

De novo genome assembly of Indian Blue Peacock (*Pavo cristatus*) from Oxford Nanopore and Illumina sequencing reads --Manuscript Draft--

Manuscript Number:	GIGA-D-18-00280
Full Title:	De novo genome assembly of Indian Blue Peacock (<i>Pavo cristatus</i>) from Oxford Nanopore and Illumina sequencing reads
Article Type:	Data Note
Funding Information:	
Abstract:	<p>Background</p> <p><i>Pavo cristatus</i> the Indian blue peacock are geographically found distributed in natural habitats of South Asia. Peacock has been described among the bird species as one of the most elegant, majestic and beautiful bird (Fig. 1). Since prehistoric times they have been described or depicted in Indian culture and has been adopted as the national bird of India. Its length varies from 92-125 centimeter (without train), weighing about 4-8 Kilograms and lives up to 20 years in the wild. The avian species have been very important in the fields of phylogenetics, developmental studies, sexual reproduction and speciation. Avian genomics have contributed immensely towards understanding the vertebrate genome evolution. Here we present the first draft genome sequence of <i>P. cristatus</i>, yet another important bird species to further add values and gain insight into avian genomics.</p> <p>Findings</p> <p>For the first time in avian genomics, long reads using Oxford Nanopore technology have been used for the whole genome assembly. We sequenced different DNA insert size libraries from Illumina and long read Nanopore technologies from the peacock DNA. We performed de novo genome assembly by integrating the reads from Illumina short insert, long insert, multiple mate-pair reads along with Nanopore long reads using multiple genome improvement tools. A draft of the peacock genome of about 0.915 Gigabases (Gb) with a N50 of 0.23 Megabases (Mb) was assembled. Annotations with other avian species, protein families, KEGG were performed for functional understanding by in-silico approaches. Proteins were compared against Chicken, Turkey and Human to obtain evolutionary similarities and uniqueness of the <i>Pavo</i> species.</p> <p>Conclusions</p> <p>Our most important findings from the genome sequence of <i>P. cristatus</i> is to decipher the different gene families and to understand their role in body pattern development and other features that truly makes this bird unique. The genome sequence also gives insights on its genetic lineage and evolution with relation to other avian members. Several hypothesis and theories have been discussed with respect to sexual selection; now with the understanding of the genome sequence, some of these evolutionary theories will be better understood. The genome will also support future studies on population genetics and breeding for species conservation as well as in understanding its evolutionary ecology and sexual dimorphism. The comparative genomics with other avian species and specifically with <i>Gallus gallus</i> (Chicken) and <i>Meleagris gallopavo</i> (Turkey) have shown insights into the gene families and their conserved domains. <i>Pavo</i> proteins were also compared with human to understand the functional components that were conserved after the speciation split.</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:	Subhradip Karmakar, PhD
First Author Secondary Information:	
Order of Authors:	Subhradip Karmakar, PhD
	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
Order of Authors Secondary Information:	
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist . Information essential to interpreting the data presented should be made available in the figure legends.	
Have you included all the information requested in your manuscript?	
Resources	Yes
A description of all resources used, including antibodies, cell lines, animals	

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

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

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,*} and Subhradip
9 Karmakar^{1,*}

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

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

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

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

24 **Abstract**

25 **Background**

1
2 26 *Pavo cristatus* the Indian blue peacock are geographically found distributed in natural
3
4
5 27 habitats of South Asia. Peacock has been described among the bird species as one of the most
6
7 28 elegant, majestic and beautiful bird (Fig. 1). Since prehistoric times they have been described
8
9
10 29 or depicted in Indian culture and has been adopted as the national bird of India. Its length
11
12 30 varies from 92-125 centimeter (without train), weighing about 4-8 Kilograms and lives up to
13
14 31 20 years in the wild. The avian species have been very important in the fields of
15
16
17 32 phylogenetics, developmental studies, sexual reproduction and speciation. Avian genomics
18
19 33 have contributed immensely towards understanding the vertebrate genome evolution. Here
20
21
22 34 we present the first draft genome sequence of *P. cristatus*, yet another important bird species
23
24 35 to further add values and gain insight into avian genomics.
25
26
27 36

28 29 37 **Findings**

30
31 38 For the first time in avian genomics, long reads using Oxford Nanopore technology have
32
33
34 39 been used for the whole genome assembly. We sequenced different DNA insert size libraries
35
36 40 from Illumina and long read Nanopore technologies from the peacock DNA. We performed
37
38
39 41 *de novo* genome assembly by integrating the reads from Illumina short insert, long insert,
40
41 42 multiple mate-pair reads along with Nanopore long reads using multiple genome
42
43
44 43 improvement tools. A draft of the peacock genome of about 0.915 Gigabases (Gb) with a
45
46 44 N50 of 0.23 Megabases (Mb) was assembled. Annotations with other avian species, protein
47
48
49 45 families, KEGG were performed for functional understanding by insilico approaches.
50
51 46 Proteins were compared against Chicken, Turkey and Human to obtain evolutionary
52
53
54 47 similarities and uniqueness of the *Pavo* species.
55
56 48

57 58 49 **Conclusions**

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

Our most important findings from the genome sequence of *P. cristatus* is to decipher the different gene families and to understand their role in body pattern development and other features that truly makes this bird unique. The genome sequence also gives insights on its genetic lineage and evolution with relation to other avian members. Several hypothesis and theories have been discussed with respect to sexual selection; now with the understanding of the genome sequence, some of these evolutionary theories will be better understood. The genome will also support future studies on population genetics and breeding for species conservation as well as in understanding its evolutionary ecology and sexual dimorphism. The comparative genomics with other avian species and specifically with *Gallus gallus* (Chicken) and *Meleagris gallopavo* (Turkey) have shown insights into the gene families and their conserved domains. *Pavo* proteins were also compared with human to understand the functional components that were conserved after the speciation split.

63 **Introduction**

64 *Pavo cristatus* commonly known as the blue Indian peacock obtained the status of National
65 Bird of India in 1963. Peacocks have been distributed widely in Asian countries. The bird is
66 very popular as it symbolizes beauty, love, grace and pride (Gadagkar, R; Kushwaha, et al.).
67 It has been referred in ancient literatures of India and has been found closely associated with
68 the life and culture of the peoples from South East Asia and particularly India (Kadgaonkar,
69 SB). Peacocks are native to South Asia but have been introduced into many other countries
70 usually as exhibits in park, zoos and also large number of aviculturists raise and breed these
71 species as pets (Brickle, N; Jackson, C).

72
73 International Chicken Genome Sequencing Consortium sequenced the *Gallus gallus* genome
74 in 2004, and this laid the foundation for study of avian genomics. A decade later, the avian
75 genome consortium assembled 48 genomes of wide variety of avian species to understand the
76 evolutionary relationships of modern birds (Zhang, G., et al.). Raw sequencing data for each
77 species were generated at from coverage of 6X for zebra finch to a maximum 160X for
78 budgerigar. The genome size varied from 1.04 to 1.26 Gb (<http://avian.genomics.cn/en/>). The
79 studies on different bird species have provided a new perspective on vertebrate genome
80 evolution. These genomes have also helped in improving the annotation of mammalian
81 genomes. There are several distinguishing as well as unique features between chicken and
82 human genome including genome size which is of one third of humans, conserved synteny
83 blocks complete absence of active short interspersed nucleotide elements (SINE) to mention
84 a few (International Chicken Genome sequencing consortium).

85
86 Despite the wealth of information from avian genomes sequencing projects, it is very
87 important to genome sequence other new species to add value into aves and vertebrate

1 88 genomics. For the first time in bird genomics, Oxford Nanopore technology has been used to
2 89 sequence a bird genome in this present study. The long read chemistry will help in better
3
4 90 genome assembly of the TEs and repeat rich. The peacock genome will aid in understanding
5
6
7 91 about the uniqueness of this birds genome in comparison to other bird species. Comparative
8
9
10 92 genomes will help in understanding the development of this species, sexual selection and its
11
12 93 evolutionary relationships with other birds. The characterization of the genes involved in sex
13
14 94 determination could provide relevant information for the selective breeding of the peafowl
15
16 95 populations. We have unraveled some of the genomic signatures and thus have reported
17
18
19 96 unique gene pools of this bird by performing comparative genomics. Further different data
20
21
22 97 types will improve the assembly and gene/genome characterization will help to address the
23
24 98 sexual selection theory and key answers relevant to the evolution of this bird.
25
26
27 99
28
29 100

31 101 **Materials and methods**

34 102 **Sample collection and extraction of DNA**

36 103 The whole blood of male peacock was collected from Kanpur zoo, India after obtaining the
37
38
39 104 necessary ethical and institutional approval. 20µl of Proteinase K (PK) solution was taken
40
41 105 into a 1.5ml micro centrifuge tube. 200µl of blood was added and briefly mixed. 200µl of cell
42
43
44 106 lysis buffer was added to the tube, mixed by vortexing for 10seconds; incubated at 56°C for
45
46 107 10minutes. ReliaPrep™ Binding Column was placed into an empty collection tube. 250µl of
47
48
49 108 Binding Buffer (BBA) was added, capped the tube, and mixed by vortexing for 10 seconds
50
51 109 with a vortex mixer. Contents of the tube were added to the ReliaPrep™ binding column,
52
53
54 110 capped and placed in a refrigerated micro centrifuge. These were then centrifuged for 1
55
56 111 minute at maximum speed and flow through was discarded. Binding column was placed into
57
58
59 112 a fresh collection tube. 500µl of column wash solution was added to the column and
60
61
62
63
64
65

113 centrifuged for 3 minutes at maximum speed; Flow through was again discarded. Column
114 washing is repeated thrice. Columns were then placed in a nuclease free clean 1.5ml micro
115 centrifuge tube. 100 µl of Nuclease-Free Water was then added to the column and
116 centrifuged for an additional 1 minute at maximum speed. Column was discarded and the
117 elute was saved. The concentration and purity of the extracted DNA was evaluated using
118 Nanodrop Spectrophotometer (Thermo Scientific) and Qubit flurometer and integrity was
119 checked on a 0.8% agarose gel. The DNA sample was aliquoted for library preparation on
120 two different platforms: Illumina HiSeq4000 and Oxford Nanopore Technologies (ONT).

121

122 **HiSeq Paired-End library preparation and sequencing**

123 Whole genome sequencing (WGS) libraries were prepared with Illumina-compatible
124 NEXTflex DNA sequencing kit (BIOO Scientific, Austin, Texas, U.S.A.). Briefly, approx. 1
125 µg of genomic DNA was sheared using Covaris S2 sonicator (Covaris, Woburn,
126 Massachusetts, USA) to generate approx. fragment size distribution from 300 to 600 basepair
127 (bp). The fragment size distribution was checked on Agilent 2200 Tape Station with D1000
128 DNA screen tapes and reagents (Agilent Technologies, Palo Alto, CA, USA) and
129 subsequently purified using HighPrep magnetic beads (Magbio Genomics Inc, USA). The
130 purified fragments were end-repaired, adenylated and ligated to Illumina multiplex barcode
131 adaptors as per NEXTflex DNA sequencing kit protocol (BIOO Scientific, Austin, Texas,
132 USA).

133

134 The adapter-ligated DNA was purified with HighPrep beads (MagBio Genomics, Inc,
135 Gaithersburg, Maryland, USA) and then size selected on 2% low melting agarose gel and
136 cleaned using MinElute column (QIAGEN). The resultant fragments were amplified for 10
137 cycles of PCR using Illumina-compatible primers provided in the NEXTFlex DNA

138 sequencing kit. The final PCR product (sequencing library) was purified with HighPrep
139 beads, followed by library quality control check. The Illumina-compatible sequencing library
140 was initially quantified by Qubit fluorometer (Thermo Fisher Scientific, MA, USA) and its
141 fragment size distribution was analyzed on Agilent TapeStation. Finally, the sequencing
142 library was accurately quantified by quantitative PCR using Kapa Library Quantification Kit
143 (Kapa Biosystems, Wilmington, MA, USA). The qPCR-quantified library was subjected to
144 sequencing on an Illumina sequencer for 150 bp paired-end chemistry.

146 The Illumina-compatible sequencing library for the samples has a fragment size range
147 between 275 to 425 bp for Paired-End Short Insert (PE-SI) and 350 bp to 650bp for Paired-
148 End Long Insert (PE-LI). As the combined adapter size is approximately 120bp, the effective
149 user-defined insert size is 155 to 305 bp and 230 to 530 bp for PE-SI and PE-LI respectively.
150 Libraries were sequencing in Illumina HiSeq platform with 150*2 chemistry. The short reads
151 of Paired-End Short Insert (PE-SI), Paired-End Long Insert (PE-LI) and Mate-Pair (MP) from
152 Illumina HiSeq platform.

155 **Mate-Pair library preparation and sequencing**

156 Mate Pair sequencing library was prepared with Illumina-compatible Nextera Mate Pair
157 Sample Preparation Kit (Illumina Inc., Austin, TX, U.S.A.). Briefly, approx. 4 ug of genomic
158 DNA was simultaneously fragmented and tagged with Mate Pair adapters in a Transposon
159 based Tagmentation step. Tagmented DNA was then purified using AMPure XP Magnetic
160 beads (Beckman Coulter Life Sciences, Indianapolis, IN, U.S.A.) followed by Strand
161 Displacement to fill gaps in the Tagmented DNA. Strand displaced DNA was further purified
162 with AMPure XP beads before size-selecting the 3-5 Kilobases (Kb), 5-7 Kb & 7-10 Kb

163 fragments on low melting agarose gel. The fragments were circularized in an overnight blunt-
164 end intra-molecular ligation step, which will result in circularization of DNA with the insert
165 flanked mate pair adapter junction.

