

De novo genome assembly of the Indian Blue Peacock (*Pavo cristatus*), from Oxford Nanopore and Illumina sequencing

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00280R3
Full Title:	De novo genome assembly of the Indian Blue Peacock (<i>Pavo cristatus</i>), from Oxford Nanopore and Illumina sequencing
Article Type:	Data Note
Funding Information:	
Abstract:	<p>Background <i>Pavo cristatus</i>, the Indian peafowl are located in natural habitats of South Asia. The male blue peacock bird is known for its elegance, majestic looks and beauty. Since prehistoric times they have been described in Indian culture and has been adopted as the national bird of India. In this study, we present the first draft genome sequence of the peacock using Illumina and Oxford Nanopore technologies (ONT).</p> <p>Findings ONT sequencing resulted in approximately 2.3-fold sequencing coverage, whereas Illumina generated 150 bp paired-end sequence data at 284.6-fold sequencing coverage from five libraries. Subsequently, we generated de-novo genome assembly of the peacock genome with a 0.915 Gigabases (Gb) with a scaffold N50 of 0.23 Megabases (Mb). We also predicted that the peacock genome contains 23,153 protein-coding genes and 75.3 Mb (7.33%) of repetitive sequences.</p> <p>Conclusions We report a high-quality genome assembly of the peacock using a hybrid assembly generated from Illumina and ONT sequencing platforms. Long read chemistry generated from ONT was found to be useful in addressing challenges related to de-novo assembly particularly at regions containing repetitive sequences that span longer than the read length, and which cannot be resolved using only short-read-based assembly. The contig assembly on the short reads from Illumina resulted in an N50 of 1639 bases, whereas using 2.3x coverage from ONT increased the N50 by nine fold to 14,749 bases. The initial contig assembly based on Illumina sequencing reads alone resulted in total of 685,241 contigs. Further scaffolding on assembled contigs using both Illumina and ONT sequencing reads resulted in a final assembly having 15,025 super scaffolds with a N50 of about 0.23 Mb. The completeness of our genome assembly was verified with the fact that 95% of proteins predicted by homology were matched to those submitted in public repository. Further in concordance with other phylogenetic studies, the avian phylogeny on the conserved genes showed <i>P. cristatus</i> being closest with <i>Gallus gallus</i> followed by <i>Meleagris gallopavo</i> and <i>Anas platyrhynchos</i>.</p>
Corresponding Author:	Subhradip Karmakar, PhD All India Institute of Medical Sciences New Delhi, Delhi INDIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	All India Institute of Medical Sciences
Corresponding Author's Secondary Institution:	
First Author:	Ruby Dhar
First Author Secondary Information:	
Order of Authors:	Ruby Dhar Ashikh Seethy

	Karthikeyan Pethusamy
	Vishwajeet Rohil
	Sunil Singh
	Kakali Purkayastha
	Sandeep Goswami
	Rakesh Singh
	Indrani Mukherjee
	Ankita Raj
	Tryambak Srivastava
	Sovon Acharya
	Balaji Rajashekhar, Ph.D
	Subhradip Karmakar, PhD
Order of Authors Secondary Information:	
Response to Reviewers:	<p>Date: 6 Feb. 2019</p> <p>Dear Dr. Scott, Thank you for suggesting improvements in the article. We have addressed all the reviewer’s comments. Below are point-by-point response for the corrections raised by the reviewers for the manuscript titled “De-novo genome assembly of the Indian Blue Peacock (Pavo cristatus), from Oxford Nanopore and Illumina sequencing”</p> <p>Page 2</p> <p>1.Line 6-9: Remove the two sentences: “The findings from avian genomics ...” to “next generation sequencing technologies” Reply: The two sentences from “The findings from avian genomics ...” to “next generation sequencing technologies” are removed.</p> <p>2.Lines 13: Recommend remove the sentence: “For the first time in avian genomics”, since it is redundant with the sentence immediate preceding it. Reply: The sentence starting from “For the first time in avian genomics” is removed.</p> <p>3.Line 19: Recommend change “75,315,566” bp to “75.3 Mb” Reply: The text is changed from “75,315,566” bp to “75.3 Mb”</p> <p>Page 3</p> <p>4.Line 6: Replace “reliability” with “completeness” Reply: The word is replaced from “reliability” with “completeness”</p> <p>5.Line 8-10: This sentence should be re-written to comment on the significance of this result. Is this expected or unexpected? Reply: The sentence is modified as “Further in concordance with other phylogenetic studies, the avian phylogeny on the conserved genes showed P. cristatus being closest with Gallus gallus followed by Meleagris gallopavo and Anas platyrhynchos”</p> <p>Page 4</p> <p>6.Line 5: Need citation(s) for peacock references in ancient Indian literatures. Probably secondary scholarly works and not primary (ancient Indian) literature references. Reply: Reference Kadgoankar, 1993 have been added here.</p> <p>7.Line 14: In the version that I reviewing there is an empty line between the paragraph that ends, “within the orthologous regions [8]”, and the paragraph that begins, “Despite the wealth of information”. Reply: The sentence got removed unintentionally in our previous version, we have included the missing sentence in this version.</p>

8.Line 20: Replace “construction” with “sequencing” or “assembly”
Reply: The word “construction” is replaced with “assembly”

Page 5

9.Line 10: Change to “A ReliaPrep™”.
Reply: changed to “A ReliaPrep™”

Page 6

10.Line 21: Remove “accurately”
Reply: The word “accurately” is removed

Page 7

11.Line 1-5: This paragraph needs to be rewritten in the past tense to match the rest of the section. “had” instead of “has”, “was” instead of “is”.
Reply: In the paragraph lines 1-5 the sentences are modified in past tense.

Page 8

12.Line 13: “The library mix”
Reply: The sentence modified as “The library mix”

13.Line 14: “The eluted library”
Reply: The sentence modified as “The eluted library”

14.Line 15: “The whole genome library was prepared” or “The whole genome libraries were prepared”
Reply: The sentences is modified as “The whole genome libraries were prepared”

15.Line 22: “bcl2fastq (Illumina)”
Reply: The text is modified to bcl2fastq (Illumina)

16.Line 24: The citation, “(Andrews, S., 2010)” is in the wrong style.
Reply: “(Andrews, S., 2010)” is changed to citation [10] and in reference the following citation is added “Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.”

Page 9

17.The phrase “Oxford Nanopore” needs to be replaced with “ONT” or whichever abbreviation the authors choose to use.
Reply: the text “Oxford Nanopore” is changed to “ONT”

Page 10

18.Line 3: “hard masked with the G. gallus repeat library using Repeatmasker (www.repeatmasker.org/).” Proper citation for Repeatmasker is found here: <http://repeatmasker.org/faq.html#faq3>.
Reply: The reference “Smit, AFA, Hubley, R and Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>” is modified

19.Line 6: Replace “obtained” with “identified”
Reply: The word “obtained” is replaced with “identified”

20.Line 10: Proper citations for Augustus found here:

<http://augustus.gobics.de/references>

Reply: the reference is replaced with “[Stanke, M., Diekhans, M., Baertsch, R. and Hausler, D., 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5), pp.637-644. (<http://augustus.gobics.de/>)”

21.Line 18: Proper citation for JGI portal <https://genome.jgi.doe.gov/pages/citeUs.jsf>

Reply: the reference is replaced with “Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I.V. and Dubchak, I., 2013. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic acids research*, 42(D1), pp.D26-D31.”

22.Line 24: The URLs for Pfam-A database and Pfam scan tools are out of date.

Proper citation for Pfam is at the bottom of this page: <http://pfam.xfam.org/>.

Reply: the reference is replaced with “El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A. and Sonnhammer, E.L.L., 2018. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1), pp.D427-D432.”

Page 11

23.Line 6: There are two papers at the top of this web page

(<http://avian.genomics.cn/en/jsp/database.shtml>) that should be cited as sources for this data.

Reply: Replaced as “avian phylogenomics project [”

And the following citation added

“Zhang, G., Li, B., Li, C., Gilbert, M.T.P., Jarvis, E.D. and Wang, J., 2014. Comparative genomic data of the Avian Phylogenomics Project. *GigaScience*, 3(1), p.26.
Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y., Faircloth, B.C., Nabholz, B., Howard, J.T. and Suh, A., 2015. Phylogenomic analyses data of the avian phylogenomics project. *GigaScience*, 4(1), p.4.”

24.Lines 17,18,20,21 There are papers that should be cited for clustal, Gblock, Phylip and IQ-tree. I don't see them cited here. The papers are listed on the tools' websites.

Reply: Following citation were added,

“Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R. and Thompson, J.D., 2007. Clustal W and Clustal X version 2.0. *bioinformatics*, 23(21), pp.2947-2948.

Talavera, G. and Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, 56(4), pp.564-577.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164-166.

L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh, 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol.*, 32:268-274.”

Page 12

25.Line 8: Papers to cite when using LAST: <http://last.cbrc.jp/doc/last-papers.html>

Reply: Following citation was added, “Frith, M.C. and Kawaguchi, R., 2015. Split-alignment of genomes finds orthologies more accurately. *Genome biology*, 16(1), p.106.” is added

26.Line 12: Paper to cite when using Circos: If you are using Circos, please cite us:

Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics. *Genome Res* (2009) 19:1639-1645 | download citation

Reply: Following citation was added “Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A., 2009. Circos: an

information aesthetic for comparative genomics. Genome research, 19(9):1639-45.”

Page 13

27.Line 4: Remove “(mega base)”.

Reply: The text “(mega base)” is removed.

28.Line 9: Change “1.02 GB (giga base)” to “1.02 Gb”, remove the “(giga base)”.

Reply: The text “1.02 GB (giga base)” is changed to “1.02 Gb”

29.Line 11: Change “>=1 Mbp” to “>= 1 Mb”.

Reply: The text “>=1 Mbp” is changed to “>= 1 Mb”.

30.Line 16: In accordance with my prior comment, change “75,315,566 bp” to “75 Mb”.

Reply: The text is changed from “75,315,566 bp” to “75 Mb”

31.Line 17: Change 56,511,635 bp to “56 Mb”.

Reply: The text is changed from “56,511,635 bp” to “56 Mb”.

Page 14

32.Line 14: Change to “The detailed annotations”.

Reply: The text is changed to “The detailed annotations”

33.Line 23: “humans” not “Humans”

Reply: The word is changed to “human”

Page 15

Page 16

34.Line 3: “Zn” to “zinc”

Reply: The text is changed from “Zn” to “zinc”

35.Line 16: “de-novo”, is italicized and not hyphenated elsewhere in the manuscript, except for Line 19 of Page 4. Needs to be consistent. Probably should use “de novo”.

