Research Article

High-Quality Genome Assembly of the Endemic, Threatened White-Bellied Sholakili Sholicola albiventris (Muscicapidae: Blanford, 1868) From the Shola Sky Islands, India

Vinay K L[1], Chiti Arvind[2], Naman Goyal[2], Robin V. Vijayan[2]

1. Department of Biological Sciences, Louisiana State University, United States; 2. Department of Biology, Indian Institute of Science Education and Research, Tirupati, Tirumala - Tirupati, India

The White-bellied Sholakili (*Sholicola albiventris*) is an endemic, elevational restricted species occurring in the Shola Sky Islands of the Western Ghats of India. This unique understory bird, with a complex vocal repertoire, exhibits impacts of gene flow due to anthropogenic habitat fragmentation. Here, we present the first genome assembly for *Sholicola albiventris*, which was assembled using a combination of Nanopore and Illumina sequences. The final assembly is 1.083 Gbp, consisting of 975 scaffolds with an N50 of 68.64Mbp and L50 of 6. Our genome assembly's completeness is supported by a high number of BUSCOs (99.9%) and a total of 4887 ultraconserved element (UCE) loci retrieved. We also report the complete mitochondrial genome comprising 13 protein-coding genes, 22 tRNAs, and 2 rRNAs. We identified 11.82% of the nuclear genome as repetitive and 36,000 putative genes, with 12017 genes functionally annotated. Our assembly showed a great synteny between *Taeniopygia guttata* and *Gallus gallus* chromosome level assemblies. This reference will be pivotal for investigating landscape connectivity, sub-population genetics, local adaptation, and conservation genetics of this high-elevation, range-restricted endemic bird species.

Corresponding authors: K L Vinay, klvinay.777@gmail.com; V.V. Robin, robin@iisertirupati.ac.in

Introduction

The White-bellied Sholakili (*Sholicola albiventris*) is a range-restricted, vulnerable passerine endemic to the high-elevation montane forests of the Shola Sky Islands of the Western Ghats of India.[1][2] A phylogenetic study by Robin et al., 2017[1] revealed that the White-bellied Sholakili and one of its two congeneric sister species, the Nilgiri Sholakili (*Sholicola major*), are separated by an ancient biogeographic barrier within the Western Ghats. Formerly classified under the genera *Myiomela, Brachypteryx*, and *Callene*[3][4][5], these species were determined to have a closer phylogenetic relationship within the taxa from the Himalayas and Southeast Asia, necessitating the establishment of a new genus. This slaty blue, monomorphic bird is often difficult to spot in the dense shola understory that it inhabits but is frequently heard due to its characteristic loud song. Notably, the White-bellied Sholakili possesses a highly complex birdsong, adding to its unique behavioral and ecological traits[6].

The habitat of the White-bellied Sholakili originally comprised a bi-phasic mosaic of shola grasslands and shola forests. These shola ecosystems, characterized by patches of stunted tropical montane forests interspersed with open grasslands, create a unique landscape that supports a variety of endemic flora and fauna[7]. However, anthropogenic activities have significantly altered this landscape with increased invasive timber and agricultural lands over the past few decades[8]. A study using microsatellite markers on the White-bellied Sholakili indicates a recent genetic differentiation in terms of shared alleles (DPS), which may have resulted from anthropogenic fragmentation[9]. This genetic differentiation suggests that the population of the White-bellied Sholakili is isolated, resulting in reduced gene flow and loss of genetic diversity. Such genetic consequences can have long-term effects on the viability of the species, making conservation efforts even more difficult[10].

A high-quality reference genome for the White-bellied Sholakili will facilitate the conservation efforts and provide researchers with valuable resources for assessing population structure at a finer scale. This will enhance our understanding of how landscape modification affects species distribution and help uncover the regions of the genome involved in the complex song production. Here, we describe bShoAlb1.1, a de novo assembly constructed from a wild-caught White-bellied Sholakili. Using a hybrid assembly strategy with Nanopore Long Read technology and Illumina Short Read sequences, we have assembled the first published reference genome for the *Sholicola* genus.

Methods

Sample collection and DNA extraction

A female *S. albiventris* was captured using a mist-netting protocol[11]. Blood was drawn from the ulnar vein and stored in the Queen's lysis buffer. DNA extraction was performed using the Qiagen Blood and Tissue kit (Qiagen, Hilden, Germany) with slight modifications according to the manufacturer's protocol. The concentration of the extracted DNA was measured using a Qubit 4 (Thermo Fisher Scientific Inc., USA) fluorometer, and its integrity was assessed on a 1% agarose gel. DNA was then sequenced on an R10 flowcell on the PromethION, targeting approximately 80x coverage for Oxford Nanopore long reads (ONT) and approximately 30x for 150 bp paired-end short reads on Illumina NovaSeq 6000.