166

167 The circularized DNA was sheared using Covaris S220 sonicator (Covaris, Woburn,
168 Massachusetts, USA) to generate approx. fragment size distribution from 300 bp to 1000 bp.

169 The sheared DNA was purified to collect the Mate pair junction positive fragments using
170 Dynabeads M-280 Streptavidin Magnetic beads (Thermo Fisher Scientific, Waltham, MA,
171 U.S.A.). The purified fragments were end-repaired, adenylated and ligated to Illumina
172 multiplex barcode adaptors as per Nextera Mate Pair Sample Preparation Kit protocol.

173

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

183

184 **Nanopore MinION library preparation and sequencing**

185 Genomic DNA (1.5µg) was end-repaired (NEBnext ultra II end repair kit, New England
186 Biolabs, MA, USA), cleaned up with 1x AmPure beads (Beckmann Coulter,USA). Adapter
187 ligation were performed for 20 minutes using NEB blunt/ TA ligase (New England Biolabs,

188 MA, USA). Library mix were cleaned up using 0.4X AmPure beads (Beckmann Coulter,
189 USA) and eluted in 25 µl of elution buffer. Eluted Library were used for sequencing. Whole
190 genome library were prepared by using ligation sequencing kit SQK-LSK108-Oxford
191 Nanopore Technology (ONT) from Oxford Nanopore Technology. Sequencing were
192 performed on MinION Mk1b (Oxford Nanopore Technologies, Oxford, UK) using SpotON
193 flow cell (FLO-MIN106) in a 48hr sequencing protocol on MinKNOW 1.1.20 from ONT.

194

195 **Illumina raw data QC and processing**

196 The Illumina reads were de-multiplexed using Illumina bcl2fastq. The Illumina generated
197 raw data for genomic libraries was quality checked using FastQC (Andrews, S., 2010). The
198 paired-end Illumina reads were processed for clipping the adapter and low-quality bases
199 using customized script which retains minimum 70% bases/reads with Phred score ($Q \geq 30$ in
200 each base position) with a read length of 50 bp. The MP libraries were trimmed for adapter
201 and low-quality base trimming from the 3'-end using PLATANUS internal trimmer (Kajitani,
202 R., et al.).

203

204 **Nanopore reads base calling and processing**

205 Base calling was performed using Metrichor V.2.43.1 is a cloud based analysis tool provided
206 by Oxford Nanopore Technology software suite. The Nanopore reads were processed using
207 Poretools (Loman, NJ., et al.) for converting fast5 files to fasta format. For further
208 quantification and analysis the 2D reads or 1D high quality reads were selected for further
209 assembly.

210

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

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

212 The quality checked Nanopore reads were error-corrected using Illumina PE reads. For error-
213 correction the Illumina PE-reads were aligned to the Nanopore reads by using BWA aligner
214 (Li, H., et al.). The paired-end reads were assembled using Abyss (Birol. I., et al.) followed
215 by contig extension using Nanopore reads using SSPACE-LongRead (Boetzer, M., et al.).
216 Super scaffolding of the assembled scaffold was performed using SSPACE (Boetzer, M., et
217 al.) and PLATANUS using the Nanopore and Matepair data. Final draft genome resulted
218 after gap closure by GAPCLOSER and PLATANUS gap_close tool using Illumina data. The
219 genome size was estimated using a k-mer distribution plot using JELLYFISH (Marcais, G., et
220 al.). The repetitive elements were identified in the final assembled draft genome using Repeat
221 Masker tool. The draft genome was hardmasked by using reference genomic repeats of *G.*
222 *gallus*. The assembly and annotation workflow overview has been represented as Figure 2.

223

224 **Simple sequence repeats prediction**

225 Final assembled scaffolds were analysed for Simple Sequence Repeats (SSR) identification.
226 SSRs like the di, tri, tetra, penta and hexa-nucleotide repeats in the genome were obtained
227 using MISA (Version 1.0.0).

228

229 **Genome prediction and annotation**

230 Gene models was predicted on a hard masked draft genome, where the repetitive elements in
231 the draft genome were masked using genomic repeats of *G. gallus* with Repeatmasker tool
232 and further genes were predicted using AUGUSTUS with *G. gallus* as a reference model.
233 The predicted proteins were annotated by using BLASTP (Altshul, S., et al.) against all
234 Avian sequences downloaded from UniProt Protein Database.

235

236 **Pathway Analysis of the draft genome**

237 The predicted proteins were searched against the KEGG-KAAS server (Moriya, Y., et al.) for
238 pathway analysis. *G. gallus* (chicken), *Meleagris gallopavo* (turkey), *Taeniopygia guttata*
239 (zebra finch), *Falco peregrinus* (peregrine falcon) were used as reference organism for
240 pathway identification. The EuKaryotic Orthologous Groups (KOGs) were predicted using
241 homology based approach.

243 **Mitochondrial genome assembly and annotation**

244 The generated scaffolds from the draft assembly were aligned against the *P. cristatus*
245 mitochondria genome and the mapped reads were filtered and stitched using ABACUS
246 software using the same reference (Zhou, TC., et al.). Further gap closure were performed
247 with 3-7kb MP reads to generate an complete assembled mitochondrial genome. MITOS
248 (Bernt, M., et al.) was used for gene annotation. Circular plot generated using GenomeVx
249 (<http://wolfe.ucd.ie/GenomeVx/>) representing the localization of the gene in the assembled
250 mitochondrial genome.

252 **Phylogenetic tree construction**

253 The assembled Peacock mitochondrial genome was searched against 695 avian mitochondrial
254 genomes downloaded from NCBI. Based on the Blast-N homology results (with query
255 coverage > 100, subject coverage > 95, % identity >85 and with 1% gaps allowed in the
256 sequences). 51 mitochondrial genome sequences along with our assembled mitochondrial
257 genome were filtered. Multiple sequence alignment with default parameters were performed
258 using MUSCLE global sequence aligner. Phylogenetic trees were constructed using IQ-
259 TREE version 1.5.6 (www.iqtree.org). The parameters used for phylogenetic tree
260 construction were ultrafast bootstrap (UFBoot, using the -bb option of 1000 replicates), and a
261 standard substitution model (-m MFP) was given for tree generation. The generated trees

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

262 from IQ-TREE tool were visualized using Figtree (<http://tree.bio.ed.ac.uk/software/figtree/>)
263 and the Brach-support values were recorded from the output “.treefile”. The trees were
264 modified for better visualization under Trees section increasing order nodes were applied.

265 266 **Protein domain analysis**

267 Predicted proteins from Peacock, Chicken and Turkey with sequence length greater than 100
268 amino acids were considered for protein domain analysis. All the protein sequences from
269 each organism were searched against Pfam-A database using Pfam scan for protein domain
270 identification.

271 272 **Avian protein families**

273 The protein sequences of 48 avian genomes was downloaded from the link
274 <http://avian.genomics.cn/en/jsp/database.shtml> apart from the predicted proteins of the draft
275 genome. Sequences greater than 100 amino acids from all the avian genomes were filtered
276 and concatenated to a single fasta file. These sequences were clustered using CD-HIT (Fu, L.,
277 et al.) with 90% alignment coverage for the shorter sequence with a length difference cutoff
278 of 0.9. The single copy ortholog gene family present across all organisms and genes unique to
279 peacock were filtered and annotated.

280 281 **Genome conservation analysis**

282 The assembled draft genome was aligned against the *G. gallus* genome using Chromosomer
283 tool. Draft chromosomes were constructed based on alignment between fragments (scaffolds)
284 using a reference genome. These reordered assembled genome was aligned against the
285 Chicken genome using LAST aligner with NEAR (finding short-and-strong (near-identical)
286 similarities.) parameter allowing for substitution and gap frequencies leading to the

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

287 identification of orthologs. These query mapped regions were filtered with greater than 1% of
288 the maximum length for visualization using Circos.

289

290

291 **Results**

292 **DNA Sequencing data**

293 Five libraries of 150 bp paired-end from Illumina HiSeq technology were generated. The
294 short-insert reads of 489,114,747 accounted to genome coverage of 146.7X and long-insert
295 reads of 302,884,819 sequences was about 90.9X coverage with a total coverage of 236X.
296 Sequencing of three mate-pairs of 3-5Kb, 5-7Kb of and 7-10Kb yielded 72,915,033,
297 47,440,144 and 36,464,628 reads respectively with an approximate coverage of 21.9, 14.2
298 and 10.9 respectively, with a total of 156 million reads of 47X coverage. Oxford Nanopore
299 technology was used to generate 366,323 long reads having of 2,398,560,283 bp with
300 coverage of 2.3X. The complete sequencing was generated to a depth of ~287X from
301 Illumina and Nanopore platforms. The coverage was based on assuming the peacock genome
302 size of about 1 Gb (Table S1).

303

304 **Genome assembly**

305 The assembly was performed on Illumina reads with Abyss *de novo* assembler that resulted
306 in ~932 Mb (mega base) of genome with an N50 of 1639 bp. The extension of the contigs
307 were performed with Nanopore reads which generated scaffolds with N50 of 14,748 bp.
308 Super scaffolding of the assembled scaffold was performed using SSPACE and PLATANUS
309 with MP libraries that generated ~916 Mb genome with the N50 value of 168,140bp. The
310 final gap closer was executed using GAPCLOSER program with MP and PE-LI libraries
311 which generated a draft genome of 1.02 GB (giga base). The draft genome assembly of *Pavo*

312 *cristatus* consists of 179,346bp scaffolds, with a N50 of 189,886bp with 37 scaffolds having
313 sequence length \geq 1Mbp. Contigs above 5000 bp have covered a genome of ~0.915 Mb with
314 N50 0.23 Mb. In the assembled genome there were ~0.4% of non ATGC characters (Table
315 1).

316
317 Complete mitochondrial genome of 16699 bp was obtained. Total of 22 tRNA, tow rRNA
318 and 13 protein coding genes were identified in the assembled genome (Fig. 3a). A 100%
319 similarity was observed with the preiously published Peacocks mitochondrial genome (Fig.
320 S1a and S1b).

321

322 **Transposable Elements (TE)**

323 In the bird genome a total of 75,315,566 bp (7.3% of the genome) was predicted to have
324 5.5% of Retroelements (with SINEs 0.08% and LINEs 4.71%), 6.25 % total interspersed
325 repeats with 0.84 % simple repeats and 0.21% low complexity regions. The DNA
326 transposons identified in the genome was 0.71% (Table S2).

327

328 **Protein coding gene annotation**

329 A total of 23,153 proteins were predicted in the assembled draft genome using AUGUSTUS.
330 Among them 95% predicted genes were annotated against the other Aves proteins. The
331 21,854 annotated proteins showed top similarity to species *Gallus gallus* (Chicken) with
332 11,398 proteins, *Meleagris gallopavo* (Common turkey) with 4059 proteins, *Amazona aestiva*
333 (Blue-fronted Amazon parrot) with 1352 proteins and *Anas platyrhynchos* (Mallard) (*Anas*
334 *boschas*) with 849 proteins. Thirteen species had about 100 to 400 annotated proteins. The
335 remaining proteins were in the range of 1 to <100 proteins in about 62 species. From the

336 annotations a total of 13,161 proteins showed similarity to uncharacterized protein annotation
337 and some of the over represented proteins were Tyrosine-protein kinase, Sulfotransferase,
338 Phosphoinositide phospholipase C, Tetraspanin, Phospholipid-transporting ATPase,
339 Olfactory receptor, Polypeptide N-acetylgalactosaminyltransferase, Transporter, Keratin,
340 Hexosyltransferase, Protein Wnt, Kinesin-like protein, Gap junction protein, Claudin, POU
341 domain protein, Sodium/hydrogen exchanger, Phospholipid-transporting ATPase, Histone-
342 lysine N-methyltransferase and others (Table S3). The gene ontology annotations showed to
343 have Gene Ontology (GO) descriptions for 18,295 proteins. Among them, 14,490 proteins
344 have Molecular Function; 11,679 have Biological Process and 13,736 proteins have Cellular
345 Component as functional categories.

346 About 17.7% of proteins were found to have pathway information against the KEGG
347 database (Table S4). Some of the overrepresented annotations were Kinases like MAPK
348 (mitogen-activated protein kinase); JNK (c-Jun N-terminal kinase); RAF (RAF proto-
349 oncogene serine/threonine-protein kinase); AKT (RAC serine/threonine-protein kinase);
350 protein kinases and different GTPases.

351 Proteins searched against the KOG annotations showed a total of 20,937 proteins having
352 annotations. Among them, the most abundant annotations include Zn-finger, transmembrane
353 receptor, ubiquitin ligase, Leucine rich repeat, Cadherin repeats, Serine/threonine protein
354 kinase, Collagens, Ankryin repeat, Fibrillins, Voltage-gated Ca²⁺ channels and Hormone
355 receptors (Table S5). The peacock proteins when searched against the human proteins
356 showed expansions in ontologies for cell morphogenesis, neuronal projection and
357 development and GTPases (Table S10 and Fig. S4).