Reply: We have used “de-novo” instead of “de novo” in the manuscript.

36.Line 16-26: This whole paragraph needs some citations, especially for the claim about different technologies improving genome assemblies. Even though you have data that supports this claim and demonstrate ONT’s use in bird genomes for the first time, this idea has been discussed before and has been the basis of at least one genome assembler in the past (eg AllPaths LG) and the topic of several reviews (see Metzker, Nature Reviews Genetics, 2010).

Reply: Citations relevant to the sentences have been included in this paragraph.

Page 17

37.Line 15: “95% homology” implies that 95% of the nucleotides match between sequences, which I don’t is demonstrated in this figure. If the claim is that 95% of the annotated (or predicted; it’s not clear in the text which set of peacock genes is meant) had a match then that should be made clearer. Looking in the abstract, I see the sentence, “The reliability of our genome assembly was verified with the fact that 95% of proteins predicted by homology were matched to those submitted in public repository.” That claim matches more closely the message communicated by Figure 7, so I would re-write this sentence to match the abstract.

Reply: The sentence is now modified as “The confidence on the predicted peacock proteins got strengthened when about 95% of them showed significant homology to various genomic features from different databases (Fig. 7).”

	<p>Page 18 38.Line 8: "Figures, Gene ontology and annotations". This sentence fragment needs to be re-written. Reply: The section is rewritten as "Additional figures included are the Peacock, Chicken and Turkey proteins showing similarity to Pfam domains, top ten Gene ontology annotations in Biological process; Cellular component and Molecular function from the Peacock proteins, and Peacock homologous proteins in humans."</p> <p>Page 21 39.Lines 16-20: This figure caption needs to be shortened, since it is partially a re-write of this section from methods. Could be re-written as "Circular image of the assembled peacock genome aligned against the G. gallus genome. The right side of the image represents the reference chicken genome and left side of the image represents the peacock genome." Reply: The text has been shortened as suggested.</p> <p>Page 22 40.Line 13: Reference 5 here, Kadgoankar, 1993, is not used in the text and provides the information that I said was missing on Page 4, Line 5. Reply: The missing citation has been added.</p> <p>41.In the reference list in general, "p" or "pp" is missing from several references (reference 12, 13, 17, maybe others) and should be made consistent throughout. Reply: We have used reference manager software to make all references consistent.</p>
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>	Yes
<p>Resources</p> <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</p>	Yes

<p>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>

[Click here to view linked References](#)

1 **De novo genome assembly of the Indian Blue Peacock**
2 **(*Pavo cristatus*), from Oxford Nanopore and Illumina**
3 **sequencing**

4
5
6 **Authors:** Ruby Dhar¹, Ashikh Seethy¹, Karthikeyan Pethusamy¹, Vishwajeet Rohil², Sunil
7 Singh¹, Kakali Purkayastha², Sandeep Goswami¹, Rakesh Singh³, Indrani Mukherjee¹, Ankita
8 Raj¹, Tryambak Srivastava¹, Sovon Acharya¹, Balaji Rajashekhar^{4,5,*} and Subhradip
9 Karmakar^{1,*}

10
11 **Affiliation:** ¹Department of Biochemistry, AIIMS, New Delhi, India. ²Vallabhbhai Patel Chest
12 Institute (VPCI), New Delhi, India. ³Kanpur Zoo, Kanpur, India. ⁴Celixa, Bangalore, India-
13 and ⁵Institute of Computer Science, University of Tartu, 50409 Tartu, Estonia

14
15 *Corresponding Authors email: balaji@ut.ee, subhradip.k@aiims.edu

16
17
18
19 **Running Title:** De novo Genome Assembly of the Peacock Bird

20
21 **Key words:** Peacock, *Pavo cristatus*, Indian National Bird, Genome Assembly, Oxford
22 Nanopore.

23
24 **Abstract**

1 **Background**

2 *Pavo cristatus*, the Indian peafowl are located in natural habitats of South Asia. The male
3 blue peacock bird is known for its elegance, majestic looks and beauty. Since prehistoric
4 times they have been described in Indian culture and has been adopted as the national bird of
5 India. In this study, we present the first draft genome sequence of the peacock using Illumina
6 and Oxford Nanopore technologies (ONT).

8 **Findings**

9 ONT sequencing resulted in approximately 2.3-fold sequencing coverage, whereas Illumina
10 generated 150 bp paired-end sequence data at 284.6-fold sequencing coverage from five
11 libraries. Subsequently, we generated de novo genome assembly of the peacock genome with
12 a 0.915 Gigabases (Gb) with a scaffold N50 of 0.23 Megabases (Mb). We also predicted that
13 the peacock genome contains 23,153 protein-coding genes and 75.3 Mb (7.33%) of repetitive
14 sequences.

16 **Conclusions**

17 We report a high-quality genome assembly of the peacock using a hybrid assembly generated
18 from Illumina and ONT sequencing platforms. Long read chemistry generated from ONT
19 was found to be useful in addressing challenges related to de novo assembly particularly at
20 regions containing repetitive sequences that span longer than the read length, and which
21 cannot be resolved using only short-read-based assembly. The contig assembly on the short
22 reads from Illumina resulted in an N50 of 1639 bases, whereas using 2.3x coverage from
23 ONT increased the N50 by nine fold to 14,749 bases. The initial contig assembly based on
24 Illumina sequencing reads alone resulted in total of 685,241 contigs. Further scaffolding on
25 assembled contigs using both Illumina and ONT sequencing reads resulted in a final

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 assembly having 15,025 super scaffolds with a N50 of about 0.23 Mb. The completeness of
2 our genome assembly was verified with the fact that 95% of proteins predicted by homology
3 were matched to those submitted in public repository. Further in concordance with other
4 phylogenetic studies, the avian phylogeny on the conserved genes showed *P. cristatus* being
5 closest with *Gallus gallus* followed by *Meleagris gallopavo* and *Anas platyrhynchos*.

6

1 Introduction

2 *Pavo cristatus* commonly known as the Indian blue peafowl are native to South Asian
3 countries. Apart from the wild, they are usually found as exhibits in park and zoo, besides
4 being raised for breeding and conservation purposes [1, 2] (Fig. 1). The peacock has been
5 widely referred in ancient Indian literatures [3]. They have been found to be closely
6 associated with the life and culture of the people from South East Asia, symbolizing beauty,
7 love, grace and pride [4, 5]. Owing to these, the peacock obtained the status as the National
8 Bird of India in 1963.

9 Genome sequencing of the avian model organism *Gallus gallus* (red junglefowl the chicken)
10 [6], as well as variety of other avian species [7] have provided a novel perspective on
11 vertebrate genome evolution. This enabled us to understand the genome structure better and
12 annotate the mammalian genome. Genome studies of *G. gallus* with respect to the human
13 have revealed an extremely high level of conservation within the orthologous regions [8] thus
14 promising of being a good candidate for studies on developmental biology, immunology and
15 vertebrate genome architecture [9, 10].

16
17 Despite the wealth of information from the existing avian genome sequencing projects, it is
18 still important to sequence genome of other new species to add value, both into avian and
19 vertebrate genomics. For the first time in avian genomics, Oxford Nanopore technology
20 (ONT or Nanopore) has been used to sequence a bird genome presented in this study. Long
21 reads have been helpful during the de novo assembly of the genome especially in the GC rich
22 repeat regions which invariably poses serious challenges in genome assembly. Comparative
23 genomics with other birds will help in understanding the uniqueness of peacock genome,
24 development of this species, complex plumage pigmentation, sexual dimorphism and its
25 evolutionary relationships with other birds. The characterization of the genes and association

1 with specific function will provide better understanding of the peafowl species. The protein
2 comparisons among the peacock, chicken and *Meleagris gallopavo* (domestic turkey) will
3 reveal conserved domains and functional annotations that are common and absent among
4 these species.

6 **Materials and methods**

7 **Sample collection and extraction of DNA**

8 The whole blood of male peacock was collected from Kanpur zoo, India after obtaining the
9 necessary ethical and institutional approval. Approximately, 20 µl of proteinase K (PK)
10 solution was taken into a 1.5 ml microcentrifuge tube, 200 µl of blood was added and briefly
11 mixed. Furthermore, 200 µl of cell lysis buffer was added to the tube, mixed by vortexing for
12 10 seconds, incubated at 56°C for 10 minutes. ReliaPrep™ Binding Column was placed into
13 an empty collection tube. Furthermore, 250 µl of Binding Buffer (BBA) was added to the
14 tube, and mixed by vortexing for 10 seconds with a vortex mixer. Contents of the tube were
15 added to the A ReliaPrep™ binding column, capped and placed in a refrigerated
16 microcentrifuge. These were then centrifuged for 1 minute at maximum speed and flow
17 through was discarded. Binding column was placed into a fresh collection tube. In addition,
18 500 µl of column wash solution was added to the column and centrifuged for 3 minutes at
19 maximum speed; flow through was again discarded. Column washing is repeated thrice.
20 Columns were then placed in a nuclease free clean 1.5 ml microcentrifuge tube. Furthermore,
21 100 µl of Nuclease-Free Water was then added to the column and centrifuged for an
22 additional 1 minute at maximum speed. Column was discarded and elute was saved. The
23 concentration and purity of the extracted DNA was evaluated using Nanodrop
24 Spectrophotometer (Thermo Scientific) and Qubit flurometer and integrity was checked on a

1 0.8% agarose gel. The DNA sample was aliquoted for library preparation on two different
2 platforms: Illumina HiSeq4000 and Oxford Nanopore Technologies (ONT).

3 4 **Library preparation and sequencing**

5 **A. Paired-End library preparation and sequencing**

6 Whole genome sequencing (WGS) libraries were prepared with Illumina-compatible
7 NEXTflex DNA sequencing kit (BIOO Scientific, Austin, TX, USA). Approximately, 1 µg of
8 genomic DNA was sheared using Covaris S2 sonicator (Covaris, Woburn, MA, USA) to
9 generate approximate fragment size distribution from 300 - 600 basepair (bp). The fragment
10 size distribution was checked on Agilent 2200 Tape Station with D1000 DNA screen tapes
11 and reagents (Agilent Technologies, Palo Alto, CA, USA) and subsequently purified using
12 HighPrep magnetic beads (MagBio Genomics Inc, USA). The purified fragments were end-
13 repaired, adenylated and ligated to Illumina multiplex barcode adaptors as per NEXTflex
14 DNA sequencing kit protocol (BIOO Scientific, Austin, TX, USA).

