datasets. NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA640810/>.
56. Karawita, A.C., Dudchenko, O., Aiden, E., Kaur, P., Burt, D., Cheng, Y., Bruxner, T. and Short, K. CAttr\_DNAZoo\_HiC\_assembly, Datasets, NCBI. [https://www.ncbi.nlm.nih.gov/genome/9057?genome\\_assembly\\_id=1951720](https://www.ncbi.nlm.nih.gov/genome/9057?genome_assembly_id=1951720).
57. Kraus, R., Fedrigo, O., Formenti, G., Mountcastle, J., Chow, W., Collins, J., Howe, K., Rhie, A., Karawita, A., Short, K. and Jarvis, E.D. *Cygnus olor* (Mute swan) genome, bCygOlo1, primary haplotype, v2. Datasets. NCBI. [https://www.ncbi.nlm.nih.gov/assembly/GCF\\_009769625.2](https://www.ncbi.nlm.nih.gov/assembly/GCF_009769625.2).
58. Karawita, A, Short, K, Black swan ISO-seq data, Datasets, NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB44719>.
59. Karawita, A, Tolf, C., Waldenström, J., Short, K, Mute swan ISO-seq data, Datasets, UQ espace, <https://doi.org/10.48610/afb4f31>.
60. Karawita, A.C., Bruxner, T.J.C., Cheng, Y., Burt, D.W. and Short, K.R., Australian Black swan genome PacBio CLR, Datasets. NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/PRJEB44719>.
61. Kraus, R., Fedrigo, O., Formenti, G., Mountcastle, J., Chow, W., Collins, J., Howe, K., Rhie, A., Karawita, A., Short, K. and Jarvis, E.D. Mute swan raw read data (PacBio CLR/Arima Hi-C Illumina, 10x Linked data/Bionano data), Datasets. [https://vgp.github.io/genomeark/Cygnus\\_olor/](https://vgp.github.io/genomeark/Cygnus_olor/).
62. Karawita, A, Short, K, RNAseq data – Black swan endothelial cells, Datasets, NCBI. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB44719>.
63. Karawita, A, Short, K, RNAseq data – Avian endothelial cells, Datasets, UQ eSpace. <https://doi.org/10.48610/58a568b>.
64. DNA zoo – Baylor College of Medicine, Hi-C of *Cygnus atratus* DNA Zoo Sample2728D. Datasets. NCBI. [https://www.ncbi.nlm.nih.gov/sra?LinkName=biosample\\_sra&from\\_uid=21582233](https://www.ncbi.nlm.nih.gov/sra?LinkName=biosample_sra&from_uid=21582233).
65. Karawita, A, Short, K, Short-read next generation DNA sequencing of Black swan AKBS03. Datasets, UQ eSpace. <https://doi.org/10.48610/f6ba161>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.