358

359

360 **Simple sequence repeats**

361 A total of 399,493 SSRs were identified from the peacock genome assembly. The largest
362 fraction of SSRs identified were mononucleotide (60.04%), followed by tetra nucleotide
363 (26%), di nucleotide (8.51%), tri nucleotide (4.31%), penta nucleotide (1.03%) and finally
364 hexa nucleotide (0.13%). Among the SSRs identified, A (49.2%) and T (44.9%) accounted
365 for 94.1% of the mono-nucleotide repeats. AT (23.8%), TA (16.5%), TG(13.7%), AC(10.6%)
366 and CA (10.32%) accounted for 75% of the di-nucleotide repeats. while TTG (9.9%), AAT
367 (9.6%), AAC (9.4%), TTA (7.1%), ATT (4.5%), TAA (3.5%), CAA (3.1%) and GGA
368 (2.69%) accounted for 49.7% of the tri-nucleotide repeats (Table S6).

369

370 **Avian protein family analysis**

371 A total of 748,544 protein sequences from 49 avian species have 653,497 protein sequences
372 of length above 100 amino acids (Table S7A). A total of 240,853 gene clusters were
373 generated of which 41 gene clusters had single copy orthologs in all avian species (Table S7B
374 and Table S7C). With the above stringent cutoff we observed 15,913 gene clusters were
375 unique to peacock species.

376

377 **Phylogeny and Genome comparisons**

378 The phylogeny of 51 mitochondrial genome sequences along with peacock genome showed a
379 clade consisting of *Pavo* species and *Gallus* (red junglefowl, Sri Lankan junglefowl, grey
380 junglefowl, grey junglefowl), *Bambusicola* (Mountain Bamboo-partridge, Chinese bamboo
381 partridge) and *Francolinus* (Chinese francolin). It can be observed many species were
382 distributed in two different clades where 34 species were found in one clade and five species
383 in other clade. Some of the endemic or native bird species like *Arborophila ardens*, *Acryllium*
384 *vulturinum* and *Numida meleagris* were found as clear outgroup of species in this study (Fig.
385 3b).

1 386 Predicted proteins from Peacock, Chicken and Turkey were searched for protein domain
2 387 analysis. 81% of the Pfams were common among the three species. About 94%, 98.4% and
3
4 388 99.7% predicted Pfam domains were identified in Peacock, Chicken and Turkey respectively.
5
6
7 389 There were 255, 69 and 14 Pfam domains found to be unique in the species mentioned above
8
9
10 390 respectively (Fig. 4).

11
12 391 The assembled Peacock genome was reordered for pseudo chromosomes generation against
13
14 392 the simple repeat masked Chicken genome (1.21 GB, Warren, WC., et al.) using
15
16
17 393 Chromosomer which generated a overall reordered Peacock genome of about 597MB. The
18
19 394 right side of the image represents the reference genome and left side of the image represents
20
21
22 395 the Peacock genome (Fig. 5).

23
24 396

25 26 397 **Conclusions**

27
28
29 398 Third generation sequencing in avian genomics where long reads having the substantiality to
30
31 399 improve genome assembly will benefit understanding the organisms in the structurally
32
33
34 400 complex regions having repeat elements and isoforms in the genome (Goodwin, S., et al.).
35
36 401 Using a combination of short reads of different insert sizes as well as mate pair reads
37
38
39 402 generated from Illumina technology along with long reads from Oxford Nanopore, we
40
41 403 obtained an improved assembly and a draft genome of the Indian Blue Peacock (*Pavo*
42
43 404 *cristatus*). In comparison to other avian genomes (Zhang, G., et al.), the current 290X
44
45
46 405 sequencing depth obtained from our study is one of the highest. With a N50 of 0.23Mb, we
47
48
49 406 presented here a reasonably reliable draft genome for the peacock species. The inclusion of
50
51 407 Nanopore reads 366,323 for scaffolding followed by subsequent gap-closing using Illumina
52
53 408 data led to a 26.2% reduction in the number of scaffolds and a 50.65% and 115% increase in
54
55
56 409 the scaffold and contig N50, respectively. With only 2.3X of long reads, a significant
57
58 410 improvement in the assembly was observed. On the contrary, the assembly contained less
59
60
61
62
63
64
65

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

411 than 0.4% of unknown nucleotides, which is very low for a draft assembly. With the addition
412 of more long reads along with transcriptomic sequencing along with scaffolding and/or gap
413 closure tools, further improvement in the assembly can be achieved.

414
415 Peacock seems to defy the Darwinian laws of natural selection. These concern were raised by
416 no other than Darwin himself. Hence, he proposed the theory of the sexual selection where
417 the female can choose for a male with a certain phenotypic feature such as brilliant color or a
418 long tail (Burgess, S.). Peacock's brilliantly colored long tail feathers seems to evolve at the
419 cost of finding its female partner thereby contributing its beneficial genes, even at the cost of
420 making itself venerable to predators. A female peafowl in turn tends to choose the mate with
421 the largest and decorated plumage, which indirectly reflects its healthiness and capacity to
422 wade off potential competitors thereby selecting the most suitable male. Peacocks beautiful
423 feathers with it all its artistry surely provides it with an advantage to impress the females
424 (Dakin, R., et al.). Understanding the formation of beautiful feathers from the genomic
425 context will help in resolving several evolutionary theories on sexual selection that have been
426 discussed on this species.

427 Pigment particles are embedded into the newly grown peacock feathers during the molting
428 season which seems to absorb light of selective wavelength there by imparting to the color of
429 the plumage (Mercedes Foster, Rennee Riedler, et al.) Pigment morphogenesis in *Gallus* has
430 been reported by a process of melanoblasts migration and colonization into feather bud where
431 they differentiate to produce the pigment, melanin (Kelsh, RN., et al.). The molecular
432 mechanisms that control the pigment cell migration have been narrowed down to proteins
433 Kit1 and FGFs which maintain the melanoblasts migration to feathers. Kit and FGF proteins
434 have also been identified in our current study (Table S3). Understanding of these proteins in
435 the patterned formation will help decode the pigment pattern morphology in peacocks. These
436 pigmented patterns play a role in communication, choice of mate and in some species it can
437 help in camouflage (Burgess, S)

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

438 It has been observed that the variations in the genome size among bird species are very low
439 (Table S8). The genome complexities of a species are influenced by the Transposable
440 elements (TE) that are believed to play a crucial role (Kapsuta, A., et al.). The long read
441 sequences have significantly helped in resolving the TEs in genome quality and assembly.
442 Peacock genome comparisons with Turkey and Chicken have showed closeness to the
443 Chicken species. The mitochondrial phylogeny also revealed similar findings. Homology
444 searches have shown several important gene family expansions such as Kinases, Zn finger
445 proteins, GTPases and others. Their roles in biology, development and evolution of the
446 Peacocks have to be further explored.

447 To summarize, we have assembled the *Pavo* genome using Illumina and Oxford nanopore
448 technology. The genome information can be valued and explored by avian enthusiasts to
449 further understand about this bird. Though not critically endangered yet, in India, peafowl
450 population is surely at a decline in the wild due to massive deforestation and habitat loss.
451 Thus is further compounded by increased poaching for meat and feathers. Our genome
452 sequencing initiative of *Pavo cristatus* is not just only from a conservational viewpoint, but
453 also to preserve a heritage associated with this bird that runs through centuries and that bears
454 a strong attachment to the national psyche.

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

456 **Availability of supporting data**

457 Supplementary data contains, read statistics, annotation, repeats identification, orthology
458 analysis, mitochondria assembly and annotation. Figures, Gene ontology. DNA and library
459 preparation protocols.

460

461 **Raw Data in SRA**

462 Raw reads (Illumina and Nanopore) are available in the Sequence Read Archive (SRA), and
463 the Whole Genome Shotgun project has been deposited at GenBank under SRA Submission
464 ID: SUB3108024, Bioproject: PRJNA413288 and Biosamples SAMN07739105 :
465 SKPea2016_SI, SAMN07739104 : SKPea2016_LI, SAMN07739101 : FPL_3_5KB,
466 SAMN07739102 : FPL_5_7KB, SAMN07739103 : FPL_7_10KB and SAMN07739107 :
467 FPL_Nano.

468

469 **Competing interests**

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

471

472 **Authors contributions**

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

477

478

479

480

481 **Acknowledgements**

1
2 482 Department of Biochemistry, AIIMS, New Delhi for providing space and infrastructure to
3
4
5 483 carry on the work. RD for providing partial funding. Kritika Tripathi for inputs with
6
7 484 manuscript editing. Raja and Sudha from Genotypic Technology Pvt. Ltd for providing
8
9
10 485 sequencing services and also Deepti, Shalini, Nagu, Abhi and Manish. BR acknowledges IUT
11
12 486 34-4.

13
14 487

15
16
17 488
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

489 **Tables**

490 Table 1. *De novo* assembly statistics of the Peacock genome.

Description	Contigs	Nanopore Scaffold	Super Scaffolds	GapClosed	>1000 Kb	>5000 Kb
Contigs	685241	281272	179346	179346	34181	15026
Maximum Length	49159	251510	2390121	2488982	2488982	2488982
Minimum Length	300	5	265	265	1000	5000
Average Length	1360	3250	5111	5729	-	-
Total Length	932162464	914363908	916720956	1027551907	954483822	915373606
Length >= 100 bp	685241	281271	179346	179346	34181	15026
Length >= 200 bp	685241	281271	179346	179346	34181	15026
Length >= 500 bp	616120	186433	93727	93727	34181	15026
Length >= 1 Kbp	363428	104479	34168	34181	34181	15026
Length >= 10 Kbp	1591	24748	9249	10311	10311	10311
Length >= 1 Mbp	0	0	27	37	37	37
Non-ATGC #	350325	42696911	49169831	4034372	4032567	3978757
Non-ATGC %	0.038	4.67	5.364	0.393	0.422	0.435
N50 value	1639	14748	168140	189886	218023	232312

491

492

493 **Figure legend**

1
2 494 **Figure 1.** The beautiful and charismatic photo of Indian blue peacock (*Pavo cristatus*) bird.

3
4
5 495 **Figure 2.** Detailed workflow for *de novo* genome assembly and annotation.

6
7 496 **Figure 3a.** Circular representation of Peacock mitochondrial genome with genes predicted.

8
9
10 497 **Figure 3b.** Phylogenetic tree generated from mitochondrial genome from 52 different avian
11 species.

12
13
14 499 **Figure 4.** Circular image of the assembled peacock genome aligned against the *G. gallus*
15 genome using Chromosomer tool. Draft chromosomes were generated by similarity between
16 scaffolds which were arranged on the reference chicken genome. Circos was used for
17 500
18
19 501
20
21
22 502
23 visualization.

24 503 **Figure 5.** Venn diagram showing common and unique Protein family domains (Pfam)
25
26 504
27 between Peacock, Chicken and Turkey proteins.

28
29 505
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

506 **References:**

1
2 507 Gadagkar, R., 2003. Is the peacock merely beautiful or also honest?. *Current Science*, 85(7),
3
4 508 pp.1012-1020.

5 509
6
7 510 Kushwaha, S., and Kumar, A. 2016. A Review on Indian Peafowl (*Pavo cristatus*) Linnaeus,
8
9 511 1758. *Journal of Wildlife and Research*, 4, 42-59.

10 512
11
12 513 Kadgaonkar, Shivendra B. 1993. The peacock in ancient Indian art and literature. *Bulletin of*
13
14 514 the Deccan College Research Institute, vol. 53, pp. 95–115. JSTOR,
15
16 515 www.jstor.org/stable/42936434.

17 516
18
19 517 Brickle, N. 2002. Habitat use, predicted distribution and conservation of green peafowl (*Pavo*
20
21 518 *muticus*) in Dak Lak Province, Vietnam. *Biological Conservation*, 105: 189-197.

22
23 519
24
25 520 Jackson, C. 2006. *Peacock*. London: Reaktion Books Ltd.

26 521
27
28 522 Zhang, G., Jarvis, E. D., and Gilbert, M. T. P. 2014. A flock of genomes. *Science* 346, 1308–
29
30 523 1309.

31 524
32
33 525 International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative
34
35 526 analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*,
36
37 527 432(7018), 695.

38 528
39
40 529 Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M.,
41
42 530 Harada, M., Nagayasu, E., Maruyama, H. and Kohara, Y., 2014. Efficient de novo assembly
43
44 531 of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*,
45
46 532 24(8), pp.1384-1395.

47 533
48
49 534 Loman, N. J. and Quinlan, A. R. 2014. Poretools: a toolkit for analyzing nanopore sequence
50
51 535 data. *Bioinformatics*, 30(23), 3399-3401.

52 536
53
54 537 Li H. and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
55
56 538 Transform. *Bioinformatics*, 25:1754-60.