15
16 The adapter-ligated DNA was purified with HighPrep beads (MagBio Genomics, Inc,
17 Gaithersburg, MD, USA) and then size selected on 2% low melting agarose gel and cleaned
18 using MinElute column (QIAGEN). The resultant fragments were amplified for 10 cycles of
19 PCR using Illumina-compatible primers provided in the NEXTFlex DNA sequencing kit. The
20 final PCR product (sequencing library) was purified with HighPrep beads, followed by
21 library quality control check. The Illumina-compatible sequencing library was initially
22 quantified by Qubit fluorometer (Thermo Fisher Scientific, MA, USA) and its fragment size
23 distribution was analyzed on Agilent TapeStation. Finally, the sequencing library was
24 quantified by quantitative PCR using Kapa Library Quantification Kit (Kapa Biosystems,

1 Wilmington, MA, USA). The qPCR-quantified library was subjected to sequencing on an
2 Illumina sequencer for 150 bp paired-end chemistry.

3
4 The Illumina-compatible sequencing library for the samples had a fragment size range
5 between 275 - 425 bp for Paired-End Short Insert (PE-SI) and 350 - 650 bp for Paired-End
6 Long Insert (PE-LI). As the combined adapter size was approximately 120 bp, the effective
7 user-defined insert size was 155 - 305 bp and 230 - 530 bp for PE-SI and PE-LI, respectively.
8 Libraries were sequenced in Illumina HiSeq platform with 150 PE chemistry.

9 10 **B. Mate-Pair library preparation and sequencing**

11 Mate Pair sequencing library was prepared with Illumina-compatible Nextera Mate Pair
12 Sample Preparation Kit (Illumina Inc., Austin, TX, USA). Approximately, 4 ug of genomic
13 DNA was simultaneously fragmented and tagged with Mate Pair adapters in a transposon-
14 based tagmentation step. Tagmented DNA was then purified using AMPure XP Magnetic
15 beads (Beckman Coulter Life Sciences, Indianapolis, IN, USA) followed by strand
16 displacement to fill gaps in the tagmented DNA. Strand displaced DNA was further purified
17 with AMPure XP beads before size-selecting the 3 - 5 kilobases (Kb), 5 - 7 Kb & 7 - 10 Kb
18 fragments on low melting agarose gel. The fragments were circularized in an overnight blunt-
19 end intra-molecular ligation step, which will result in circularization of DNA with the insert
20 mate pair adapter junction. The circularized DNA was sheared using Covaris S220 sonicator
21 (Covaris, Woburn, MA, USA) to generate approximate fragment size distribution from 300 -
22 1000 bp. The sheared DNA was purified to collect the mate pair junction positive fragments
23 using Dynabeads M-280 Streptavidin Magnetic beads (Thermo Fisher Scientific, Waltham,
24 MA, USA). The purified fragments were end-repaired, adenylated and ligated to Illumina
25 multiplex barcode adaptors as per Nextera Mate Pair Sample Preparation Kit protocol.

1
2 The adapter-ligated DNA was then amplified for 15 cycles of PCR using Illumina-compatible
3 primers. The final PCR product (sequencing library) was purified with AMPure XP beads,
4 followed by library quality control check. The Illumina compatible sequencing library was
5 initially quantified by Qubit fluorometer (Thermo Fisher Scientific, MA, USA), and its
6 fragment size distribution was analyzed on Agilent TapeStation. Finally, the sequencing
7 library was accurately quantified by quantitative PCR using Kapa Library Quantification Kit
8 (Kapa Biosystems, Wilmington, MA, USA). The qPCR quantified libraries were pooled in
9 equimolar amounts to create a final multiplexed library pool for sequencing on an Illumina
10 sequencer.

11 12 **C. Oxford Nanopore MinION library preparation and sequencing**

13 Genomic DNA (1.5µg) was end-repaired (NEBnext ultra II end repair kit, New England
14 Biolabs, MA, USA), cleaned up with 1x Ampure beads (Beckmann Coulter, USA). Adapter
15 ligations were performed for 20 minutes using NEB blunt/TA ligase (New England Biolabs,
16 MA, USA). The library mix were cleaned up using 0.4X Ampure beads (Beckmann Coulter,
17 USA) and eluted in 25 µl of elution buffer. The eluted library was used for sequencing. The
18 whole genome libraries were prepared by using ligation sequencing SQK-LSK108 Oxford
19 Nanopore sequencing kit (ONT, Oxford, UK). Sequencing was performed on MinION Mk1b
20 (ONT, Oxford, UK) using SpotON flow cell (FLO-MIN106) in a 48 hour sequencing
21 protocol on MinKNOW (1.1.20 from ONT).

22 23 **Raw data quality control and processing**

24 **A. Illumina raw data quality control and processing**

1 The Illumina reads were de-multiplexed using bcl2fastq (Illumina). The Illumina generated
2 raw data for genomic libraries was quality checked using FastQC
3 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) [11]. The paired-end Illumina
4 reads were processed for clipping the adapter and low-quality bases using customized script
5 which retains minimum 70% bases/reads with Phred score ($Q \geq 30$ in each base position) with
6 a read length of 50 bp. The MP libraries were trimmed for adapter and low-quality base
7 trimming from the 3'-end using PLATANUS internal trimmer
8 (<http://platanus.bio.titech.ac.jp/>) [12].

10 **B. Oxford Nanopore reads base calling and processing**

11 The raw data were then base-called with the cloud-based Metrichor workflow 2D Basecalling
12 plus Barcoding by Metrichor (V.2.43.1 from ONT,
13 <https://nanoporetech.com/products/metrichor>). The Oxford Nanopore reads were processed
14 using Poretools [13] for converting fast5 files to fasta format. For further quantification and
15 analysis the 2D reads or 1D high quality reads were selected for further assembly.

17 **De novo genome assembly and genome size estimation**

18 The quality checked Oxford Nanopore reads were error-corrected using Illumina PE reads.
19 For error-correction the Illumina PE-reads were aligned to the Nanopore reads by using
20 BWA aligner [14]. The paired-end reads were assembled using Abyss [15] followed by
21 contig extension using ONT reads using SSPACE-LongRead [16]. Super scaffolding of the
22 assembled scaffold was performed using SSPACE [17] and PLATANUS
23 (<http://platanus.bio.titech.ac.jp/>) using the Oxford Nanopore and Matepair data. Final draft
24 genome resulted after gap closure by GAPCLOSER
25 (<http://sourceforge.net/projects/soapdenovo2/files/GapCloser/>) and PLATANUS gap_close

1 tool (<http://platanus.bio.titech.ac.jp/>) using Illumina data. The genome size was estimated
2 using a k-mer distribution plot using JELLYFISH [18]. The assembly and annotation
3 workflow has been represented in Figure 2.

5 **Identification of repetitive elements and SSR markers**

6 Repetitive elements, retrotransposons and DNA transposons were identified in the draft
7 genome and was hard masked by using reference genomic repeats of *G. gallus* using
8 Repeatmasker tool [19]. Final assembled scaffolds were analysed for Simple Sequence
9 Repeats (SSR) identification. SSRs like the di, tri, tetra, penta and hexa-nucleotide repeats in
10 the genome were identified using MISA (Version 1.0.0) ([http://pgrc.ipk-](http://pgrc.ipk-gatersleben.de/misa/)
11 [gatersleben.de/misa/](http://pgrc.ipk-gatersleben.de/misa/)).

13 **Annotation of the draft genome**

14 Gene models were predicted on a hard-masked draft genome using AUGUSTUS [20] with *G.*
15 *gallus* as a reference model. The predicted proteins were annotated by using BLASTP [21]
16 against the NCBI NR (non-redundant) database with default parameters at E-value cutoff of
17 1E-5.

18 The predicted proteins were searched against the KEGG-KAAS server
19 (<http://www.genome.jp/tools/kaas/>) for pathway analysis [22]. *G. gallus*, *M. gallopavo*,
20 *Taeniopygia guttata* (zebra finch), *Falco peregrinus* (peregrine falcon) were used as
21 reference organism for pathway identification. The EuKaryotic Orthologous Groups (KOGs)
22 [23] were predicted using homology-based approach.

24 **Prediction of protein domains**

1 Predicted proteins from peacock, chicken and turkey with sequence length greater than 100
2 amino acids were considered for protein domain analysis. All the protein sequences from
3 each organism were searched against Pfam-A database using Pfam scan [24] for protein
4 domain identification.

6 **Identification of avian protein families**

7 A total of 748,544 protein sequences from 49 avian species (including peacock proteins from
8 this study) and others were downloaded from avian phylogenomics project [25, 26].
9 Sequences greater than 100 amino acids from all the avian genomes were selected and
10 concatenated to a single fasta file. These sequences were clustered using CD-HIT [27] with
11 70% alignment coverage for the shorter sequence with a length difference cutoff of 0.7. The
12 single copy gene family orthologs present across all avian species and not clustered peacock
13 proteins were annotated.

15 **Phylogenetic tree construction**

16 For phylogenetic tree construction we considered single copy gene clusters present as single
17 copy in all the avian species. These protein sequences from each species were concatenated
18 and were further aligned by multiple sequence alignment tool Clustalw [28]. The poorly
19 aligned positions and divergent regions were removed using Gblock tool [29]. The fasta
20 format sequences were converted to phylip format using Phylip tool [30]. Phylogenetic trees
21 were constructed using IQ-TREE version 1.5.6 [31]. The parameters used for phylogenetic
22 tree construction were ultrafast bootstrap (UFBoot, using the `-bb` option of 1000 replicates),
23 and a standard substitution model (`-st AA -m TEST`) and `alrt 1000 -nt AUTO` was given for
24 tree generation. The generated trees from IQ-TREE tool were visualized using Figtree
25 (<http://tree.bio.ed.ac.uk/software/figtree/>) and the Branch-support values were recorded from

1 the output “.treefile”. The trees were modified for better visualization under Trees section an
2 increasing order nodes were applied.**Genome conservation analysis**
3 Draft chromosome visualizations were constructed by aligning the assembled peacock
4 genome against the *G. gallus* with the Chromosomer tool
5 (<https://github.com/gtamazian/chromosomer>). The reordered assembled genome was aligned
6 against the chicken genome using LAST aligner [32] with NEAR (finding short-and-strong
7 [near-identical] similarities) parameter allowing for substitution and gap frequencies, leading
8 to the identification of orthologs. These query-mapped regions were filtered with a greater
9 than 1% of the maximum length for visualization using Circos [33].