Read pre-processing

Oxford Nanopore reads underwent quality assessment using NanoPlot (Supplementary Figure S1), followed by any reminiscent adapter removal using Porechop v0.2.4[12][13]. Adapter removed reads were subjected to quality trimming with Chopper v0.7.0, employing a quality threshold of Q>7[12]. K-mer counts were then estimated using Meryl v1.4.1[14] with k=21, and GenomeScope2 was utilized for visualizing the generated histogram to determine genome size and heterozygosity[15]. Illumina short reads quality was checked using Fastp v0.20.1[16]. Adapters and low-quality bases (Q<20) were trimmed using Trimmomatic v0.39[17].

Nuclear Genome Assembly

Trimmed and quality-filtered long reads were used to de-novo assemble the genome using Flye v2.9.3-b1797 (--nano-hq, --asm_coverage 40, --genome_size 1.16g)[18]. The 'draft' assembly was then subjected to contamination screening using the Foreign Contamination Screen (FCS-adapter and FCS-gx) suite[19] and found to have none. The draft assembly was then polished using both short reads and long reads. A total of five rounds of polishing was carried out. First, we used Medaka v1.11.3 (https://github.com/nanoporetech/medaka) to correct the assembly from Flye, which was then polished with one round of Racon v1.5.0[20]. Three rounds of polishing with POLCA v4.1.10[21] were carried out using short reads to obtain the 'polished' assembly. We removed redundant haplotypes

from the assembly using purge_haplotigs v1.1.3[22] with coverage cutoffs of -l 5, -m 50, and -h 160. (Supplementary Figure S2). We further improved the assembly using ntLink with ntlink_rounds (w=250)[23][24]. Using the de-novo assembled mitogenome (see below), we removed any contigs associated with the mitogenome from the nuclear assembly. Further, reference-based pseudochromosome scaffolding was performed using RagTag v2.0.1.[25] RagTag clusters, order, and orient assembly contigs based on a Minimap2 alignment of those contigs to a reference genome. We used the *Taeniopygia guttata* reference genome (GCA_003957565.4) to obtain the pseudochromosomes with default settings under the "scaffold" module within RagTag. We used gfastats v1.3.6[26] and compleasm v0.2.2[27] with the aves_odb10 dataset to evaluate the assembly quality and completeness after each step. We also assessed genome completeness by estimating the number of UCEs that could be retrieved from the genome. According to the online tutorial, we extracted the UCEs with Phyluce v1.7.3[28] with 1000 bp flanking regions on both sides. Final scaffolds were renamed before uploading to NCBI.

Mitochondrial genome assembly

The findMitoReference.py script within the MitoHiFi suite[29] was employed to identify the closest available mitogenome to our species. Utilizing the Snowy-browed Flycatcher (*Ficedula hyperythra)* mitogenome (NC_058320.1) identified by findMitoReference.py as a reference, we assembled the mitochondrial genome from trimmed ONT reads using MitoHiFi v3.2.1 with default settings and annotated the assembly using MitoAnnotator v3.98[30][31][32] (Supplementary Figure S3).

Genome Synteny analysis

To evaluate the validity of our reference-guided scaffolding approach, we assessed genome synteny by aligning the Ragtag-scaffolded assembly with both the *Taeniopygia guttata* (GCA_003957565.4*)* and *Gallus gallus* (GCA_016699485.1) reference genomes. *Gallus gallus* was chosen as a reference due to its frequent use as a model organism for comparative studies. Utilizing the 'nucmer' module within MUMMER v4.0.0rc1, alignments were conducted and subsequently filtered using MUMMER's delta_filter module, permitting many-to-many alignments with a minimum identity threshold of 70%[33]. The resultant tab-delimited file of alignment coordinates was generated using the show_coords module and subsequently employed in OmicCircos package v1.40.0[34] in R v4.3.1[35] for Circos plot visualization of synteny.