57 539

540 Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao,
1 541 Y., Hirst, M., Schein, J.E. and Horsman, D.E., 2009. De novo transcriptome assembly with
2 542 ABySS. *Bioinformatics*, 25(21), pp.2872-2877.
3
4 543
5
6 544 Boetzer, Marten, and Walter Pirovano. 2014. SSPACE-LongRead: scaffolding bacterial draft
7 545 genomes using long read sequence information. *BMC bioinformatics* 15.1: 211
8
9 546
10 547 Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W., 2010. Scaffolding pre-
11 548 assembled contigs using SSPACE. *Bioinformatics*, 27(4), pp.578-579.
12
13 549
14 550 Marcais, G and Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting
15 551 of occurrences of k-mers. *Bioinformatics* 27(6): 764-770.
16
17 552
18 553 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local
19 554 alignment search tool. *Journal of molecular biology*, 215(3), pp.403-410.
20
21 555
22 556
23 557 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. and Kanehisa, M., 2007. KAAS: an
24 558 automatic genome annotation and pathway reconstruction server. *Nucleic acids research*,
25 35(suppl_2), pp.W182-W185.
26
27 559
28 560 Zhou, T.C., Sha, T., Irwin, D.M. and Zhang, Y.P., 2015. Complete mitochondrial genome of
29 561 the Indian peafowl (*Pavo cristatus*), with phylogenetic analysis in phasianidae. *Mitochondrial*
30 562 *DNA*, 26(6), pp.912-913.
31
32 563
33 564 Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsche, G., Pütz, J.,
34 565 Middendorf, M. and Stadler, P.F., 2013. MITOS: improved de novo metazoan mitochondrial
35 566 genome annotation. *Molecular phylogenetics and evolution*, 69(2), pp.313-319.
36
37 567
38 568 Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W., 2012. CD-HIT: accelerated for clustering the
39 569 next-generation sequencing data. *Bioinformatics*, 28(23), pp.3150-3152.
40
41 570
42 571 Warren, W.C., Hillier, L.W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., Markovic,
43 572 C., Bouk, N., Pruitt, K.D., Thibaud-Nissen, F. and Schneider, V., 2016. A new chicken
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

573 genome assembly provides insight into avian genome structure. G3: Genes, Genomes,
1 574 Genetics, pp.g3-116.
2
3 575
4
5 576 Goodwin, S., McPherson, J.D. and McCombie, W.R., 2016. Coming of age: ten years of
6 next-generation sequencing technologies. Nature Reviews Genetics, 17(6), p.333.
7 577
8
9 578
10
11 579 Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., Storz, J.F., Antunes, A., Greenwold,
12 M.J., Meredith, R.W. and Ödeen, A., 2014. Comparative genomics reveals insights into avian
13 580 genome evolution and adaptation. Science, 346(6215), pp.1311-1320.
14 581
15
16 582
17
18 583 Burgess, S., 2001. The beauty of the peacock tail and the problems with the theory of sexual
19 selection. Journal of Creation, 15(2), pp.94-102..
20 584
21
22 585
23
24 586 Dakin, R. 2008. The role of the visual train ornament in the courtship of peafowl, *Pavo*
25 *cristatus*. Masters Abstracts International, 47/03: 97.
26 587
27 588
28
29 589 Mercedes S Foster 1975. The overlap of molting and breeding in some birds. The Condor
30 77:304-314.
31 590
32
33 591
34
35 592 Renee Riedler, Christel Pesme, James Druzik, Molly Gleeson, Ellen Pearlstein The Journal of
36 the American Institute of Conservation 2014
37 593
38 594
39
40 595
41
42 596 Kelsh, R.N., Harris, M.L., Colanesi, S. and Erickson, C.A., 2009. Stripes and belly-spots—a
43 review of pigment cell morphogenesis in vertebrates. Seminars in cell & developmental
44 597 biology. Vol. 20, No. 1, pp. 90-104.
45 598
46
47 599
48
49 600 Kapusta, A. and Suh, A., 2017. Evolution of bird genomes—a transposon's- eye
50 view. Annals of the New York Academy of Sciences, 1389(1), pp.164-185.
51 601
52
53 602
54
55 603
56
57 604 **Webservers:**
58 605 FastQC : <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
59
60 606 AUGUSTUS : <http://augustus.gobics.de/>
61
62
63
64
65

607 PLATANUS : <http://platanus.bio.titech.ac.jp/>
1
2 608 ABACUS : <http://abacas.sourceforge.net/Manual.html>
3
4 609 MISA: <http://pgrc.ipk-gatersleben.de/misa/>
5
6 610 REPEATMASKER : <http://www.repeatmasker.org/>
7
8 611 KOGs : <https://genome.jgi.doe.gov/portal/help/kogbrowser.jsf>
9
10 612 KAAS : <http://www.genome.jp/tools/kaas/>
11
12 613 MITOS : <http://mitos.bioinf.uni-leipzig.de/>
13
14 614 MUSCLE : <https://www.drive5.com/muscle/>
15
16 615 PFAMSCAN : <https://www.ebi.ac.uk/seqdb/confluence/display/THD/PfamScan>
17
18 616 LAST : <http://last.cbrc.jp/>
19
20 617 CHROMOSOMER : <https://github.com/gtamazian/chromosomer>
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 ***De novo* genome assembly of Indian Blue Peacock (*Pavo***
2 ***cristatus*), from Oxford Nanopore and Illumina sequencing**
3 **reads**

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,*} and Subhradip
9 Karmakar^{1,*}

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

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

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

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

24 **Abstract**

25 **Background**

1
2 26 *Pavo cristatus* the Indian blue peacock are geographically found distributed in natural habitats
3
4 27 of South Asia. Peacock has been described among the bird species as one of the most elegant,
5
6 28 majestic and beautiful bird (Fig. 1). Since prehistoric times they have been described in Indian
7
8 29 culture and has been adopted as the national bird of India. Its length varies from 92-125
9
10 30 centimeter (without train), weighing about 4-8 Kilograms and lives up to 20 years in the wild.
11
12 31 The avian species have been very important in the fields of phylogenetics, developmental
13
14 32 studies, sexual reproduction and speciation. Avian genomics have contributed immensely
15
16 33 towards understanding the vertebrate genome evolution. Here we present the first draft genome
17
18 34 sequence of *P. cristatus*, yet another important bird species to further add values and gain
19
20 35 insight into avian genomics.
21
22 36

24 37 **Findings**

26 38 For the first time in avian genomics, long reads using Oxford Nanopore technology have been
27
28 39 used for the whole genome assembly. We sequenced different DNA insert size libraries from
30
31 40 Illumina and long read Nanopore technologies from the peacock DNA. We performed *de novo*
32
33 41 genome assembly by integrating the reads from Illumina short insert, long insert, multiple
34
35 42 mate-pair reads along with Nanopore long reads using multiple genome improvement tools. A
36
37 43 draft of the peacock genome of about 0.915 Gigabases (Gb) with a N50 of 0.23 Megabases
38
39 44 (Mb) was assembled. Annotations with other avian species, protein families, KEGG were
40
41 45 performed for functional understanding by insilico approaches. Proteins were compared
42
43 46 against Chicken, Turkey and Human to obtain evolutionary similarities and uniqueness of the
44
45 47 *Pavo* species.
46
47 48

53 49 **Conclusions**

55 50 Our most important findings from the genome sequence of *P. cristatus* is to decipher the
56
57 51 different gene families and to understand their role in body pattern development and other
58
59
60
61
62
63
64
65

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 features that truly makes this bird unique. The genome sequence also gives insights on its
53 genetic lineage and evolution with relation to other avian members. Several hypothesis and
54 theories have been discussed with respect to sexual selection; now with the understanding of
55 the genome sequence, some of these evolutionary theories will be better understood. The
56 genome will also support future studies on population genetics and breeding for species
57 conservation as well as in understanding its evolutionary ecology and sexual dimorphism. The
58 comparative genomics with other avian species and specifically with *Gallus gallus* (Chicken)
59 and *Meleagris gallopavo* (Turkey) have shown insights into the gene families and their
60 conserved domains. *Pavo* proteins were also compared with human to understand the
61 functional components that were conserved after the speciation split.
62

63 **Introduction**

64 *Pavo cristatus* commonly known as the blue Indian peacock obtained the status of National
65 Bird of India in 1963. Peacocks have been distributed widely in Asian countries. The bird is
66 very popular as it symbolizes beauty, love, grace and pride (Gadagkar, R; Kushwaha, et al.). It
67 has been referred in ancient literatures of India and has been found closely associated with the
68 life and culture of the peoples from South East Asia and particularly India (Kadgaonkar, SB).
69 Peacocks are native to South Asia but have been introduced into many other countries usually
70 as exhibits in park, zoos and also large number of aviculturists raise and breed these species as
71 pets (Brickle, N; Jackson, C).

72
73 International Chicken Genome Sequencing Consortium sequenced the *Gallus gallus* genome
74 in 2004, and this laid the foundation for study of avian genomics. A decade later, the avian
75 genome consortium assembled 48 genomes of wide variety of avian species to understand the
76 evolutionary relationships of modern birds (Zhang, G., et al.). Raw sequencing data for each
77 species were generated at from coverage of 6X for zebra finch to a maximum 160X for
78 budgerigar. The genome size varied from 1.04 to 1.26 Gb (<http://avian.genomics.cn/en/>). The
79 studies on different bird species have provided a new perspective on vertebrate genome
80 evolution. These genomes have also helped in improving the annotation of mammalian
81 genomes. There are several distinguishing as well as unique features between chicken and
82 human genome including genome size which is of one third of humans, conserved synteny
83 blocks complete absence of active short interspersed nucleotide elements (SINE) to mention a
84 few (International Chicken Genome sequencing consortium).

85
86 Despite the wealth of information from avian genomes sequencing projects, it is very important
87 to genome sequence other new species to add value into aves and vertebrate genomics. For the

1 88 first time in bird genomics, Oxford Nanopore technology has been used to sequence a bird
2 89 genome in this present study. The long read chemistry will help in better genome assembly of
3
4 90 the TEs and repeat rich. The peacock genome will aid in understanding about the uniqueness
5
6
7 91 of this birds genome in comparison to other bird species. Comparative genomes will help in
8
9
10 92 understanding the development of this species, sexual selection and its evolutionary
11
12 93 relationships with other birds. The characterization of the genes involved in sex determination
13
14 94 could provide relevant information for the selective breeding of the peafowl populations. We
15
16
17 95 have unraveled some of the genomic signatures and thus have reported unique gene pools of
18
19 96 this bird by performing comparative genomics. Further different data types will improve the
20
21
22 97 assembly and gene/genome characterization will help to address the sexual selection theory
23
24 98 and key answers relevant to the evolution of this bird.

25
26 99

27
28
29 100

30 31 101 **Materials and methods**

32 33 102 **Sample collection and extraction of DNA**

34 103 The whole blood of male peacock was collected from Kanpur zoo, India after obtaining the
35
36 104 necessary ethical and institutional approval. 20µl of Proteinase K (PK) solution was taken into
37
38
39 105 a 1.5ml micro centrifuge tube. 200µl of blood was added and briefly mixed. 200µl of cell lysis
40
41
42 106 buffer was added to the tube, mixed by vortexing for 10seconds; incubated at 56°C for
43
44
45 107 10minutes. ReliaPrep™ Binding Column was placed into an empty collection tube. 250µl of
46
47
48 108 Binding Buffer (BBA) was added, capped the tube, and mixed by vortexing for 10 seconds
49
50
51 109 with a vortex mixer. Contents of the tube were added to the ReliaPrep™ binding column,
52
53
54 110 capped and placed in a refrigerated micro centrifuge. These were then centrifuged for 1 minute
55
56 111 at maximum speed and flow through was discarded. Binding column was placed into a fresh
57
58 112 collection tube. 500µl of column wash solution was added to the column and centrifuged for 3
59
60
61
62
63
64
65

113 minutes at maximum speed; Flow through was again discarded. Column washing is repeated
114 thrice. Columns were then placed in a nuclease free clean 1.5ml micro centrifuge tube. 100 µl
115 of Nuclease-Free Water was then added to the column and centrifuged for an additional 1
116 minute at maximum speed. Column was discarded and the elute was saved. The concentration
117 and purity of the extracted DNA was evaluated using Nanodrop Spectrophotometer (Thermo
118 Scientific) and Qubit flurometer and integrity was checked on a 0.8% agarose gel. The DNA
119 sample was aliquoted for library preparation on two different platforms: Illumina HiSeq4000
120 and Oxford Nanopore Technologies (ONT).

121

122 **HiSeq Paired-End library preparation and sequencing**

123 Whole genome sequencing (WGS) libraries were prepared with Illumina-compatible
124 NEXTflex DNA sequencing kit (BIOO Scientific, Austin, Texas, U.S.A.). Briefly, approx. 1
125 µg of genomic DNA was sheared using Covaris S2 sonicator (Covaris, Woburn,
126 Massachusetts, USA) to generate approx. fragment size distribution from 300 to 600 basepair
127 (bp). The fragment size distribution was checked on Agilent 2200 Tape Station with D1000
128 DNA screen tapes and reagents (Agilent Technologies, Palo Alto, CA, USA) and subsequently
129 purified using HighPrep magnetic beads (Magbio Genomics Inc, USA). The purified fragments
130 were end-repaired, adenylated and ligated to Illumina multiplex barcode adaptors as per
131 NEXTflex DNA sequencing kit protocol (BIOO Scientific, Austin, Texas, USA).

132

133 The adapter-ligated DNA was purified with HighPrep beads (MagBio Genomics, Inc,
134 Gaithersburg, Maryland, USA) and then size selected on 2% low melting agarose gel and
135 cleaned using MinElute column (QIAGEN). The resultant fragments were amplified for 10
136 cycles of PCR using Illumina-compatible primers provided in the NEXTFlex DNA sequencing
137 kit. The final PCR product (sequencing library) was purified with HighPrep beads, followed

138 by library quality control check. The Illumina-compatible sequencing library was initially
139 quantified by Qubit fluorometer (Thermo Fisher Scientific, MA, USA) and its fragment size
140 distribution was analyzed on Agilent TapeStation. Finally, the sequencing library was
141 accurately quantified by quantitative PCR using Kapa Library Quantification Kit (Kapa
142 Biosystems, Wilmington, MA, USA). The qPCR-quantified library was subjected to
143 sequencing on an Illumina sequencer for 150 bp paired-end chemistry.

144
145 The Illumina-compatible sequencing library for the samples has a fragment size range between
146 275 to 425 bp for Paired-End Short Insert (PE-SI) and 350 bp to 650bp for Paired-End Long
147 Insert (PE-LI). As the combined adapter size is approximately 120bp, the effective user-defined
148 insert size is 155 to 305 bp and 230 to 530 bp for PE-SI and PE-LI respectively. Libraries were
149 sequencing in Illumina HiSeq platform with 150*2 chemistry. The short reads of Paired-End
150 Short Insert (PE-SI), Paired-End Long Insert (PE-LI) and Mate-Pair (MP) from Illumina HiSeq
151 platform.