11 **Results**

12 **Genome sequencing assessment**

13 A total of five libraries from Illumina HiSeq technology of 150 bp paired-end were
14 generated. The short-insert reads of 489,114,747 accounted to genome coverage of 146.7X
15 and long-insert reads of 302,884,819 sequences was about 90.9X coverage with a total
16 coverage of 237.6X. Sequencing of three mate-pairs of 3-5 Kb, 5-7 Kb of and 7-10 Kb
17 yielded 72,915,033, 47,440,144 and 36,464,628 reads, respectively with an approximate
18 coverage of 21.9X, 14.2X and 10.9X, respectively, with a grand total of 156 million mate-
19 pair reads of 47X coverage. Oxford Nanopore technology was used to generate 366,323 long
20 reads having of 2,398,560,283 bp with coverage of 2.3X. The complete genome sequencing
21 was generated to a depth of ~287X from both Illumina and Oxford Nanopore platform (Table
22 1). The coverage was based on the assumption that the peacock genome size of about 1 Gb.

24 **Genome assembly**

1 The first assembly was performed on Illumina reads with Abyss de novo assembler that
2 resulted in ~932 Mb of genome with an N50 of 1639 bp. The extension of the contigs were
3 performed with Oxford Nanopore reads, which generated scaffolds with N50 of 14,748 bp.
4 Super scaffolding of the assembled scaffold was performed using SSPACE and PLATANUS
5 with MP libraries that generated ~916 Mb genome with the N50 value of 168,140 bp. The
6 final gap closer was executed using GAPCLOSER program with MP and PE-LI libraries,
7 which generated a draft genome of 1.02 Gb. The draft genome assembly of *Pavo cristatus*
8 consists of 179,346 bp scaffolds, with a N50 of 189,886 bp with 37 scaffolds, having
9 sequence length ≥ 1 Mb. Contigs above 5000 bp have covered a genome of ~0.915 Mb with
10 N50 0.23 Mb. In the assembled genome there were ~0.4% of non-ATGC characters (Table
11 2).

13 **Repetitive genome elements and SSR markers**

14 A total of 75 Mb (7.33%) of the peacock genome was estimated to consist of repeat
15 sequences (Table S1). In the genome about 56 Mb (5.5%) of retrotransposons (class I) were
16 identified as the NON-LTR elements (LINEs (4.7%), SINEs (0.08%) and LTR elements
17 (0.72%). Subsequently, the DNA transposons (class II) of 7,277,390 bp (0.71%) and
18 unclassified elements of about 467,719 (0.05%) were identified (Table S1). The median
19 percentages of LINEs, SINEs, LTR, DNA, unknown and total masked bases of other avian
20 birds were 3.94, 0.11, 1.31, 0.22, 0.85 and 6.93, respectively (Table S2). A total of 399,493
21 SSRs were obtained from the peacock genome assembly. The largest fraction of SSRs
22 identified were mono-nucleotide (60.04%), followed by tetra-nucleotide (26%), di-nucleotide
23 (8.51%), tri-nucleotide (4.31%), penta-nucleotide (1.03%) and finally hexa-nucleotide
24 (0.13%). Among the SSRs identified, A (49.2%) and T (44.9%) accounted for 94.1% of the
25 mono-nucleotide repeats. AT (23.8%), TA (16.5%), TG (13.7%), AC (10.6%) and CA

1 (10.32%) accounted for 75% of the di-nucleotide repeats, whereas TTG (9.9%), AAT (9.6%),
2 AAC (9.4%), TTA (7.1%), ATT (4.5%), TAA (3.5%), CAA (3.1%) and GGA (2.69%)
3 accounted for 49.7% of the tri-nucleotide repeats (Table S3).

5 **Gene prediction and annotation**

6 A total of 23,153 proteins were predicted from the assembled draft peacock genome using
7 AUGUSTUS. Among them, 21,854 (94.4%) predicted proteins showed homology to other
8 sequences from the NCBI NR database (Fig. 3). The top four organisms where the peacock
9 proteins showed homology belonged to the *G. gallus* with 11,398 proteins, *M. gallopavo* with
10 4059 proteins, *Amazona aestiva* (blue-fronted Amazon parrot) with 1352 proteins and *Anas*
11 *platyrhynchos* (mallard the duck) with 849 proteins. The detailed annotations of all the
12 proteins are available in Table S4.

14 Gene Ontology (GO) descriptions were assigned for a total of 18,294 (79%) peacock
15 proteins. Among them, 14,489 proteins have molecular function; 11,678 have biological
16 process and 13,735 proteins have cellular component as functional categories (Table S4). A
17 total of 4091 (17.7%) peacock proteins were found to have pathway information from the
18 KEGG database (Table S5), whereas a total of 20,937 (88.1%) peacock proteins found a
19 similarity against the KOG annotations (Table S6). The peacock proteins when searched
20 against the human proteins showed gene family expansions (in cell morphogenesis, neuronal
21 projection and development and GTPases (Table S7 and Fig. S3).

23 **Analysis of avian protein families**

24 A total of 748,544 protein sequences from 49 avian species have 653,497 protein sequences
25 of length above 100 amino acids (Table S8A). Based on the level of identity CD-HIT

1 clustered all the proteins into a total of 114,121 gene clusters. Among them, 68 highly
2 homologous gene clusters were present as single copy in all the 49 avian species (Table S8B
3 and Table S8C). We also observed 13,860 protein clusters of peacock species not clustered
4 with other avian species (Table S8D).

6 **Phylogenetic analysis**

7 The phylogenetic analysis of 48 avian species and the peacock proteins showed clustering of
8 the *P. cristatus* species in a clade of *G. gallus*, *M. gallopavo*, *A. platyrhynchos*, *Tinamus*
9 *guttatus* (white-throated tinamou) and *Struthio camelus* (ostrich). This is the largest clade
10 with six species having a bootstrap support of a 100. In the aforementioned clade, except the
11 mallard species all belong to flightless or low flying birds. The bootstrap support between *P.*
12 *cristatus* and *G. gallus* were 96, followed by *M. gallopavo* 100 bootstrap support (Fig. 4).

14 **Comparison with other species and databases**

15 Predicted proteins from peacock, chicken and turkey when searched for the conserved Pfam
16 protein domains showed about 81% of the domains that were common among these three
17 species (Fig. 5, Table S9). In comparison with the total Pfam domains from all the three
18 species, 94%, 98.4% and 99.7% Pfam domains were present in peacock, chicken and turkey,
19 respectively. However, 255, 69 and 14 Pfam domains were absent among the species
20 comparisons, respectively (Table S9H).

21 There were 15,470 (78%), 12,794 (85%) and 11,745 (85%) of the peacock, chicken and
22 turkey proteins found to contain a match to Pfam domains, respectively (Table S9). The
23 domain comparisons among the species showed gene family expansions such as kinases, zinc
24 finger proteins, GTPases and others in either one of the species (Fig. 6). Commonly, a total of
25 9974 peacock proteins were found to have annotation in all the four databases NCBI-NR,

1 KOG, Pfam and GO (Fig. 7). The assembled peacock genome when reordered for pseudo
2 chromosomes generation against the masked 1.21 Gb chicken genome [34] showed a 597
3 MB reordered peacock genome (Fig. 8). There are around 60 different avian species that have
4 been sequenced by using various sequencing technologies (Table S10). The sequencing depth
5 varies from as low as 6x to maximum of 390x coverage. The result obtained from different
6 bioinformatics methods to assemble the sequencing data are measured as scaffold N50 that is,
7 from 30 Kb to 14 Mb.

10 **Conclusions**

11 A rapid surge in de novo genome sequence assembly of diverse species is seen in recent
12 years [35]. This is essentially driven largely due to an affordable cost per base sequencing
13 along with the development of smarter algorithms refined and equipped to handle large data
14 sets [36-38]. The challenge of newer genome analysis pipeline lies in generating assembly
15 with lower contig numbers and longer contigs per genome. To achieve this, technologies that
16 generate longer reads or greater read depths are found to be very helpful [39]; but most
17 importantly combination of different sequencing technologies play a significant role in
18 improving genome assemblies [40] (Table S10). Libraries generated using different
19 chemistry have been found to be superior on improving assemblies [41]. Further, a
20 combination of different sequencing platform like Illumina when used in combination with
21 other technologies like Sanger sequencing, Pacbio and ONT have shown to reduce the
22 number of scaffolds even with very low coverage. Thus, we need to consider combination of
23 sequencing technologies, along with using different bioinformatics software to obtain
24 assembly with fewer number or scaffolds or closer to chromosome-level [42].

1 In comparison with other avian genomes [43], the 290X sequencing depth generated for
2 peacock is one of the highest. The final draft genome assembly of peacock resulted in N50 of
3 0.23 MB. Inclusion of 2.3X of reads from Oxford Nanopore helped the assembly to improve
4 by 26.2% reduction in the number of scaffolds and about 50.7% and 115% increase in the
5 scaffold and contig N50, respectively. The draft assembly contained less than 0.4% of
6 unknown nucleotides, which is very low for a draft assembly. Thus, we have shown for the
7 first time in avian genomics how the low-cost third generation sequencing data from Oxford
8 Nanopore can play a significant role in improving the genomes draft assembly. Assemblies
9 with longer scaffolds will further benefit in understanding the organisms with structurally
10 complex regions, repeat elements and isoforms in the genome [37].

11
12 The confidence on the predicted peacock proteins got strengthened when about 95% of them
13 showed significant homology to various genomic features from different databases (Fig. 7).
14 The phylogeny based on the conserved proteins across the avians showed that the peacock
15 being closest with chicken followed by turkey and duck, the grouping correlated to the
16 previous mitochondrial phylogeny [44]. Thus the genome sequence further gives insights on
17 its genetic lineage and evolution with relation to the other avian members. The estimated
18 median divergence time of *P. cristatus* from *G. gallus* is of about 35 million years ago
19 (MYA), whereas between *G. gallus* and *M. gallopavo* is about 37 MYA [45]. The huge gap
20 of other avians to peacock is due to non-availability of genome sequences from other avians.
21 The gap can be by sequencing other avian species. Among the vertebrates, it has been
22 observed that the variations in TEs among avians are very low [46] (Table S8). The genome
23 complexities of a species are influenced by the transposable elements (TE) that are believed
24 to play a crucial role [47]. In this peacock genome assembly, inclusion of Oxford Nanopore
25 long read sequences has significantly improved the assembly, thus, helping in resolving

1 across the repetitive regions in genome. Their roles in development and evolution of the
2 peacocks need to be further explored.