Repeat masking and Genome annotation

RepeatModeler v2.0.2[36], within Dfam Transposable Element Tools (TETools)[37] container v1.88[38] was employed (with - LTRStruct) to create a species-specific library of transposable elements and repetitive sequences of *Sholicola albiventris*. This species-specific library was merged with existing repeat libraries sourced from Dfam 0th and 3rd partitions[39] and RepBase Repeat Masker libraries (v20181026)[40][41]. The resulting combined library was then used in identifying and soft masking (-xsmall) the repeat elements using the RepeatMasker v4.1.2-p1 (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015 http://www.repeatmasker.org). We utilized BRAKER v3.03 to predict the gene models without the RNAseq data. BRAKER 3 predictions are based on successive training using GeneMark-EP+ and AUGUSTUS with extrinsic information of homologous protein sequences. To predict the locations of the genes, we employed the ProtHint pipeline within the BRAKER and trained with AUGUSTUS using vertebrate amino acid sequences from Vertebrata_OrthoDB_10[42][43][44][45][46][47][48][49][50]. We ran InterProScan-5.66-98.0 (with AntiFam-7.0, CDD-3.20, Coils-2.2.1, FunFam-4.3.0, Gene3D-4.3.0, Hamap-2023_05, MobiDBLite-2.0, NCBIfam-13.0, PANTHER-18.0, Pfam-36.0, PIRSF-3.10, PIRSR-2023_05, PRINTS-42.0, ProSitePatterns-2023_05, ProSiteProfiles-2023_05, SFLD-4, SMART-9.0, SUPERFAMILY-1.75) [51] and eggNOG-mapper v2 (with eggNOG DB v5.0.2) on the protein domains identified by BRAKER[52][53]. Before the functional annotation, we sanitized the gff3 files using gfftk (https://github.com/nextgenusfs/gfftk). We then used funannotate v1.8.15 with the outputs from BRAKER, InterProScan 5, and eggNOG-mapper to annotate the genomes functionally. Statistics on the produced annotation have been generated using AGAT v1.0.0[54].

Results and Discussion

Genome assembly

Long read sequencing yielded 11.095 million reads (89GB) with a read length N50 of 10,371bp, longest read of 428.013Kbp (Supplementary Figure S1), and mean read quality of 15.4, leading to an estimated long-read depth of ~77x. Additionally, short read sequencing yielded 221.97 million reads, and post quality and adapter trimming 221.95 million reads were retained, totaling 28.45GB with an estimated

read depth of ~25x for Illumina reads. GenomeScope2 estimated the genome size to be 1.16GB with a heterozygosity of 0.55%[15] (Figure 1b).

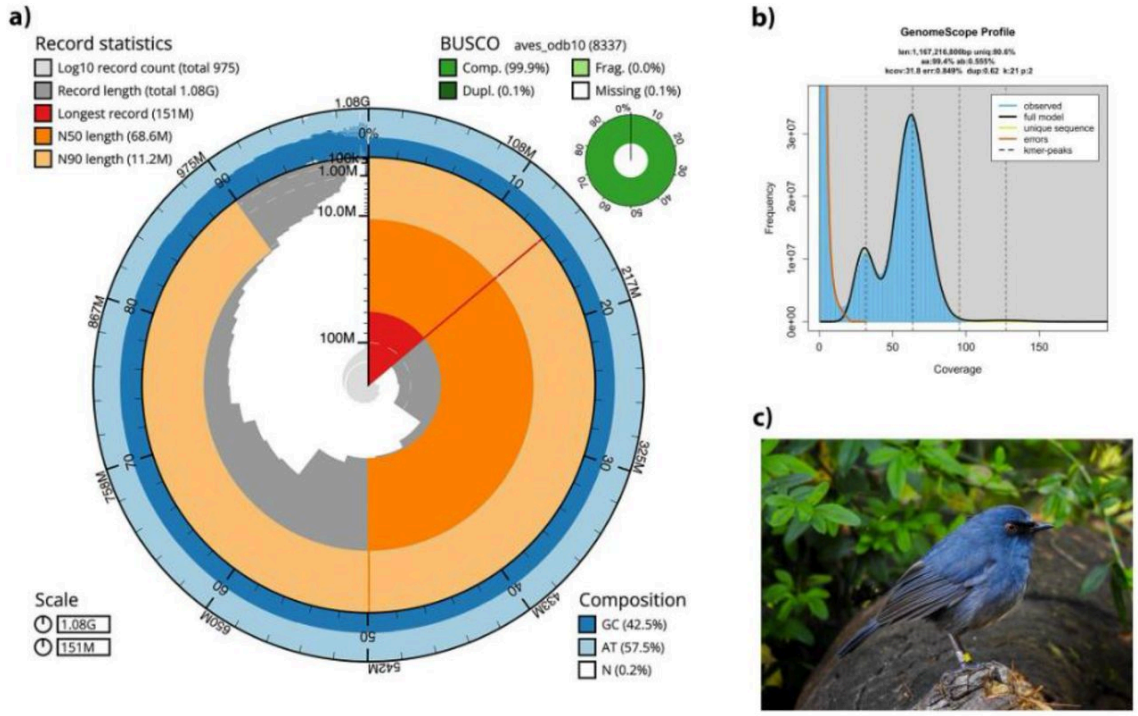

Figure 1. Genome characteristics of assembly bShoAlb1.1. a) BlobToolKit[55] snail plot showing a graphical representation of the quality metrics for the *S. albiventris* assembly (bShoAlb1.1). The circle plot represents the total size of the assembly. From the inside out, the central plot covers length-related metrics. The red line represents the size of the longest scaffold; all other scaffolds are arranged in size order, moving clockwise around the plot. Dark and light orange arcs show the scaffold N50 and scaffold N90 values. The dark versus light blue area around it shows mean, maximum, and minimum GC versus AT content. The BUSCO matrix is obtained from the compleasm. b) GenomeScope 2.0 profile of the k-mer spectra at k = 21 obtained using Meryl. The bimodal pattern observed corresponds to a diploid genome, and the k-mer profile matches that of low (<1%) heterozygosity. c) Photograph of a ringed *Sholicola albiventris* – Yellow - from Kodaikanal, India, picture credits: Vinay K L.