152

153

154 **Mate-Pair library preparation and sequencing**

155 Mate Pair sequencing library was prepared with Illumina-compatible Nextera Mate Pair
156 Sample Preparation Kit (Illumina Inc., Austin, TX, U.S.A.). Briefly, approx. 4 ug of genomic
157 DNA was simultaneously fragmented and tagged with Mate Pair adapters in a Transposon
158 based Tagmentation step. Tagmented DNA was then purified using AMPure XP Magnetic
159 beads (Beckman Coulter Life Sciences, Indianapolis, IN, U.S.A.) followed by Strand
160 Displacement to fill gaps in the Tagmented DNA. Strand displaced DNA was further purified
161 with AMPure XP beads before size-selecting the 3-5 Kilobases (Kb), 5-7 Kb & 7-10 Kb
162 fragments on low melting agarose gel. The fragments were circularized in an overnight blunt-

163 end intra-molecular ligation step, which will result in circularization of DNA with the insert
164 flanked mate pair adapter junction.

165
166 The circularized DNA was sheared using Covaris S220 sonicator (Covaris, Woburn,
167 Massachusetts, USA) to generate approx. fragment size distribution from 300 bp to 1000 bp.

168 The sheared DNA was purified to collect the Mate pair junction positive fragments using
169 Dynabeads M-280 Streptavidin Magnetic beads (Thermo Fisher Scientific, Waltham, MA,
170 U.S.A.). The purified fragments were end-repaired, adenylated and ligated to Illumina
171 multiplex barcode adaptors as per Nextera Mate Pair Sample Preparation Kit protocol.

172
173 The adapter-ligated DNA was then amplified for 15 cycles of PCR using Illumina-compatible
174 primers. The final PCR product (sequencing library) was purified with AMPure XP beads,
175 followed by library quality control check. The Illumina compatible sequencing library was
176 initially quantified by Qubit fluorometer (Thermo Fisher Scientific, MA, USA) and its
177 fragment size distribution was analyzed on Agilent TapeStation. Finally, the sequencing library
178 was accurately quantified by quantitative PCR using Kapa Library Quantification Kit (Kapa
179 Biosystems, Wilmington, MA, USA). The qPCR quantified libraries were pooled in equimolar
180 amounts to create a final multiplexed library pool for sequencing on an Illumina sequencer.

181

182 **Nanopore MinION library preparation and sequencing**

183 Genomic DNA (1.5µg) was end-repaired (NEBnext ultra II end repair kit, New England
184 Biolabs, MA, USA), cleaned up with 1x AmPure beads (Beckmann Coulter,USA). Adapter
185 ligation were performed for 20 minutes using NEB blunt/ TA ligase (New England Biolabs,
186 MA, USA). Library mix were cleaned up using 0.4X AmPure beads (Beckmann Coulter, USA)
187 and eluted in 25 µl of elution buffer. Eluted Library were used for sequencing. Whole genome

188 library were prepared by using ligation sequencing kit SQK-LSK108-Oxford Nanopore
189 Technology (ONT) from Oxford Nanopore Technology. Sequencing were performed on
190 MinION Mk1b (Oxford Nanopore Technologies, Oxford, UK) using SpotON flow cell (FLO-
191 MIN106) in a 48hr sequencing protocol on MinKNOW 1.1.20 from ONT.

Illumina raw data QC and processing

194 The Illumina reads were de-multiplexed using Illumina bcl2fastq. The Illumina generated raw
195 data for genomic libraries was quality checked using FastQC (Andrews, S., 2010). The paired-
196 end Illumina reads were processed for clipping the adapter and low-quality bases using
197 customized script which retains minimum 70% bases/reads with Phred score ($Q \geq 30$ in each
198 base position) with a read length of 50 bp. The MP libraries were trimmed for adapter and low-
199 quality base trimming from the 3'-end using PLATANUS internal trimmer (Kajitani, R., et al.).

Nanopore reads base calling and processing

202 Base calling was performed using Metrichor V.2.43.1 is a cloud based analysis tool provided
203 by Oxford Nanopore Technology software suite. The Nanopore reads were processed using
204 Poretools (Loman, NJ., et al.) for converting fast5 files to fasta format. For further
205 quantification and analysis the 2D reads or 1D high quality reads were selected for further
206 assembly.

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

209 The quality checked Nanopore reads were error-corrected using Illumina PE reads. For error-
210 correction the Illumina PE-reads were aligned to the Nanopore reads by using BWA aligner
211 (Li, H., et al.). The paired-end reads were assembled using Abyss (Birol, I., et al.) followed by
212 contig extension using Nanopore reads using SSPACE-LongRead (Boetzer, M., et al.). Super

1 scaffolding of the assembled scaffold was performed using SSPACE (Boetzer, M., et al.) and
2 PLATANUS using the Nanopore and Matepair data. Final draft genome resulted after gap
3 closure by GAPCLOSER and PLATANUS gap_close tool using Illumina data. The genome
4 size was estimated using a k-mer distribution plot using JELLYFISH (Marcais, G., et al.). The
5 repetitive elements were identified in the final assembled draft genome using Repeat Masker
6 tool. The draft genome was hardmasked by using reference genomic repeats of *G. gallus*. The
7 assembly and annotation workflow overview has been represented as Figure 2.
8
9
10
11
12
13
14
15
16
17
18

19 **Simple sequence repeats prediction**

20 Final assembled scaffolds were analysed for Simple Sequence Repeats (SSR) identification.
21
22 SSRs like the di, tri, tetra, penta and hexa-nucleotide repeats in the genome were obtained using
23
24 MISA (Version 1.0.0).
25
26
27
28
29
30

31 **Genome prediction and annotation**

32 Gene models was predicted on a hard masked draft genome, where the repetitive elements in
33
34 the draft genome were masked using genomic repeats of *G. gallus* with Repeatmasker tool and
35
36 further genes were predicted using AUGUSTUS with *G. gallus* as a reference model. The
37
38 predicted proteins were annotated by using BLASTP (Altshul, S., et al.) against all Avian
39
40 sequences downloaded from UniProt Protein Database.
41
42
43
44
45
46
47
48

49 **Pathway Analysis of the draft genome**

50 The predicted proteins were searched against the KEGG-KAAS server (Moriya, Y., et al.) for
51
52 pathway analysis. *G. gallus* (chicken), *Meleagris gallopavo* (turkey), *Taeniopygia guttata*
53
54 (zebra finch), *Falco peregrinus* (peregrine falcon) were used as reference organism for
55
56
57
58
59
60
61
62
63
64
65

237 pathway identification. The EuKaryotic Orthologous Groups (KOGs) were predicted using
238 homology based approach.

239

240 **Mitochondrial genome assembly and annotation**

241 The generated scaffolds from the draft assembly were aligned against the *P. cristatus*
242 mitochondria genome and the mapped reads were filtered and stitched using ABACUS
243 software using the same reference (Zhou, TC., et al.). Further gap closure were performed with
244 3-7kb MP reads to generate an complete assembled mitochondrial genome. MITOS (Bernt, M.,
245 et al.) was used for gene annotation. Circular plot generated using GenomeVx
246 (<http://wolfe.ucd.ie/GenomeVx/>) representing the localization of the gene in the assembled
247 mitochondrial genome.

248

249 **Phylogenetic tree construction**

250 The assembled Peacock mitochondrial genome was searched against 695 avian mitochondrial
251 genomes downloaded from NCBI. Based on the Blast-N homology results (with query
252 coverage > 100, subject coverage > 95, % identity >85 and with 1% gaps allowed in the
253 sequences). 51 mitochondrial genome sequences along with our assembled mitochondrial
254 genome were filtered. Multiple sequence alignment with default parameters were performed
255 using MUSCLE global sequence aligner. Phylogenetic trees were constructed using IQ-TREE
256 version 1.5.6 (www.iqtree.org). The parameters used for phylogenetic tree construction were
257 ultrafast bootstrap (UFBoot, using the `-bb` option of 1000 replicates), and a standard
258 substitution model (`-m MFP`) was given for tree generation. The generated trees from IQ-TREE
259 tool were visualized using Figtree (<http://tree.bio.ed.ac.uk/software/figtree/>) and the Brach-
260 support values were recorded from the output “.treefile”. The trees were modified for better
261 visualization under Trees section increasing order nodes were applied.

262

1
2 **263 Protein domain analysis**

3
4
5 264 Predicted proteins from Peacock, Chicken and Turkey with sequence length greater than 100
6
7 265 amino acids were considered for protein domain analysis. All the protein sequences from each
8
9
10 266 organism were searched against Pfam-A database using Pfam scan for protein domain
11
12 267 identification.

13
14
15 268

16
17 **269 Avian protein families**

18
19 270 The protein sequences of 48 avian genomes was downloaded from the link
20
21
22 271 <http://avian.genomics.cn/en/jsp/database.shtml> apart from the predicted proteins of the draft
23
24 272 genome. Sequences greater than 100 amino acids from all the avian genomes were filtered and
25
26
27 273 concatenated to a single fasta file. These sequences were clustered using CD-HIT (Fu, L., et
28
29 274 al.) with 90% alignment coverage for the shorter sequence with a length difference cutoff of
30
31
32 275 0.9. The single copy ortholog gene family present across all organisms and genes unique to
33
34 276 peacock were filtered and annotated.

35
36 277

37
38
39 **278 Genome conservation analysis**

40
41 279 The assembled draft genome was aligned against the *G. gallus* genome using Chromosomer
42
43
44 280 tool. Draft chromosomes were constructed based on alignment between fragments (scaffolds)
45
46 281 using a reference genome. These reordered assembled genome was aligned against the Chicken
47
48
49 282 genome using LAST aligner with NEAR (finding short-and-strong (near-identical)
50
51 283 similarities.) parameter allowing for substitution and gap frequencies leading to the
52
53
54 284 identification of orthologs. These query mapped regions were filtered with greater than 1% of
55
56 285 the maximum length for visualization using Circos.

57
58 286

287

288 **Results**

289 **DNA Sequencing data**

290 Five libraries of 150 bp paired-end from Illumina HiSeq technology were generated. The short-
291 insert reads of 489,114,747 accounted to genome coverage of 146.7X and long-insert reads of
292 302,884,819 sequences was about 90.9X coverage with a total coverage of 236X. Sequencing
293 of three mate-pairs of 3-5Kb, 5-7Kb of and 7-10Kb yielded 72,915,033, 47,440,144 and
294 36,464,628 reads respectively with an approximate coverage of 21.9, 14.2 and 10.9
295 respectively, with a total of 156 million reads of 47X coverage. Oxford Nanopore technology
296 was used to generate 366,323 long reads having of 2,398,560,283 bp with coverage of 2.3X.
297 The complete sequencing was generated to a depth of ~287X from Illumina and Nanopore
298 platforms. The coverage was based on assuming the peacock genome size of about 1 Gb (Table
299 S1).

301 **Genome assembly**

302 The assembly was performed on Illumina reads with Abyss *de novo* assembler that resulted in
303 ~932 Mb (mega base) of genome with an N50 of 1639 bp. The extension of the contigs were
304 performed with Nanopore reads which generated scaffolds with N50 of 14,748 bp. Super
305 scaffolding of the assembled scaffold was performed using SSPACE and PLATANUS with
306 MP libraries that generated ~916 Mb genome with the N50 value of 168,140bp. The final gap
307 closer was executed using GAPCLOSER program with MP and PE-LI libraries which
308 generated a draft genome of 1.02 GB (giga base). The draft genome assembly of *Pavo cristatus*
309 consists of 179,346bp scaffolds, with a N50 of 189,886bp with 37 scaffolds having sequence
310 length ≥ 1 Mbp. Contigs above 5000 bp have covered a genome of ~0.915 Mb with N50 0.23
311 Mb. In the assembled genome there were ~0.4% of non ATGC characters (Table 1).

312

1
2
3
4 313 Complete mitochondrial genome of 16699 bp was obtained. Total of 22 tRNA, tow rRNA and
5
6 314 13 protein coding genes were identified in the assembled genome (Fig. 3a). A 100% similarity
7
8 315 was observed with the preiously published Peacocks mitochondrial genome (Fig. S1a and S1b).
9

10
11
12 316

13
14 317 **Transposable Elements (TE)**
15

16
17 318 In the bird genome a total of 75,315,566 bp (7.3% of the genome) was predicted to have 5.5%
18
19 319 of Retroelements (with SINEs 0.08% and LINEs 4.71%), 6.25 % total interspersed repeats with
20
21 320 0.84 % simple repeats and 0.21% low complexity regions. The DNA transposons identified in
22
23 321 the genome was 0.71% (Table S2).
24
25

26 322

27
28
29 323 **Protein coding gene annotation**
30

31 324 A total of 23,153 proteins were predicted in the assembled draft genome using AUGUSTUS.
32
33 325 Among them 95% predicted genes were annotated against the other Aves proteins. The 21,854
34
35 326 annotated proteins showed top similarity to species *Gallus gallus* (Chicken) with 11,398
36
37 327 proteins, *Meleagris gallopavo* (Common turkey) with 4059 proteins, *Amazona aestiva* (Blue-
38
39 328 fronted Amazon parrot) with 1352 proteins and *Anas platyrhynchos* (Mallard) (*Anas boschas*)
40
41 329 with 849 proteins. Thirteen species had about 100 to 400 annotated proteins. The remaining
42
43 330 proteins were in the range of 1 to <100 proteins in about 62 species. From the annotations a
44
45 331 total of 13,161 proteins showed similarity to uncharacterized protein annotation and some of
46
47 332 the over represented proteins were Tyrosine-protein kinase, Sulfotransferase, Phosphoinositide
48
49 333 phospholipase C, Tetraspanin, Phospholipid-transporting ATPase, Olfactory receptor,
50
51 334 Polypeptide N-acetylgalactosaminyltransferase, Transporter, Keratin, Hexosyltransferase,
52
53 335 Protein Wnt, Kinesin-like protein, Gap junction protein, Claudin, POU domain protein,
54
55
56
57
58
59
60
61
62
63
64
65

336 Sodium/hydrogen exchanger, Phospholipid-transporting ATPase, Histone-lysine N-
337 methyltransferase and others (Table S3). The gene ontology annotations showed to have Gene
338 Ontology (GO) descriptions for 18,295 proteins. Among them, 14,490 proteins have Molecular
339 Function; 11,679 have Biological Process and 13,736 proteins have Cellular Component as
340 functional categories.