3
4
5
6
7 4 The genome information of peacock can be valued and explored by avian enthusiasts to
8
9
10 5 further understand about the avian world. Though not yet critically endangered in India,
11
12 6 peafowl population is surely at a declining trend in the wild due to massive deforestation,
13
14 7 habitat loss [48] and increased poaching for meat and feathers. Our genome sequencing
15
16
17 8 initiative of *Pavo cristatus* is not only valuable from a conservational viewpoint, but also to
18
19 9 preserve a heritage associated with this bird that runs through centuries and that bears a
20
21
22 10 strong attachment to the national psyche.

23 24 25 11 26 27 12 **Availability of supporting data**

28
29
30 13 Supplementary data contains, read statistics, annotation, repeats identification, orthology
31
32 14 analysis, assembly and annotation. Additional figures included are the peacock, chicken and
33
34
35 15 turkey proteins showing similarity to Pfam domains, top ten Gene ontology annotations in
36
37 16 Biological process; Cellular component and Molecular function from the Peacock proteins,
38
39
40 17 and Peacock homologous proteins in humans. Additional data are available from
41
42 18 <https://biit.cs.ut.ee/supplementary/peacock/>

43 44 45 19 46 47 20 **Raw Data and genome assembly in SRA**

48
49 21 Raw reads (Illumina and Oxford Nanopore) are available in the Sequence Read Archive
50
51 22 (SRA), and the Whole Genome Shotgun project has been deposited at GenBank under SRA
52
53
54 23 Submission ID: SUB3108024, Bioproject: PRJNA413288 and Biosamples
55
56
57 24 SUB3108018/SAMN07739105 : SKPea2016_SI, SUB3108017/SAMN07739104 :
58
59 25 SKPea2016_LI, SUB3107930/SAMN07739101 : FPL_3_5KB,

1 SUB3108015/SAMN07739102 : FPL_5_7KB, SUB3108016/SAMN07739103 :
2 FPL_7_10KB and SUB3108020/SAMN07739107 : FPL_Nano (Table 1). The de novo
3 genome assembly can be accessed under SUB4504869/ SAMN07739105.

5 **Competing interests**

6 The author(s) declare that they have no competing interests.

8 **Authors contributions**

9 RD, AS, KP performed wet lab experiments; RD designed work plan, experiments and
10 logistics; SS, VR, KP SG IM and AR assisted with the work; RS provided samples from bird;
11 BR, SK performed data analysis and interpretation; BR, SK drafted the manuscript and SK
12 overseen the whole project.

15 **Acknowledgements**

16 Department of Biochemistry, AIIMS, New Delhi for providing space and infrastructure to
17 carry on the work. RD for providing partial funding. Genotypic Technology and their team
18 for providing sequencing services. BR acknowledges IUT 34-4.

1 **Tables**

2 Table 1. Raw data statistics of Illumina HiSeq and Nanopore reads of the peacock genome.

Sample	Platform	Library and chemistry	Number of reads	Coverage	SRA ID
SO_6221_SKPea2016_SI	HiSeq	PE – SI (150 * 2)	489114747	146.73	SUB3108018, SAMN07739105
SO_6221_SKPea2016_LI	HiSeq	PE – LI (150 * 2)	302884819	90.87	SUB3108017, SAMN07739104
SO_6221_FPL_3_5KB	HiSeq	MP (150 * 2)	72915033	21.87	SUB3107930, SAMN07739101
SO_6221_FPL_5_7KB	HiSeq	MP (150 * 2)	47440144	14.23	SUB3108015, SAMN07739102
SO_6221_FPL_7_10KB	HiSeq	MP (150 * 2)	36464628	10.94	SUB3108016, SAMN07739103
SO_6221_NP	Nanopore	5 - 341124	366323	2.3	SUB3108020, SAMN07739107

3
4 Abbreviations used, PE = Paired end, SI = Short Insert, LI = Long insert, MP = Mate pair, NP = Nano pore and

5 KB = Kilo Bases

6 Table 2. De novo assembly statistics of the peacock genome.

Description	Contigs	Nanopore Scaffold	Super Scaffolds	GapClosed	>1000 Kb	>5000 Kb
Contigs	685,241	281,272	179,346	179,332	34,178	15,025
Maximum Length	49,159	251,510	2,390,121	2,488,982	2,488,982	2,488,982
Minimum Length	300	5	265	265	1000	5000
Average Length	1360	3250	5111	5729	-	-
Total Length	932,162,464	914,363,908	916,720,956	1,027,510,962	954,449,349	915,342,012
Length >= 100 bp	685,241	281,271	179,346	179,332	34,178	15,025
Length >= 200 bp	685,241	281,271	179,346	179,332	34,178	15,025
Length >= 500 bp	616,120	186,433	93,727	93,718	34,178	15,025
Length >= 1 Kb	363,428	104,479	34,168	34,178	34,178	15,025
Length >= 10 Kb	1591	24,748	9249	10,310	10,310	10,310
Length >= 1 Mb	0	0	27	37	37	37
Non-ATGC #	350,325	42,696,911	49,169,831	4,043,129	4,040,790	3,986,487
Non-ATGC %	0.038	4.67	5.36	0.393	0.423	0.436
N50 value	1639	14,748	168,140	190,304	218,023	232,312

7

8

1 **Figure legend**

2 **Figure 1.** The beautiful and charismatic photo of Indian blue peacock (*Pavo cristatus*) bird.

3 **Figure 2.** Detailed workflow for de novo whole genome assembly and annotation.

4 **Figure 3.** Peacock proteins showing homology. Pie chart showing significant similarity
5 scores of peacock proteins against the NR database.

6 **Figure 4.** Phylogenetic tree generated from homologous proteins from 49 different avian
7 species.

8 **Figure 5.** Venn diagram showing common and not present protein family domains (Pfam)
9 between peacock, chicken and turkey proteins.

10 **Figure 6.** Heatmap showing protein family (Pfam) distributed in peacock, chicken or turkey
11 species. The number represents the Pfam domain count predicted from the protein sequences.
12 Pfam domains of 50 and above identified in any one of the species are compared in the
13 heatmap.

14 **Figure 7.** Venn diagram showing peacock proteins showing significant homology to NR
15 database, KOG, Pfam and GO ontologies.

16 **Figure 8.** Circular image of the assembled peacock genome aligned against the *G. gallus*
17 genome. The right side of the image represents the reference chicken genome and left side of
18 the image represents the peacock genome.

1 **References:**

- 2 1. Brickle NW. Habitat use, predicted distribution and conservation of green peafowl
3 (Pavo muticus) in Dak Lak Province, Vietnam. *Biological*
4 *Conservation*. 2002;105:189–97.
- 5 2. Jackson CE. *Peacock*. London: Reaktion, 2006.
- 6 3. Kadgaonkar, Shivendra B. The peacock in ancient Indian art and literature. *Bulletin of*
7 *the Deccan College Research Institute*. 1993;53:95–115.
- 8 4. Gadagkar R. Is the peacock merely beautiful or also honest? *Current*
9 *Science*. 2003;85:1012–20.
- 10 5. Kushwaha S, Kumar A. A Review on Indian Peafowl (*Pavo cristatus*) Linnaeus,
11 1758. *Journal of Wildlife Research*. 2016;4:42–59.
- 12 6. Hillier LW, Miller W, Birney E et al. Sequence and comparative analysis of the
13 chicken genome provide unique perspectives on vertebrate evolution. *International*
14 *Chicken Genome Sequencing Consortium* (ed.). *Nature*. 2004;432:695–716.
- 15 7. Zhang G, Jarvis ED, Gilbert MTP. A flock of genomes. *Science*. 2014;346:1308–9.
- 16 8. Bejerano G, Pheasant M, Makunin I et al. Ultraconserved elements in the human
17 genome. *Science*. 2004;304:1321–5.
- 18 9. Burt DW. Emergence of the Chicken as a Model Organism: Implications for
19 *Agriculture and Biology*. *Poultry Science*. 2007;86:1460–71.
- 20 10. Furlong RF. Insights into vertebrate evolution from the chicken genome
21 sequence. *Genome Biology*. 2005;6:207.
- 22 11. Andrews S. *FastQC A Quality Control tool for High Throughput Sequence*
23 *Data*. Babraham.ac.uk 2010.
- 24 12. Kajitani R, Toshimoto K, Noguchi H et al. Efficient de novo assembly of highly
25 heterozygous genomes from whole-genome shotgun short reads. *Genome*
26 *Research*. 2014;24:1384–95.
- 27 13. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence
28 data. *Bioinformatics*. 2014;30:3399–401.

1 14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
2 transform. *Bioinformatics*. 2009;25:1754–60.

3 15. Birol I, Jackman SD, Nielsen CB et al. De novo transcriptome assembly with
4 ABySS. *Bioinformatics*. 2009;25:2872–7.

5 16. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes
6 using long read sequence information. *BMC Bioinformatics*. 2014;15:211.

7 17. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
8 occurrences of k-mers. *Bioinformatics*. 2011;27:764–70.

9 18. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013-
10 2015. www.repeatmasker.org.

11 19. Stanke M, Diekhans M, Baertsch R et al. Using native and syntenically mapped
12 cDNA alignments to improve de novo gene finding. *Bioinformatics*. 2008;24:637–44.

13 20. Boetzer M, Henkel CV, Jansen HJ et al. Scaffolding pre-assembled contigs using
14 SSPACE. *Bioinformatics*. 2010;27:578–9.

15 21. Altschul SF, Gish W, Miller W et al. Basic local alignment search tool. *Journal of*
16 *molecular biology*. 1990;215:403–10.

17 22. Moriya Y, Itoh M, Okuda S et al. KAAS: an automatic genome annotation and
18 pathway reconstruction server. *Nucleic Acids Research*. 2007;35:W182–5.

19 23. Nordberg H, Cantor M, Dusheyko S et al. The genome portal of the Department of
20 Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*. 2013;42:D26–
21 31.

22 24. El-Gebali S, Mistry J, Bateman A et al. The Pfam protein families database in
23 2019. *Nucleic Acids Research*. 2018;47:D427–32.

24 25. Zhang G, Li B, Li C et al. Comparative genomic data of the Avian Phylogenomics
25 Project. *GigaScience*. 2014;3:26.

26 26. Jarvis ED, Mirarab S, Aberer AJ et al. Phylogenomic analyses data of the avian
27 phylogenomics project. *GigaScience*. 2015;4:4.

28 27. Fu L, Niu B, Zhu Z et al. CD-HIT: accelerated for clustering the next-generation
29 sequencing data. *Bioinformatics*. 2012;28:3150–2.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 28. Larkin MA, Blackshields G, Brown NP et al. Clustal W and Clustal X version
2 2.0. *Bioinformatics*. 2007;23:2947–8.