341 About 17.7% of proteins were found to have pathway information against the KEGG database
342 (Table S4). Some of the overrepresented annotations were Kinases like MAPK (mitogen-
343 activated protein kinase); JNK (c-Jun N-terminal kinase); RAF (RAF proto-oncogene
344 serine/threonine-protein kinase); AKT (RAC serine/threonine-protein kinase); protein kinases
345 and different GTPases.

346 Proteins searched against the KOG annotations showed a total of 20,937 proteins having
347 annotations. Among them, the most abundant annotations include Zn-finger, transmembrane
348 receptor, ubiquitin ligase, Leucine rich repeat, Cadherin repeats, Serine/threonine protein
349 kinase, Collagens, Ankryin repeat, Fibrillins, Voltage-gated Ca²⁺ channels and Hormone
350 receptors (Table S5). The peacock proteins when searched against the human proteins showed
351 expansions in ontologies for cell morphogenesis, neuronal projection and development and
352 GTPases (Table S10 and Fig. S4).

353

354

355 **Simple sequence repeats**

356 A total of 399,493 SSRs were identified from the peacock genome assembly. The largest
357 fraction of SSRs identified were mononucleotide (60.04%), followed by tetra nucleotide
358 (26%), di nucleotide (8.51%), tri nucleotide (4.31%), penta nucleotide (1.03%) and finally hexa
359 nucleotide (0.13%). Among the SSRs identified, A (49.2%) and T (44.9%) accounted for
360 94.1% of the mono-nucleotide repeats. AT (23.8%), TA (16.5%), TG(13.7%), AC(10.6%) and

361 CA (10.32%) accounted for 75% of the di-nucleotide repeats. while TTG (9.9%), AAT (9.6%),
362 AAC (9.4%), TTA (7.1%), ATT (4.5%), TAA (3.5%), CAA (3.1%) and GGA (2.69%)
363 accounted for 49.7% of the tri-nucleotide repeats (Table S6).

364

365 **Avian protein family analysis**

366 A total of 748,544 protein sequences from 49 avian species have 653,497 protein sequences of
367 length above 100 amino acids (Table S7A). A total of 240,853 gene clusters were generated of
368 which 41 gene clusters had single copy orthologs in all avian species (Table S7B and Table
369 S7C). With the above stringent cutoff we observed 15,913 gene clusters were unique to
370 peacock species.

371

372 **Phylogeny and Genome comparisons**

373 The phylogeny of 51 mitochondrial genome sequences along with peacock genome showed a
374 clade consisting of *Pavo* species and *Gallus* (red junglefowl, Sri Lankan junglefowl, grey
375 junglefowl, grey junglefowl), *Bambusicola* (Mountain Bamboo-partridge, Chinese bamboo
376 partridge) and *Francolinus* (Chinese francolin). It can be observed many species were
377 distributed in two different clades where 34 species were found in one clade and five species
378 in other clade. Some of the endemic or native bird species like *Arborophila ardens*, *Acryllium*
379 *vulturinum* and *Numida meleagris* were found as clear outgroup of species in this study (Fig.
380 3b).

381 Predicted proteins from Peacock, Chicken and Turkey were searched for protein domain
382 analysis. 81% of the Pfams were common among the three species. About 94%, 98.4% and
383 99.7% predicted Pfam domains were identified in Peacock, Chicken and Turkey respectively.
384 There were 255, 69 and 14 Pfam domains found to be unique in the species mentioned above
385 respectively (Fig. 4).

1 386 The assembled Peacock genome was reordered for pseudo chromosomes generation against
2 387 the simple repeat masked Chicken genome (1.21 GB, Warren, WC., et al.) using Chromosomer
3
4 388 which generated a overall reordered Peacock genome of about 597MB. The right side of the
5
6
7 389 image represents the reference genome and left side of the image represents the Peacock
8
9
10 390 genome (Fig. 5).

11 391

14 392 **Conclusions**

16
17 393 Third generation sequencing in avian genomics where long reads having the substantiality to
18
19 394 improve genome assembly will benefit understanding the organisms in the structurally
20
21
22 395 complex regions having repeat elements and isoforms in the genome (Goodwin, S., et al.).
23
24 396 Using a combination of short reads of different insert sizes as well as mate pair reads generated
25
26
27 397 from Illumina technology along with long reads from Oxford Nanopore, we obtained an
28
29 398 improved assembly and a draft genome of the Indian Blue Peacock (*Pavo cristatus*). In
30
31
32 399 comparison to other avian genomes (Zhang, G., et al.), the current 290X sequencing depth
33
34 400 obtained from our study is one of the highest. With a N50 of 0.23Mb, we presented here a
35
36 401 reasonably reliable draft genome for the peacock species. The inclusion of Nanopore reads
37
38
39 402 366,323 for scaffolding followed by subsequent gap-closing using Illumina data led to a 26.2%
40
41 403 reduction in the number of scaffolds and a 50.65% and 115% increase in the scaffold and contig
42
43
44 404 N50, respectively. With only 2.3X of long reads, a significant improvement in the assembly
45
46 405 was observed. On the contrary, the assembly contained less than 0.4% of unknown nucleotides,
47
48
49 406 which is very low for a draft assembly. With the addition of more long reads along with
50
51 407 transcriptomic sequencing along with scaffolding and/or gap closure tools, further
52
53 408 improvement in the assembly can be achieved.

54 409

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

410 Peacock seems to defy the Darwinian laws of natural selection. These concern were raised by
411 no other than Darwin himself. Hence, he proposed the theory of the sexual selection where the
412 female can choose for a male with a certain phenotypic feature such as brilliant color or a long
413 tail (Burgess, S.). Peacock's brilliantly colored long tail feathers seems to evolve at the cost
414 of finding its female partner thereby contributing its beneficial genes, even at the cost of
415 making itself venerable to predators. A female peafowl in turn tends to choose the mate with
416 the largest and decorated plumage, which indirectly reflects its healthiness and capacity to
417 wade off potential competitors thereby selecting the most suitable male. Peacocks beautiful
418 feathers with it all its artistry surely provides it with an advantage to impress the females
419 (Dakin, R., et al.). Understanding the formation of beautiful feathers from the genomic context
420 will help in resolving several evolutionary theories on sexual selection that have been discussed
421 on this species.

422 Pigment particles are embedded into the newly grown peacock feathers during the molting
423 season which seems to absorb light of selective wavelength there by imparting to the color of
424 the plumage (Mercedes Foster, Rennee Riedler, et al.) Pigment morphogenesis in *Gallus* has
425 been reported by a process of melanoblasts migration and colonization into feather bud where
426 they differentiate to produce the pigment, melanin (Kelsh, RN., et al.). The molecular
427 mechanisms that control the pigment cell migration have been narrowed down to proteins Kit1
428 and FGFs which maintain the melanoblasts migration to feathers. Kit and FGF proteins have
429 also been identified in our current study (Table S3). Understanding of these proteins in the
430 patterned formation will help decode the pigment pattern morphology in peacocks. These
431 pigmented patterns play a role in communication, choice of mate and in some species it can
432 help in camouflage (Burgess, S)

433 It has been observed that the variations in the genome size among bird species are very low
434 (Table S8). The genome complexities of a species are influenced by the Transposable elements
435 (TE) that are believed to play a crucial role (Kapsuta, A., et al.). The long read sequences have
436 significantly helped in resolving the TEs in genome quality and assembly. Peacock genome
437 comparisons with Turkey and Chicken have showed closeness to the Chicken species. The

1 438 mitochondrial phylogeny also revealed similar findings. Homology searches have shown
2
3 439 several important gene family expansions such as Kinases, Zn finger proteins, GTPases and
4
5 440 others. Their roles in biology, development and evolution of the Peacocks have to be further
6
7 441 explored.

8
9 442 To summarize, we have assembled the *Pavo* genome using Illumina and Oxford nanopore
10
11 443 technology. The genome information can be valued and explored by avian enthusiasts to further
12
13 444 understand about this bird. Though not critically endangered yet, in India, peafowl population
14
15 445 is surely at a decline in the wild due to massive deforestation and habitat loss. Thus is further
16
17 446 compounded by increased poaching for meat and feathers. Our genome sequencing initiative
18
19 447 of *Pavo cristatus* is not just only from a conservational viewpoint, but also to preserve a
20
21 448 heritage associated with this bird that runs through centuries and that bears a strong attachment
22
23
24 449 to the national psyche.
25
26
27
28

29 450
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
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

451 **Availability of supporting data**

452 Supplementary data contains, read statistics, annotation, repeats identification, orthology
453 analysis, mitochondria assembly and annotation. Figures, Gene ontology. DNA and library
454 preparation protocols.

455

456 **Raw Data in SRA**

457 Raw reads (Illumina and Nanopore) are available in the Sequence Read Archive (SRA), and
458 the Whole Genome Shotgun project has been deposited at GenBank under SRA Submission
459 ID: SUB3108024, Bioproject: PRJNA413288 and Biosamples SAMN07739105 :
460 SKPea2016_SI, SAMN07739104 : SKPea2016_LI, SAMN07739101 : FPL_3_5KB,
461 SAMN07739102 : FPL_5_7KB, SAMN07739103 : FPL_7_10KB and SAMN07739107 :
462 FPL_Nano.

463

464 **Competing interests**

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

466

467 **Authors contributions**

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

472

473

474

475

476 **Acknowledgements**

1
2 477 Department of Biochemistry, AIIMS, New Delhi for providing space and infrastructure to carry
3
4
5 478 on the work. RD for providing partial funding. Kritika Tripathi for inputs with manuscript
6
7 479 editing. Raja and Sudha from Genotypic Technology Pvt. Ltd for providing sequencing
8
9
10 480 services and also Deepti, Shalini, Nagu, Abhi and Manish. BR acknowledges IUT 34-4.

11
12 481

13
14
15 482
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

483 **Tables**

484 Table 1. *De novo* assembly statistics of the Peacock genome.

Description	Contigs	Nanopore Scaffold	Super Scaffolds	GapClosed	>1000 Kb	>5000 Kb
Contigs	685241	281272	179346	179346	34181	15026
Maximum Length	49159	251510	2390121	2488982	2488982	2488982
Minimum Length	300	5	265	265	1000	5000
Average Length	1360	3250	5111	5729	-	-
Total Length	932162464	914363908	916720956	1027551907	954483822	915373606
Length >= 100 bp	685241	281271	179346	179346	34181	15026
Length >= 200 bp	685241	281271	179346	179346	34181	15026
Length >= 500 bp	616120	186433	93727	93727	34181	15026
Length >= 1 Kbp	363428	104479	34168	34181	34181	15026
Length >= 10 Kbp	1591	24748	9249	10311	10311	10311
Length >= 1 Mbp	0	0	27	37	37	37
Non-ATGC #	350325	42696911	49169831	4034372	4032567	3978757
Non-ATGC %	0.038	4.67	5.364	0.393	0.422	0.435
N50 value	1639	14748	168140	189886	218023	232312

485

486

487 **Figure legend**

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

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

490 **Figure 3a.** Circular representation of Peacock mitochondrial genome with genes predicted.

491 **Figure 3b.** Phylogenetic tree generated from mitochondrial genome from 52 different avian
species.

493 **Figure 4.** Circular image of the assembled peacock genome aligned against the *G. gallus*
genome using Chromosomer tool. Draft chromosomes were generated by similarity between
scaffolds which were arranged on the reference chicken genome. Circos was used for
visualization.

497 **Figure 5.** Venn diagram showing common and unique Protein family domains (Pfam) between
Peacock, Chicken and Turkey proteins.

499

500 **References:**

1
2 501 Gadagkar, R., 2003. Is the peacock merely beautiful or also honest?. Current Science, 85(7),
3
4 502 pp.1012-1020.

5 503
6
7 504 Kushwaha, S., and Kumar, A. 2016. A Review on Indian Peafowl (*Pavo cristatus*) Linnaeus,
8
9 505 1758. Journal of Wildlife and Research, 4, 42-59.

10 506
11
12 507 Kadgaonkar, Shivendra B. 1993. The peacock in ancient Indian art and literature. Bulletin of
13
14 508 the Deccan College Research Institute, vol. 53, pp. 95–115. JSTOR,
15
16 509 www.jstor.org/stable/42936434.

17 510
18
19 511 Brickle, N. 2002. Habitat use, predicted distribution and conservation of green peafowl (*Pavo*
20
21 512 *muticus*) in Dak Lak Province, Vietnam. Biological Conservation, 105: 189-197.

22 513
23
24 514 Jackson, C. 2006. Peacock. London: Reaktion Books Ltd.

25 515
26
27 516 Zhang, G., Jarvis, E. D., and Gilbert, M. T. P. 2014. A flock of genomes. Science 346, 1308–
28
29 517 1309.

30 518
31
32 519 International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative
33
34 520 analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature,
35
36 521 432(7018), 695.

37 522
38
39 523 Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M.,
40
41 524 Harada, M., Nagayasu, E., Maruyama, H. and Kohara, Y., 2014. Efficient de novo assembly
42
43 525 of highly heterozygous genomes from whole-genome shotgun short reads. Genome research,
44
45 526 24(8), pp.1384-1395.

46 527
47
48 528 Loman, N. J. and Quinlan, A. R. 2014. Poretools: a toolkit for analyzing nanopore sequence
49
50 529 data. Bioinformatics, 30(23), 3399-3401.

51 530
52
53 531 Li H. and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
54
55 532 Transform. Bioinformatics, 25:1754-60.

56 533

534 Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao,
1 535 Y., Hirst, M., Schein, J.E. and Horsman, D.E., 2009. De novo transcriptome assembly with
2 ABySS. *Bioinformatics*, 25(21), pp.2872-2877.
3
4 536
5 537
6
7 538 Boetzer, Marten, and Walter Pirovano. 2014. SSPACE-LongRead: scaffolding bacterial draft
8 genomes using long read sequence information. *BMC bioinformatics* 15.1: 211
9 539
10
11 540
12 541 Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W., 2010. Scaffolding pre-
13 assembled contigs using SSPACE. *Bioinformatics*, 27(4), pp.578-579.
14 542
15
16 543
17
18 544 Marcais, G and Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting
19 of occurrences of k-mers. *Bioinformatics* 27(6): 764-770.
20 545
21
22 546
23 547 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local
24 alignment search tool. *Journal of molecular biology*, 215(3), pp.403-410.
25 548
26
27 549
28
29 550 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. and Kanehisa, M., 2007. KAAS: an
30 automatic genome annotation and pathway reconstruction server. *Nucleic acids research*,
31 551 35(suppl_2), pp.W182-W185.
32 552
33
34 553
35
36 554 Zhou, T.C., Sha, T., Irwin, D.M. and Zhang, Y.P., 2015. Complete mitochondrial genome of
37 the Indian peafowl (*Pavo cristatus*), with phylogenetic analysis in phasianidae. *Mitochondrial*
38 555 *DNA*, 26(6), pp.912-913.
39
40 556
41
42 557
43 558 Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsche, G., Pütz, J.,
44 Middendorf, M. and Stadler, P.F., 2013. MITOS: improved de novo metazoan mitochondrial
45 559 genome annotation. *Molecular phylogenetics and evolution*, 69(2), pp.313-319.
46 560
47
48 561
49
50 562 Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W., 2012. CD-HIT: accelerated for clustering the next-
51 563 generation sequencing data. *Bioinformatics*, 28(23), pp.3150-3152.
52
53
54 564
55
56 565 Warren, W.C., Hillier, L.W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., Markovic,
57 C., Bouk, N., Pruitt, K.D., Thibaud-Nissen, F. and Schneider, V., 2016. A new chicken genome
58 566
59
60
61
62
63
64
65

567 assembly provides insight into avian genome structure. G3: Genes, Genomes, Genetics, pp.g3-
1 568 116.
2
3 569
4
5 570 Goodwin, S., McPherson, J.D. and McCombie, W.R., 2016. Coming of age: ten years of next-
6 generation sequencing technologies. Nature Reviews Genetics, 17(6), p.333.
7 571
8
9 572
10 573 Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., Storz, J.F., Antunes, A., Greenwold,
11 M.J., Meredith, R.W. and Ödeen, A., 2014. Comparative genomics reveals insights into avian
12 574 genome evolution and adaptation. Science, 346(6215), pp.1311-1320.
13 575
14
15 576
16
17 577 Burgess, S., 2001. The beauty of the peacock tail and the problems with the theory of sexual
18 578 selection. Journal of Creation, 15(2), pp.94-102..
19
20
21 579
22
23 580 Dakin, R. 2008. The role of the visual train ornament in the courtship of peafowl, *Pavo*
24 581 *cristatus*. Masters Abstracts International, 47/03: 97.
25
26 582
27
28 583 Mercedes S Foster 1975. The overlap of molting and breeding in some birds. The Condor
29 584 77:304-314.
30
31
32 585
33
34 586 Renee Riedler, Christel Pesme, James Druzik, Molly Gleeson, Ellen Pearlstein The Journal of
35 587 the American Institute of Conservation 2014
36
37
38 588
39
40 589
41
42 590 Kelsh, R.N., Harris, M.L., Colanesi, S. and Erickson, C.A., 2009. Stripes and belly-spots—a
43 591 review of pigment cell morphogenesis in vertebrates. Seminars in cell & developmental
44 592 biology. Vol. 20, No. 1, pp. 90-104.
45
46
47 593
48
49 594 Kapusta, A. and Suh, A., 2017. Evolution of bird genomes—a transposon's- eye view. Annals
50 595 of the New York Academy of Sciences, 1389(1), pp.164-185.
51
52
53 596
54
55 597
56
57 598 **Webservers:**
58 599 FastQC : <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
59
60 600 AUGUSTUS : <http://augustus.gobics.de/>
61
62
63
64
65

601 PLATANUS : <http://platanus.bio.titech.ac.jp/>
1
2 602 ABACUS : <http://abacas.sourceforge.net/Manual.html>
3
4 603 MISA: <http://pgrc.ipk-gatersleben.de/misa/>
5
6 604 REPEATMASKER : <http://www.repeatmasker.org/>
7
8 605 KOGs : <https://genome.jgi.doe.gov/portal/help/kogbrowser.jsf>
9
10 606 KAAS : <http://www.genome.jp/tools/kaas/>
11
12 607 MITOS : <http://mitos.bioinf.uni-leipzig.de/>
13
14 608 MUSCLE : <https://www.drive5.com/muscle/>
15
16 609 PFAMSCAN : <https://www.ebi.ac.uk/seqdb/confluence/display/THD/PfamScan>
17
18 610 LAST : <http://last.cbrc.jp/>
19
20 611 CHROMOSOMER : <https://github.com/gtamazian/chromosomer>
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



Figure 2

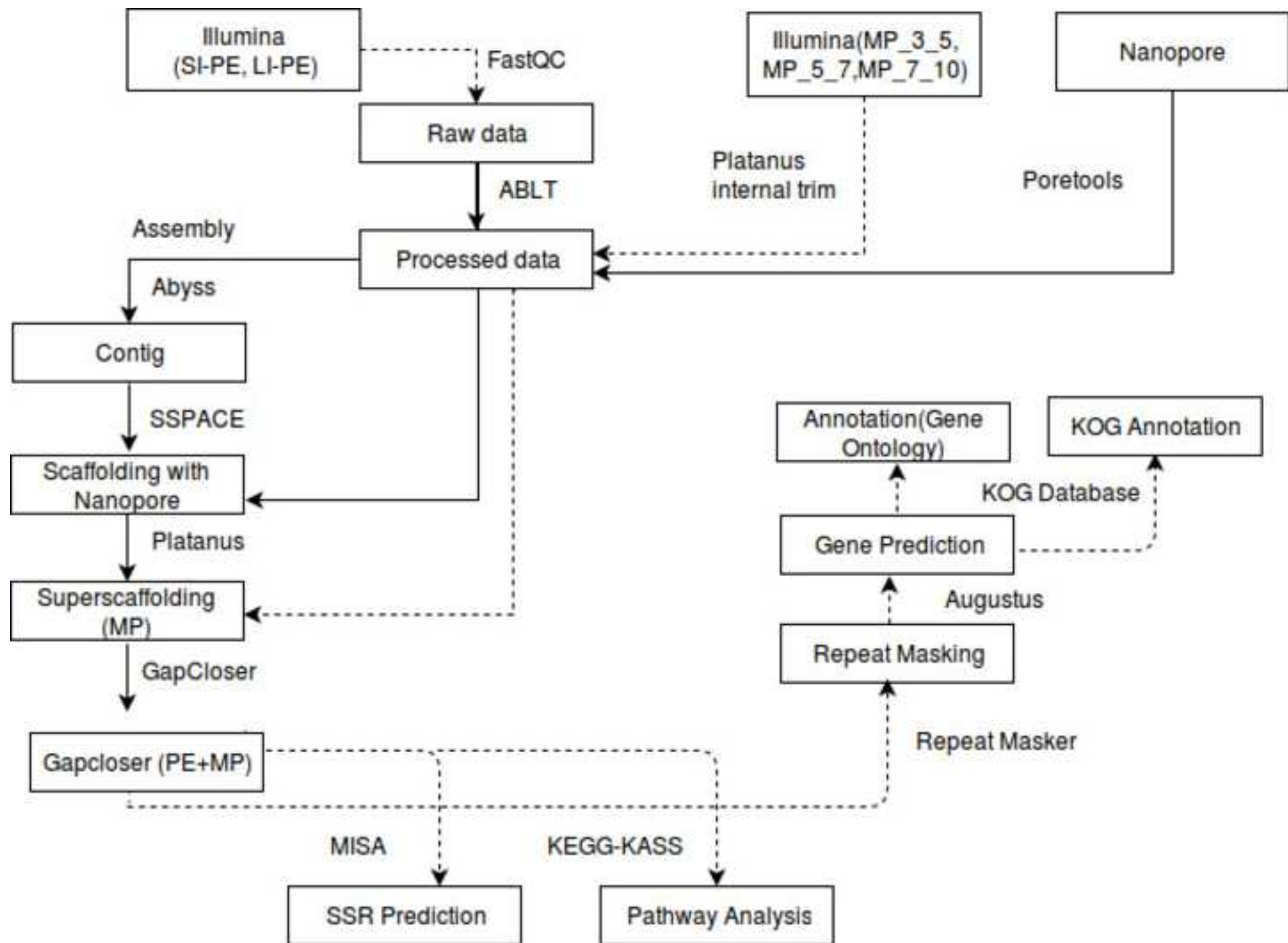


Figure 3a

[Click here to download Figure Fig 3a. Mitochondria genome with genes.pdf](#)