3
4 3 29. Talavera G, Castresana J. Improvement of Phylogenies after Removing Divergent and
5 4 Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic*
6 5 *Biology*. 2007;56:564–77.

7
8
9
10 6 30. Felsenstein J. PHYLIP (Phylogeny Inference Package) Version
11 7 3.57c. <http://evolution.genetics.washington.edu/phylip.html> 1993.

12
13
14 8 31. Nguyen L-T, Schmidt HA, von Haeseler A et al. IQ-TREE: A Fast and Effective
15 9 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular*
16 10 *Biology and Evolution*. 2014;32:268–74.

17
18
19
20 11 32. Frith MC, Kawaguchi R. Split-alignment of genomes finds orthologies more
21 12 accurately. *Genome Biology*. 2015;16:106.

22
23
24 13 33. Krzywinski M, Schein J, Birol I et al. Circos: an information aesthetic for
25 14 comparative genomics. *Genome research*. 2009;19:1639–45.

26
27
28
29 15 34. Warren WC, Hillier LW, Tomlinson C et al. A New Chicken Genome Assembly
30 16 Provides Insight into Avian Genome Structure. *G3: Genes, Genomes,*
31 17 *Genetics*. 2016;7:109–17.

32
33
34 18 35. Peona V, Weissensteiner MH, Suh A. How complete are “complete” genome
35 19 assemblies?-An avian perspective. *Molecular Ecology Resources* 2018;18:1188–95.

36
37
38
39 20 36. Muir P, Li S, Lou S et al. The real cost of sequencing: scaling computation to keep
40 21 pace with data generation. *Genome Biology*. 2016;17:53.

41
42
43 22 37. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-
44 23 generation sequencing technologies. *Nature Reviews Genetics*. 2016;17:333–51.

45
46
47 24 38. Levy SE, Myers RM. Advancements in Next-Generation Sequencing. *Annual Review*
48 25 *of Genomics and Human Genetics*. 2016;17:95–115.

49
50
51 26 39. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome
52 27 Assembly. *Genomics, Proteomics & Bioinformatics*. 2016;14:265–79.

53
54
55
56 28 40. Rice ES, Green RE. New Approaches for Genome Assembly and Scaffolding. *Annual*
57 29 *Review of Animal Biosciences*. 2018;7.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 41. Weissensteiner MH, Pang AWC, Bunikis I et al. Combination of short-read, long-
2 read, and optical mapping assemblies reveals large-scale tandem repeat arrays with
3 population genetic implications. *Genome Research*. 2017;27:697–708.
4
5
6 42. Sohn J, Nam J-W. The present and future of de novo whole-genome assembly.
7 *Briefings in Bioinformatics*. 2016:bbw096.
8
9
10 43. Zhang G, Li C, Li Q et al. Comparative genomics reveals insights into avian genome
11 evolution and adaptation. *Science*. 2014;346:1311–20.
12
13
14 44. Dalloul RA, Long JA, Zimin AV et al. Multi-Platform Next-Generation Sequencing
15 of the Domestic Turkey (*Meleagris gallopavo*): Genome Assembly and Analysis.
16 Roberts RJ (ed.). *PLoS Biology* 2010;8:e1000475.
17
18
19
20 45. Kumar S, Stecher G, Suleski M et al. TimeTree: A Resource for Timelines,
21 Timetrees, and Divergence Times. *Molecular Biology and Evolution*. 2017;34:1812–
22 9.
23
24
25
26 46. Sotero-Caio CG, Platt RN, Suh A et al. Evolution and Diversity of Transposable
27 Elements in Vertebrate Genomes. *Genome Biology and Evolution*. 2017;9:161–77.
28
29
30
31 47. Kapusta A, Suh A. Evolution of bird genomes—a transposon’s-eye view. *Annals of the*
32 *New York Academy of Sciences*. 2016;1389:164–85.
33
34
35 48. Ramesh K, McGowan P. On the current status of Indian Peafowl *Pavo cristatus*
36 (*Aves: Galliformes: Phasianidae*): keeping the common species common. *Journal of*
37 *Threatened Taxa*. 2009;1:106–8.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65



Figure 2

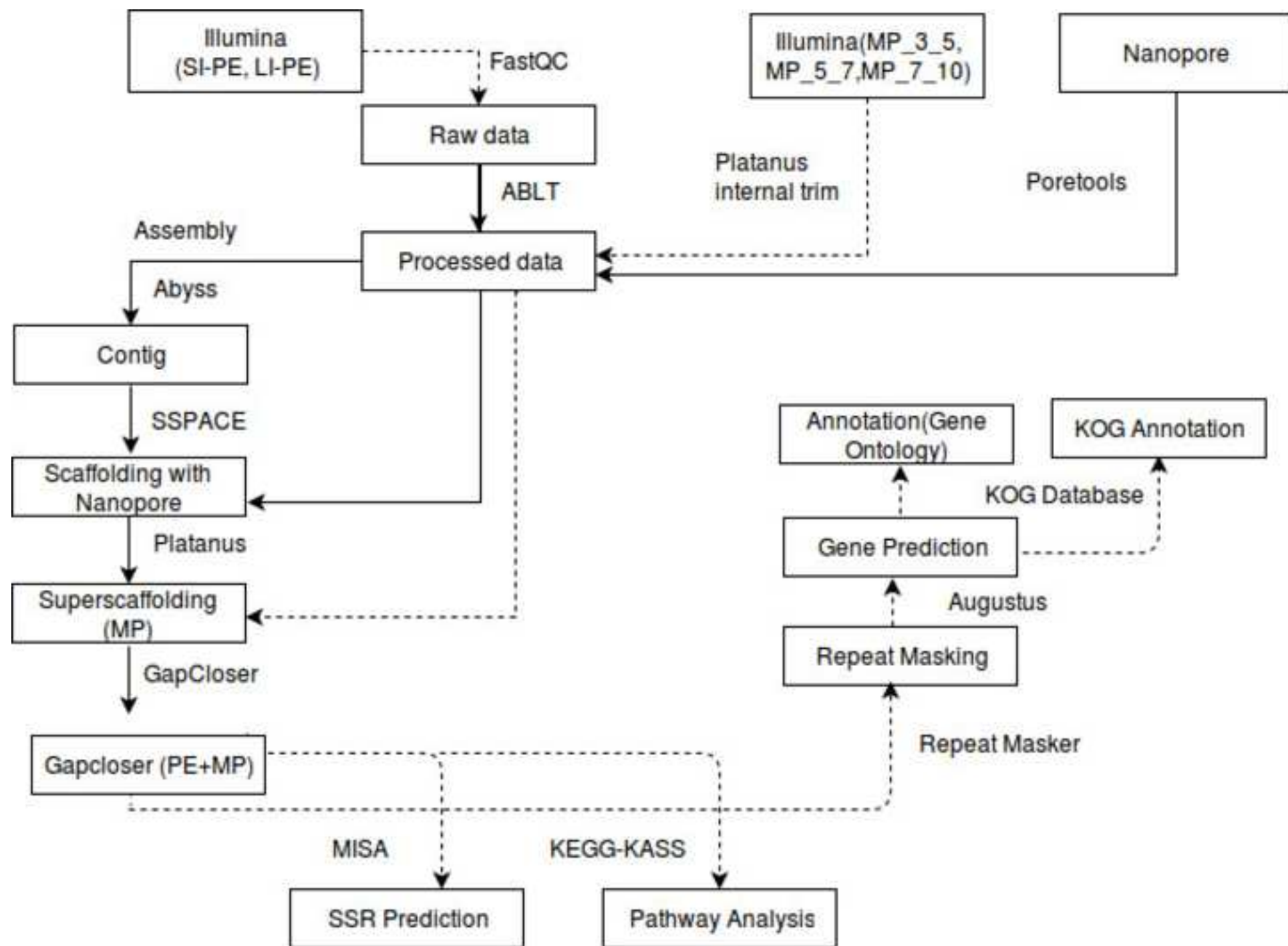


Figure 3

[Click here to access/download;Figure;Fig 3; Similarity scores against the Uniprot database.pdf](#)

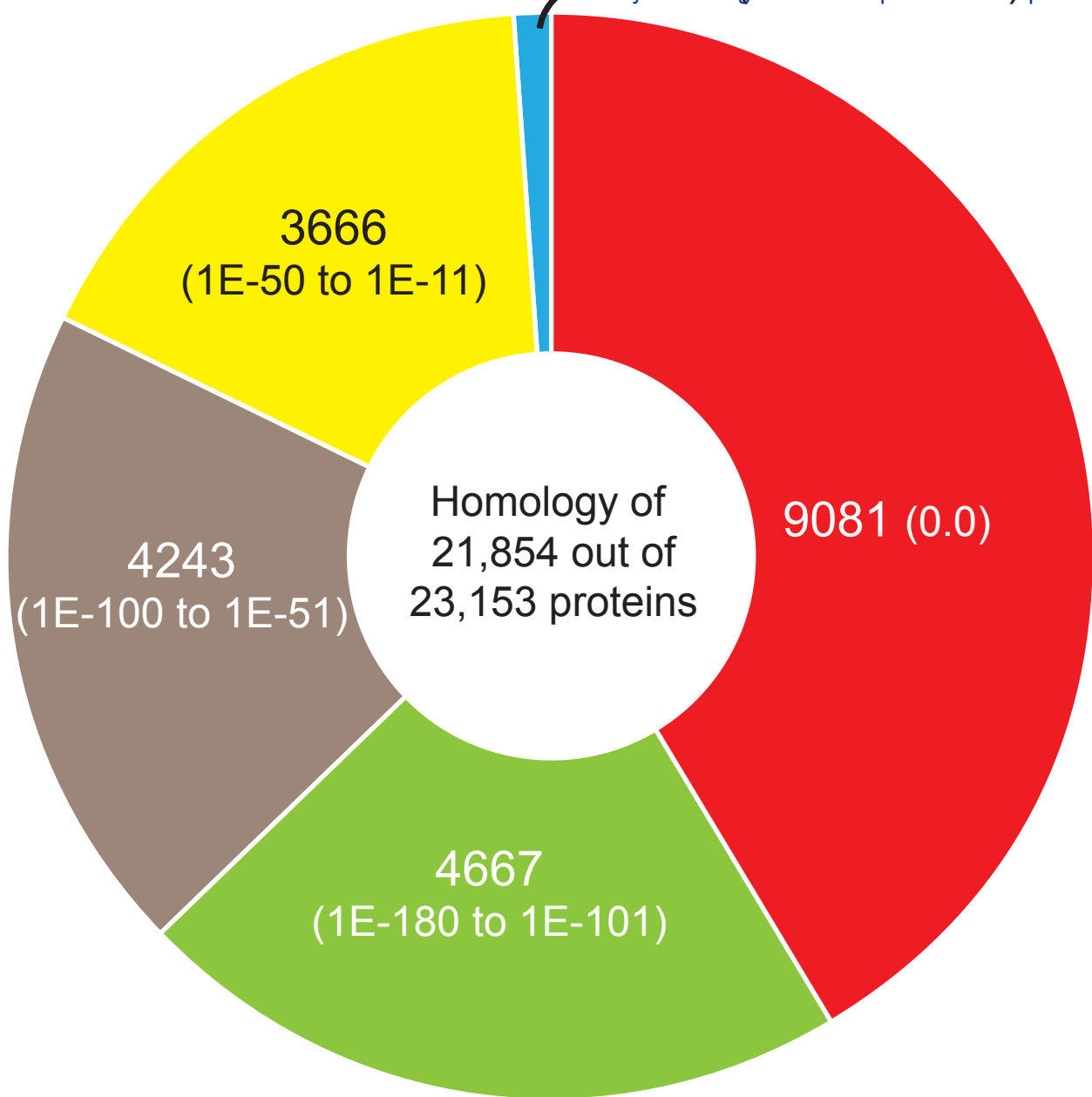
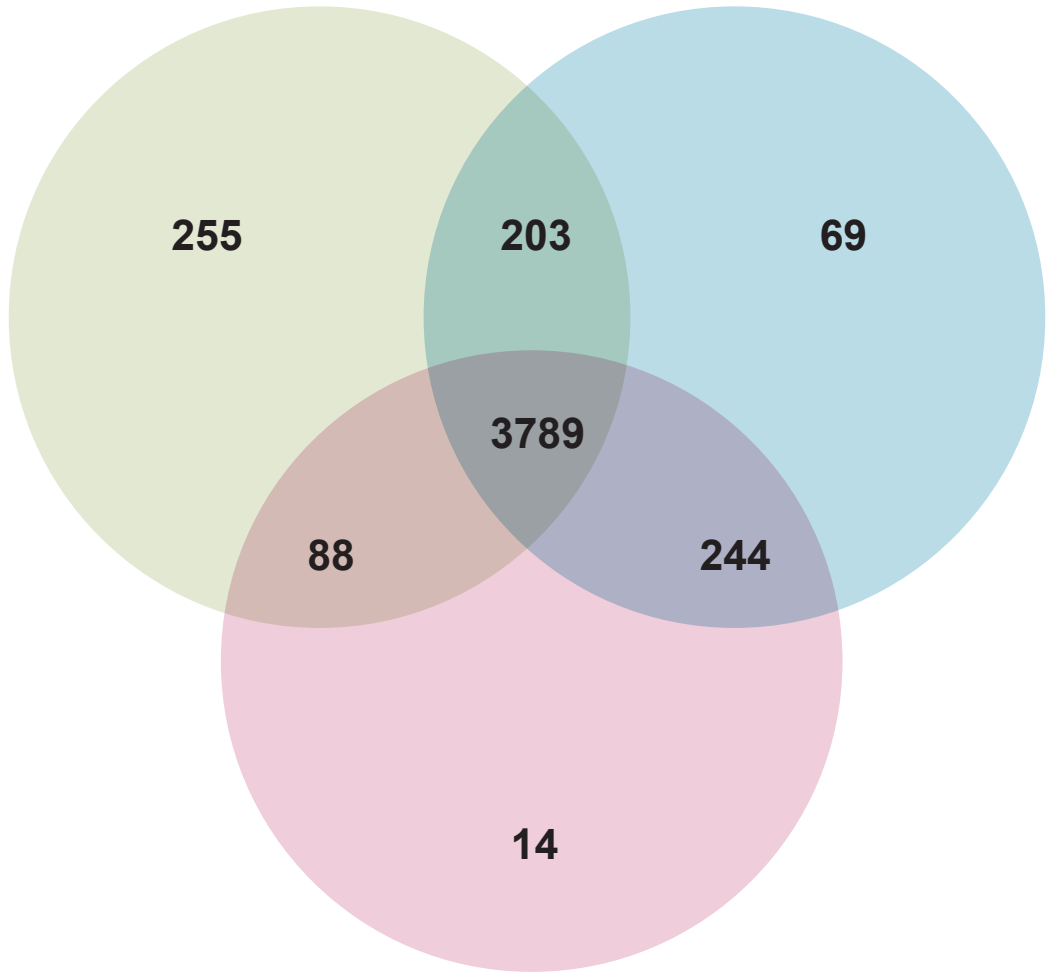
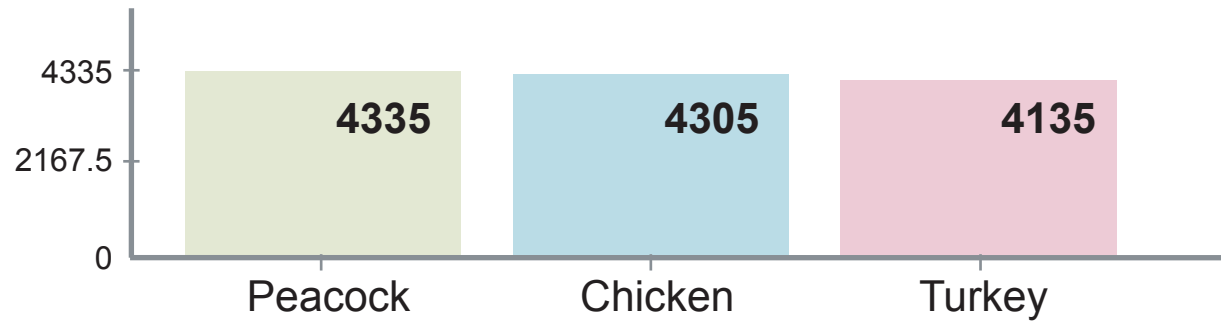


Figure 5

[Click here to access/download;Figure;Fig 5.](#) Unique Pfam domains common with chicken



Turkey



Number of Pfam domains unique to 1 species or shared between 2 or all 3



Figure 6



Click here to access/download/figure;Figure6. Heatmap

Figure 6

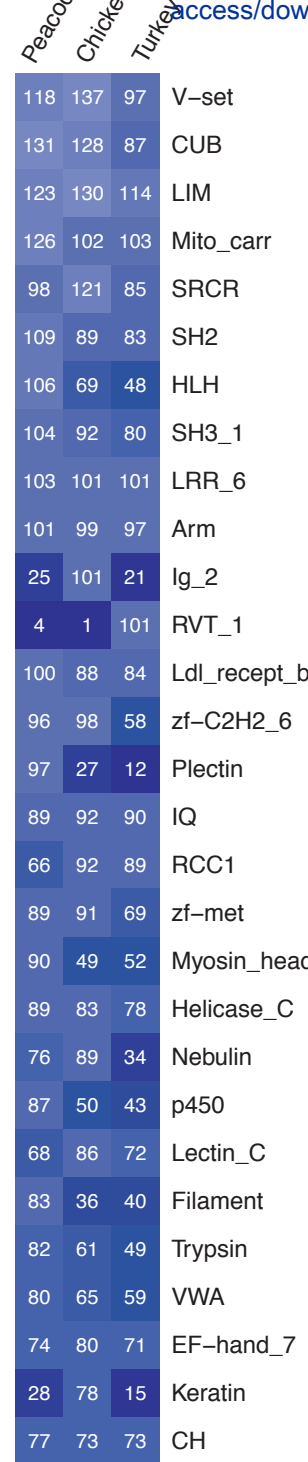
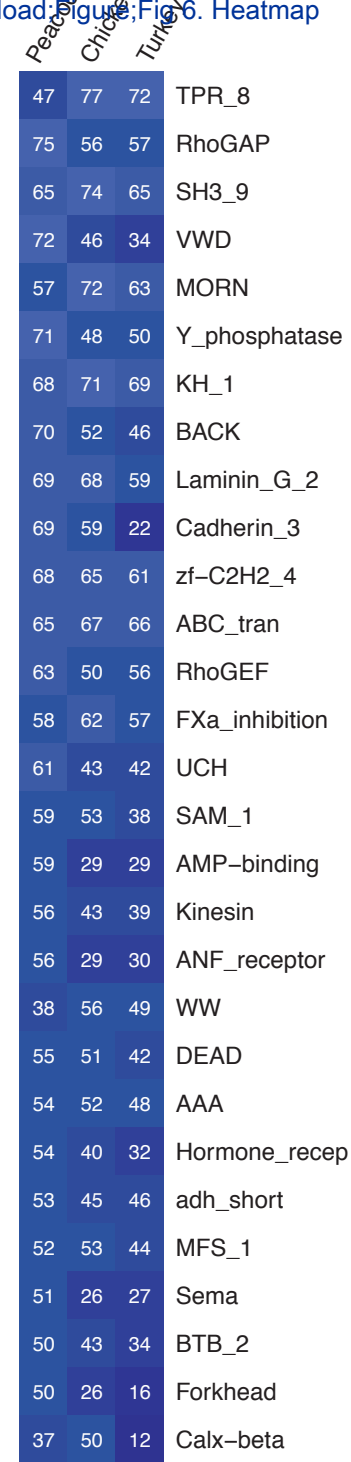
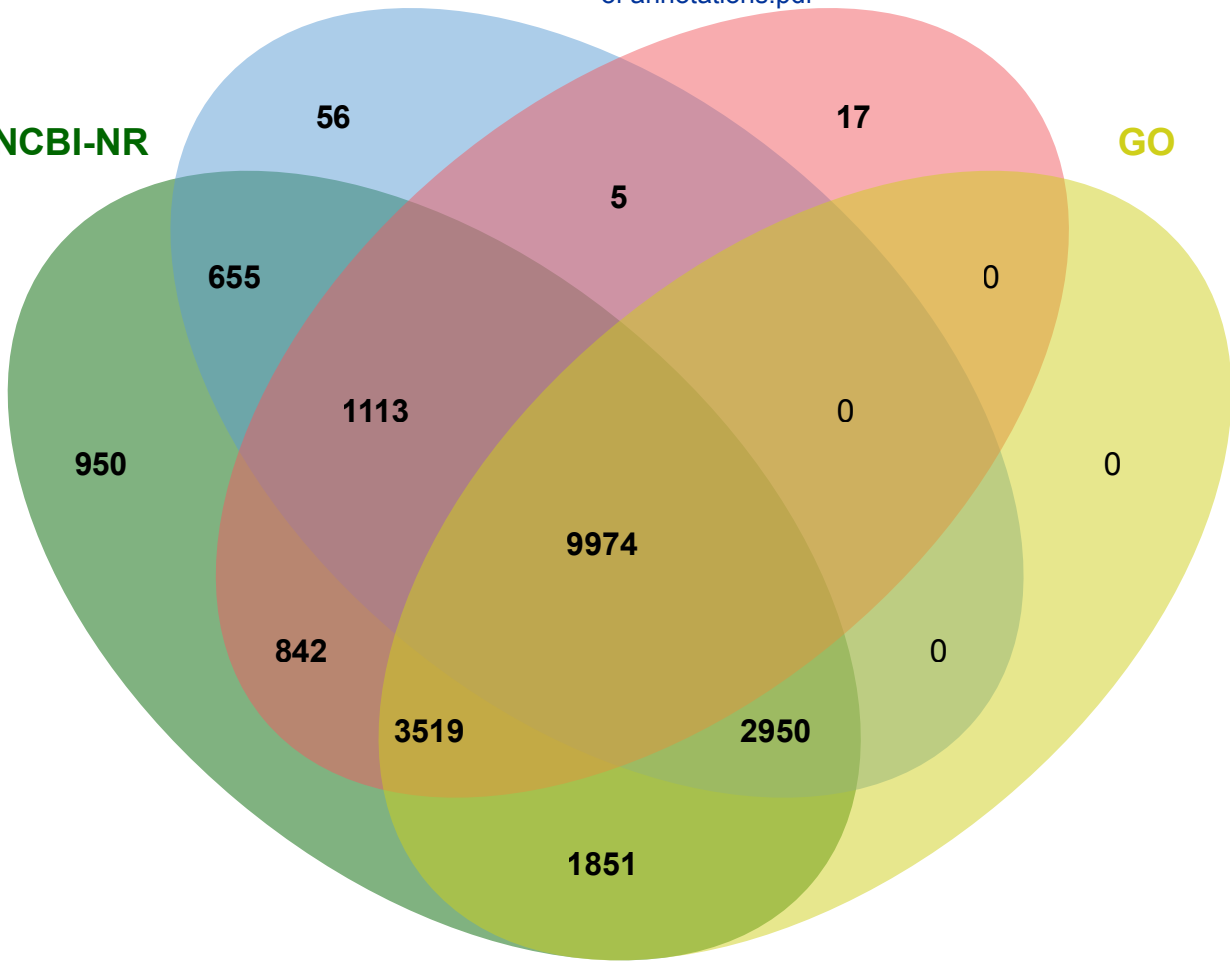