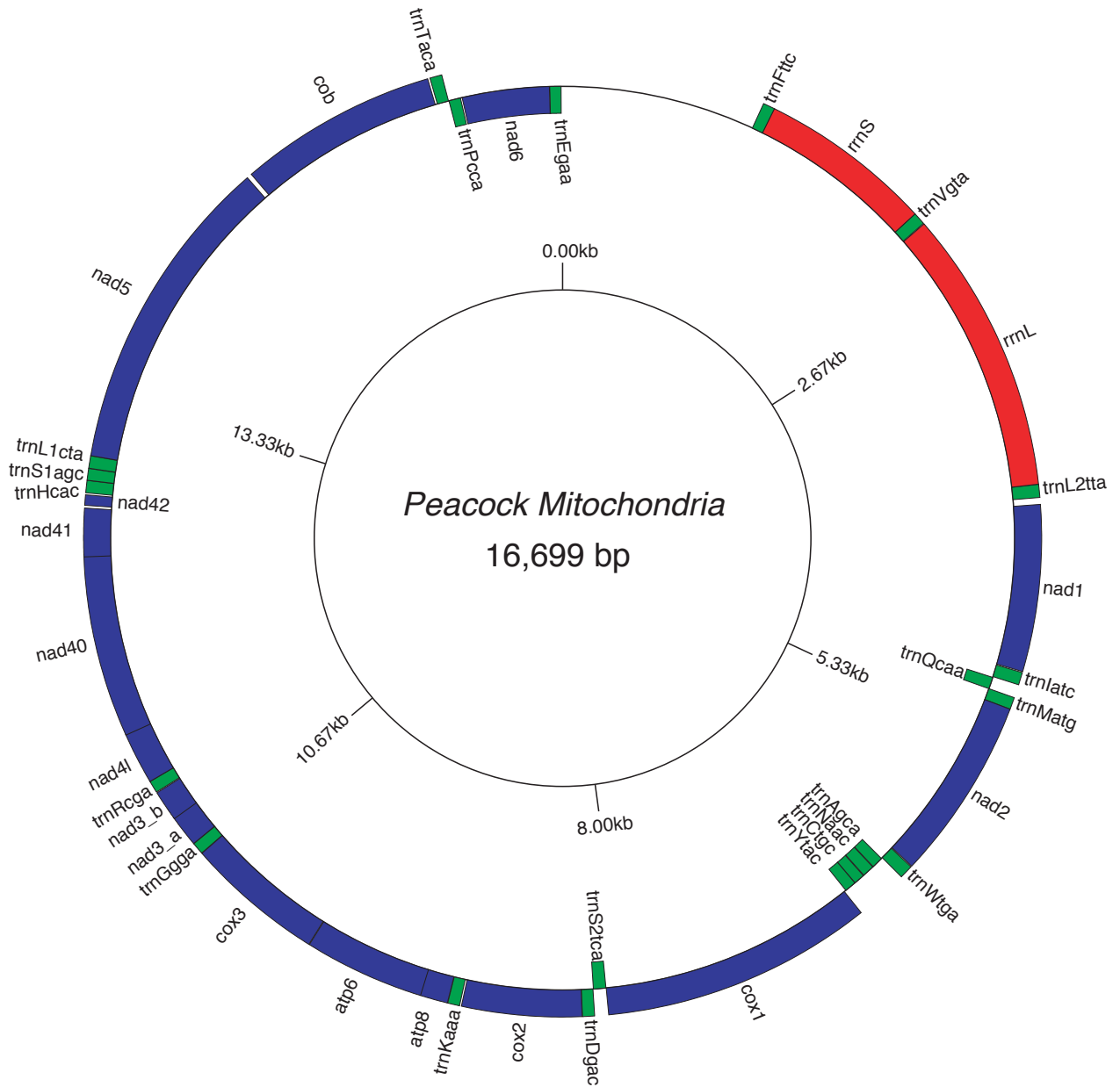
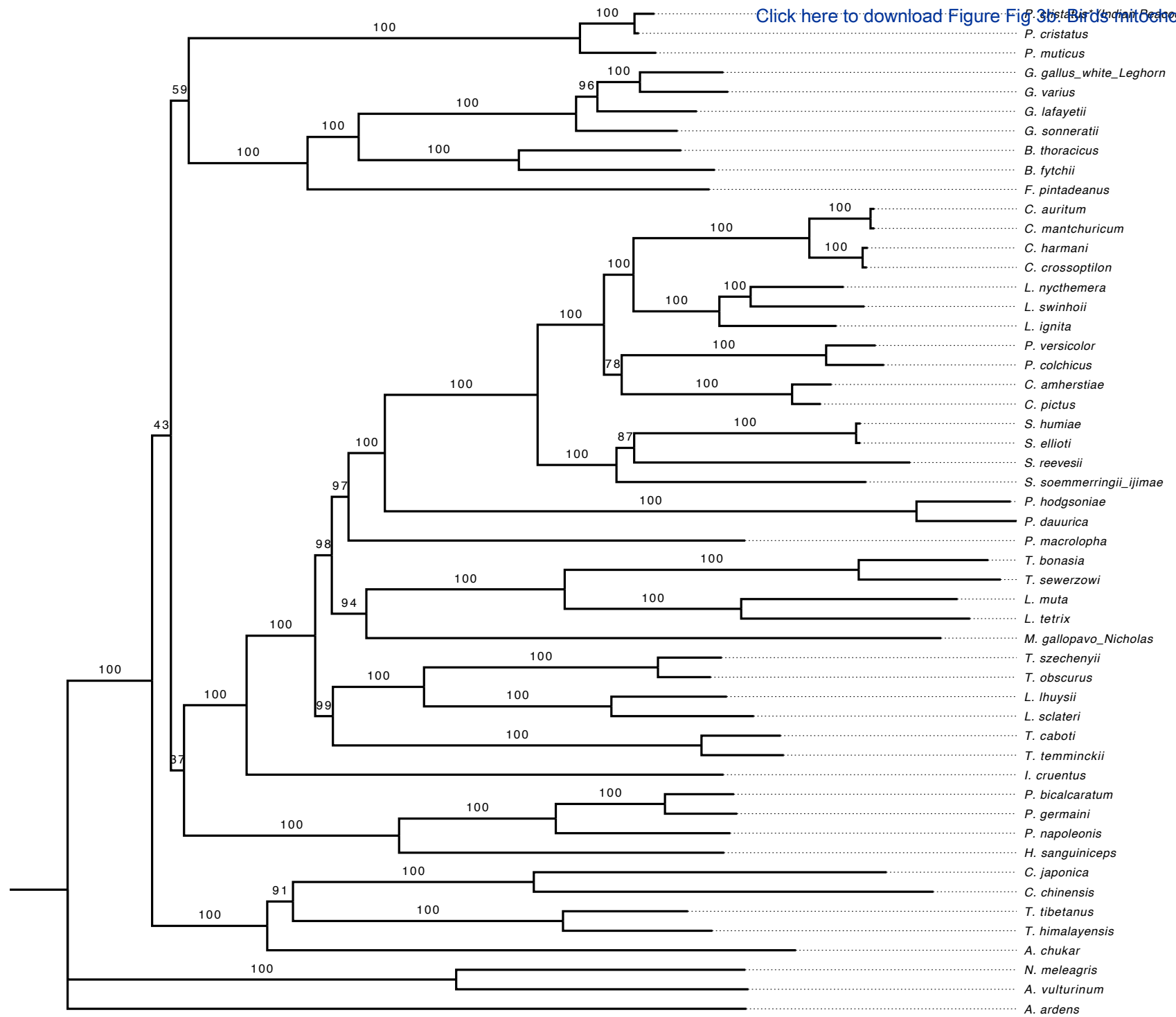
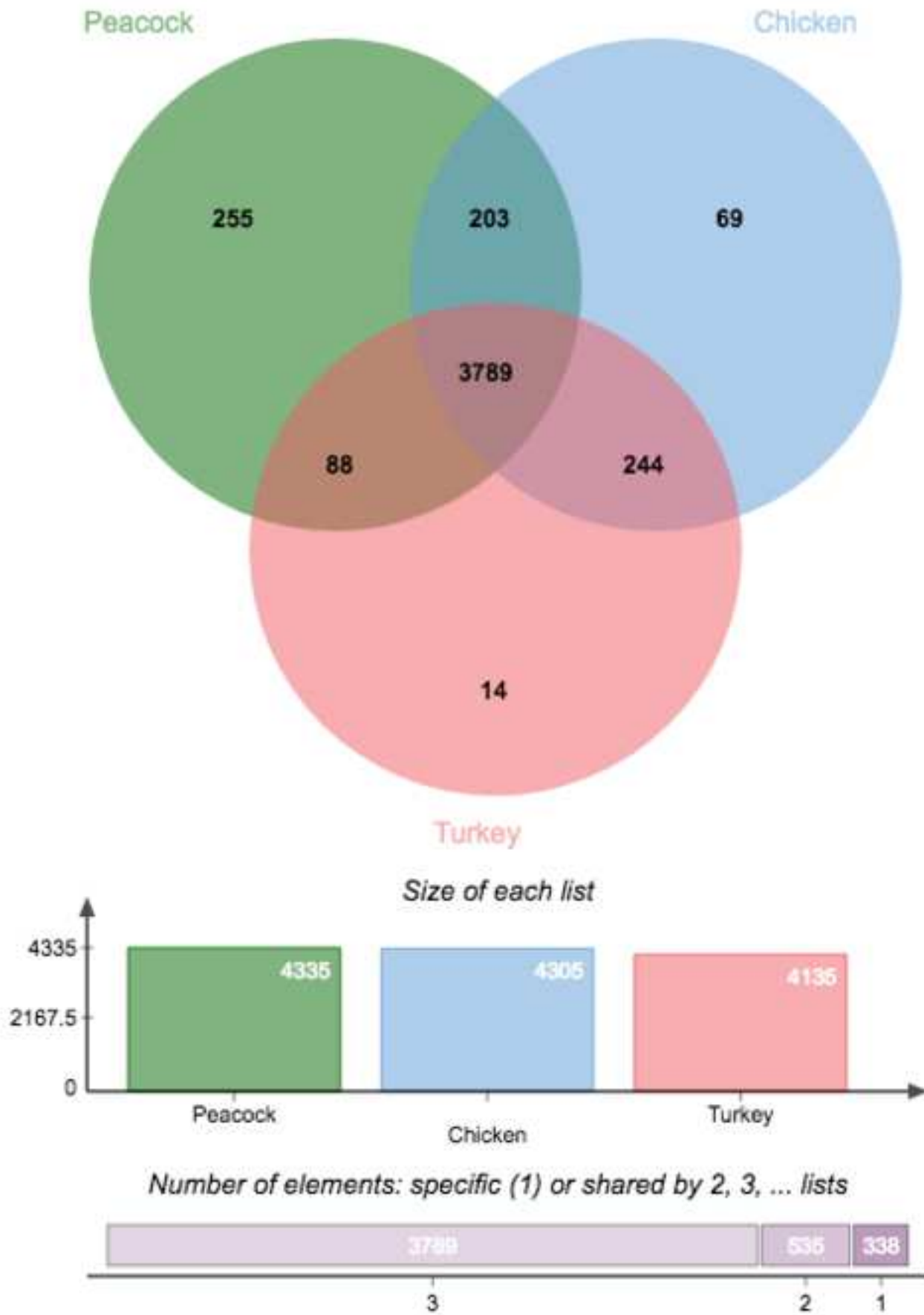
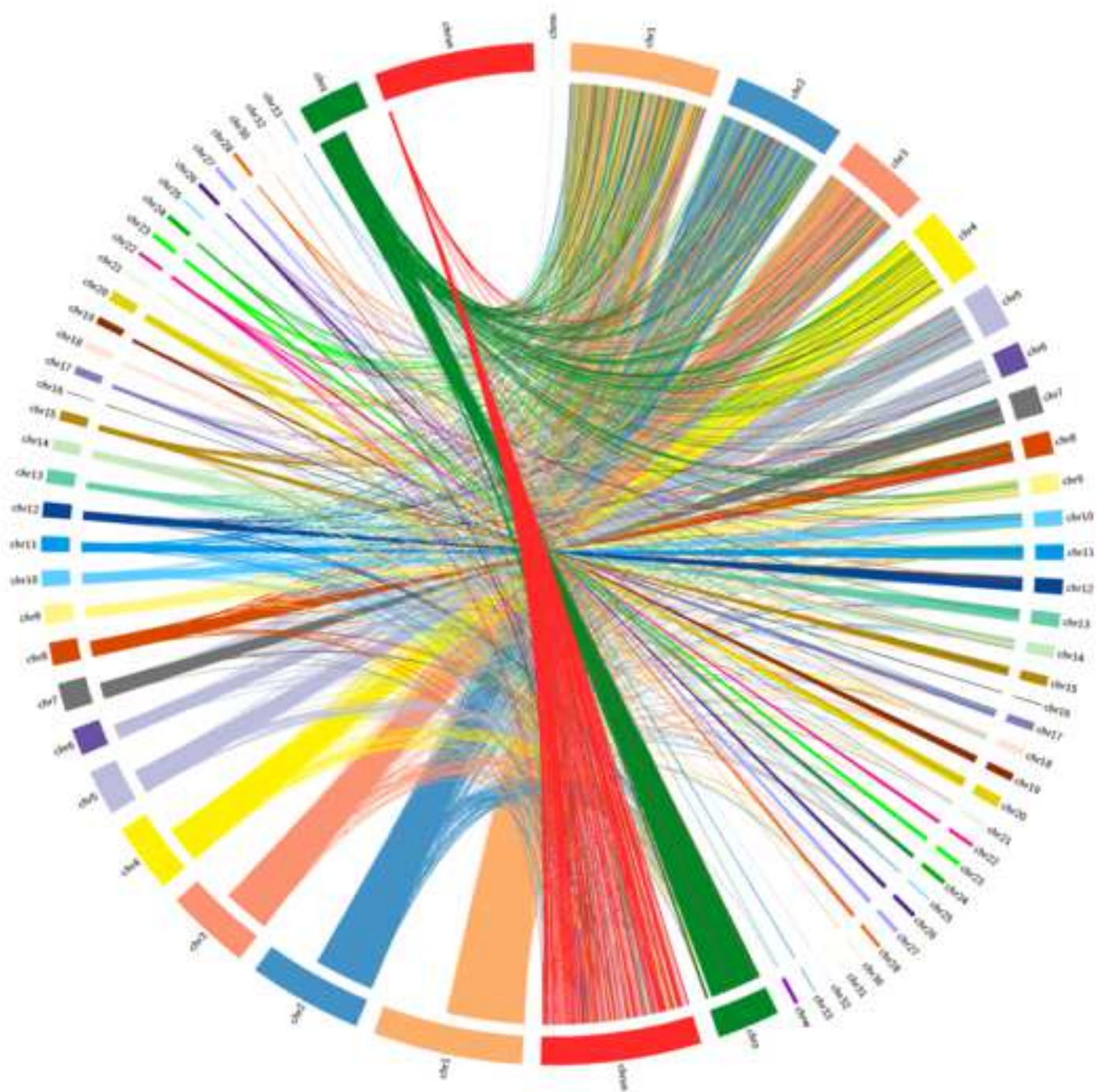


Figure 3b



0.03







[Click here to access/download](#)

Supplementary Material

[Supplementary_Notes_Mitochondria.docx](#)





[Click here to access/download](#)

Supplementary Material

[Description of all the tables and figures.docx](#)



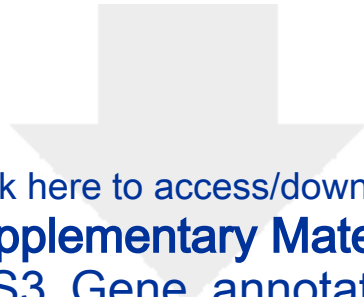


[Click here to access/download](#)

Supplementary Material

[Table_S1_ReadStats_Table_S2_TEs.xlsx](#)





Click here to access/download
Supplementary Material
Table_S3_Gene_annotation.xlsx



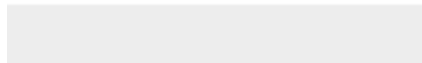


Click here to access/download
Supplementary Material
Table_S4_KEGG_annotation.xlsx





Click here to access/download
Supplementary Material
Table_S5_KOG_annotation.xlsx

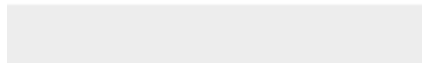




[Click here to access/download](#)

Supplementary Material

[Table_S6_BlastVsHumanProteins.xlsx](#)





Click here to access/download
Supplementary Material
Table_S7_SSR.xlsx







[Click here to access/download](#)

Supplementary Material

[Table_S9_Species_for_mitochondriaxlsx.xlsx](#)





Click here to access/download
Supplementary Material
Table_S10_Pfam_Analysis.xlsx