Figure 6

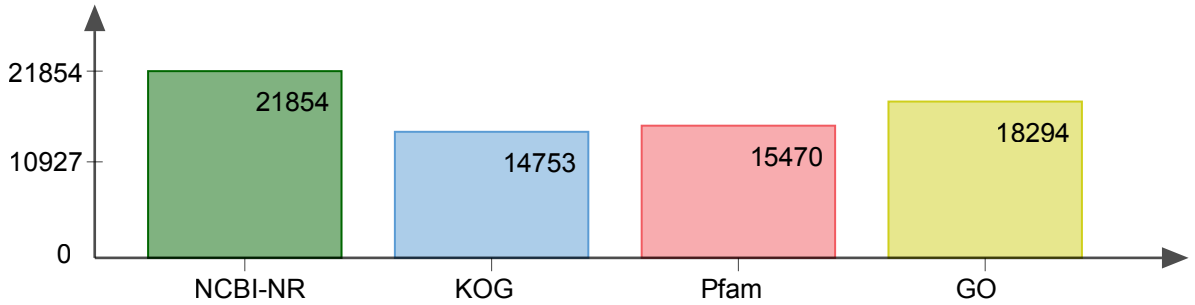


NCBI-NR

GO



Proteins annotated from different sources



Number of common proteins: specific to 1 or shared by 2, 3, or 4 annotations

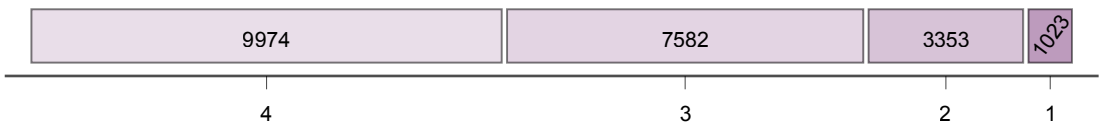
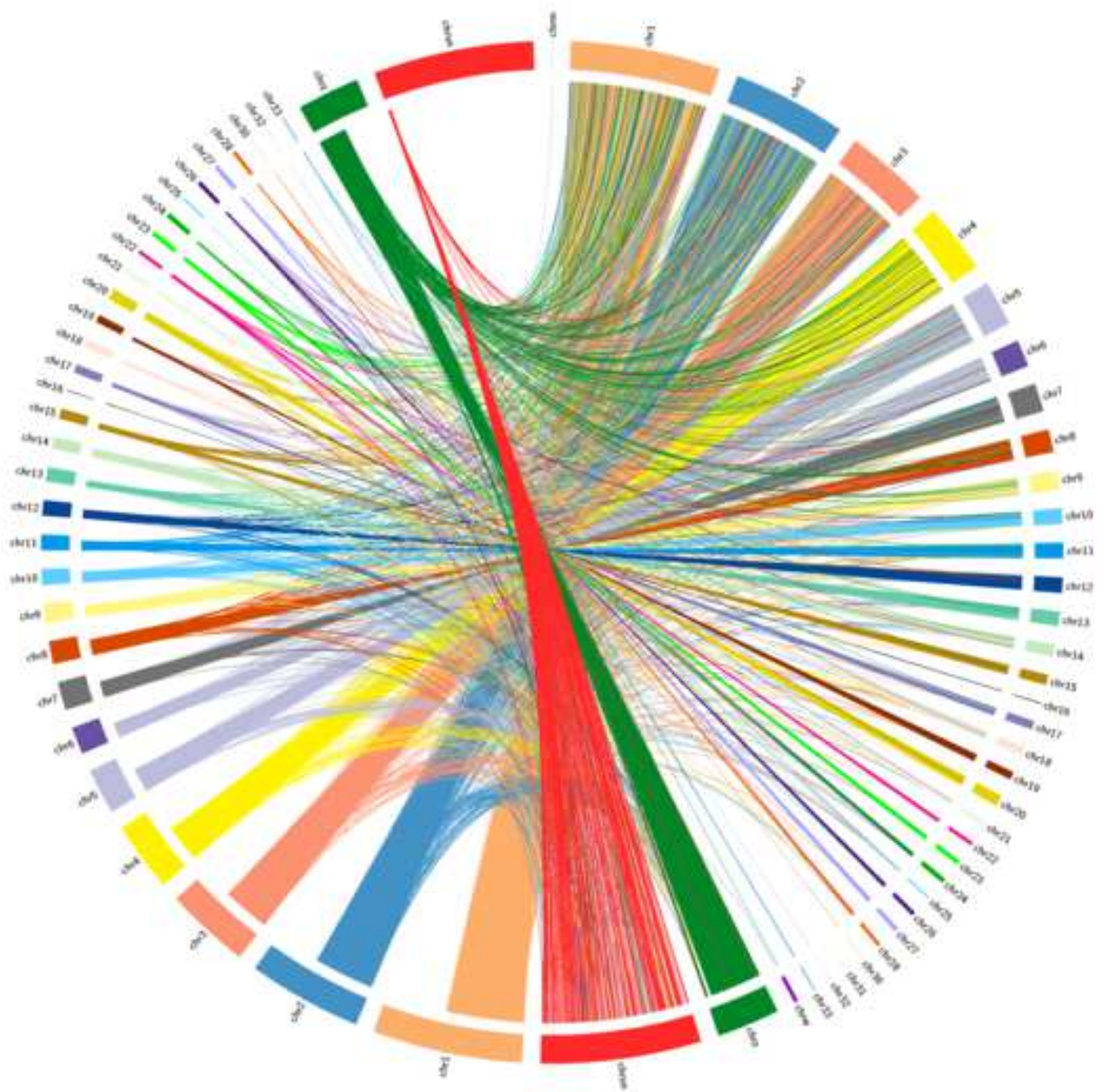



Figure 8

[Click here to access/download;Figure;Fig 8. Peacock scaffolds against Gallus circular synteny.png](#)





Click here to access/download
Supplementary Material
Supplementary_Description of all the tables and
figures.docx



[Click here to access/download](#)

Supplementary Material

[Table_S1_ReadStats_Table_S2_TEs.xlsx](#)





Click here to access/download
Supplementary Material
Table_S3_Repeats.xlsx

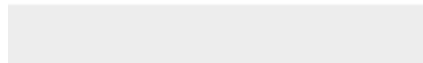




[Click here to access/download](#)

Supplementary Material

[Table_S4_Gene_annotations_of_peacock_proteins.xlsx](#)



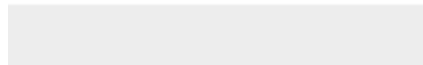


Click here to access/download
Supplementary Material
Table_S5_KEGG_annotation.xlsx





Click here to access/download
Supplementary Material
Table_S6_KOG_annotation.xlsx

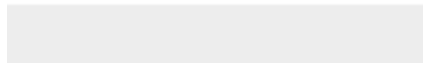




[Click here to access/download](#)

Supplementary Material

[Table_S7_BlastVsHumanProteins.xlsx](#)





Click here to access/download
Supplementary Material
Table_S8_Orthologous_proteins

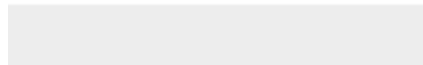




Click here to access/download

Supplementary Material

Table_S9_Pfam_Analysis.xlsx





[Click here to access/download](#)

Supplementary Material

[Table_S10_Bird_Species_with_counts.xlsx](#)

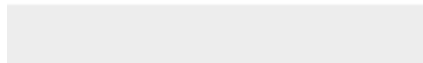




[Click here to access/download](#)

Supplementary Material

Fig S1. Proteins showing similarity to Pfam domains.pdf





[Click here to access/download](#)

Supplementary Material

Fig S2. Gene Ontology of top 10 WGS.png





Click here to access/download
Supplementary Material
Fig S3.Peacock vs Human_GO.pdf