Our initial 'draft' assembly from Flye resulted in 2337 contigs, totaling 1.088 Gbp. After six rounds of long-read and short-read-based polishing, the number of contigs was reduced to 2036. Further, refinement through haplotig purging based on read coverage resulted in an improvement of contiguity and the number of contigs reduced to 1580, accompanied by a reduction in genome size to 1.081 Gbp and the contig N50 of 22.51Mbp. Minimizer graph-based scaffolding and orientation of

contigs produced a total of 1254 scaffolds and 1293 contigs with a scaffold N50 of 33.76Mbp. Notably, fifty percent of the assembly was covered within the ten largest scaffolds (L50). The final assembly, integrated into pseudochromosomes, comprises 975 scaffolds with scaffold N50 of 68.64 Mbp, with the largest scaffold measuring 150.74 Mbp and an L50 of 6 scaffolds. See supplementary table S1 for a comparison of genome contiguity statistics of the recently published bird genomes.

Compleasm results showed that 99.81% (n = 8321) of the avian Benchmarking Universal Single-Copy Orthologs (BUSCO) were present in the final assembly. Detailed genome metrics and BUSCO scores are reported in Table 1. We recovered 4887 UCEs (96.9%, n = 5040), indicating the assembly's high overall recovery and completeness, comparable to loci recovered from other Muscicapidae family genomes[56].

Features	Draft assembly	Polished	Haplotigs purged	Scaffolded	Final (bShoAlb1.1)
# scaffolds	N/A	N/A	N/A	1254	975
Total scaffold length (Gbp)	N/A	N/A	N/A	1.083	1.083
Scaffold N50 (Mbp)	N/A	N/A	N/A	33.76	68.64
Scaffold L50	N/A	N/A	N/A	10	6
Largest scaffold (Mbp)	N/A	N/A	N/A	98.66	150.74
# Gaps in scaffolds	N/A	N/A	N/A	39	318
# contigs	2337	2036	1580	1293	1293
Total contig length (Gbp)	1.088	1.087	1.081	1.080	1.080
Contig N50 (Mbp)	22.52	22.51	22.51	26.82	26.82
Contig L50	13	13	13	11	11
Largest contig (Mbp)	8.27	8.27	8.27	9.75	9.75
	aves_odb10 BUSCOs				
Single	99.69%, 8311	99.70%, 8312	99.80%, 8320	99.81%, 8321	99.81%, 8321
Duplicated	0.22%, 18	0.20%, 17	0.11%, 9	0.11%, 9	0.11%, 9
Fragmented Class 1	0.02%, 2	0.02%, 2	0.02%, 2	0.01%, 1	0.01%, 1
Fragmented Class 2	0.00%, 0	0.00%, 0	0.00%, 0	0.00%, 0	0.00%, 0
Missing	0.07%, 6	0.07%, 6	0.07%, 6	0.07%, 6	0.07%, 6
Total	8337				

Table 1. Quality metrics for the assembly of *Sholicola albiventris* at various stages of the assembly pipeline with BUSCO scores

Mitochondrial assembly

Our final complete circularized mitochondrial assembly obtained from MitoHiFi has a length of 16771bp (Supplementary Figure S3). The mitochondrial genome is similar to many other reported avian genomes with 13 protein-coding genes, 22 tRNAs, and 2 rRNAs with a GC content of 46%[56][57][58][59]. Among the total annotated genes within the mitochondrial genome, 28 were on the heavy chain, and nine were on the light chain with a single non-coding control region (D-loop) of 1199bp length.

Genome synteny

Our synteny analysis generally shows a large synteny between the pseudochromosomes of *Sholicola albiventris* and chromosomes of two other species (Figure 2). We recovered a high degree of one-to-one synteny between *Sholicola albiventris* and *Taeniopygia guttata*; however, it is worth noting that there was an absence of one-to-one synteny with the *Gallus gallus* assembly for the larger chromosomes (Figure 2). This indicates a likelihood of chromosomal splitting, attributed to the distinct phylogenetic relationship between taxa[60] and varying numbers of chromosomes.

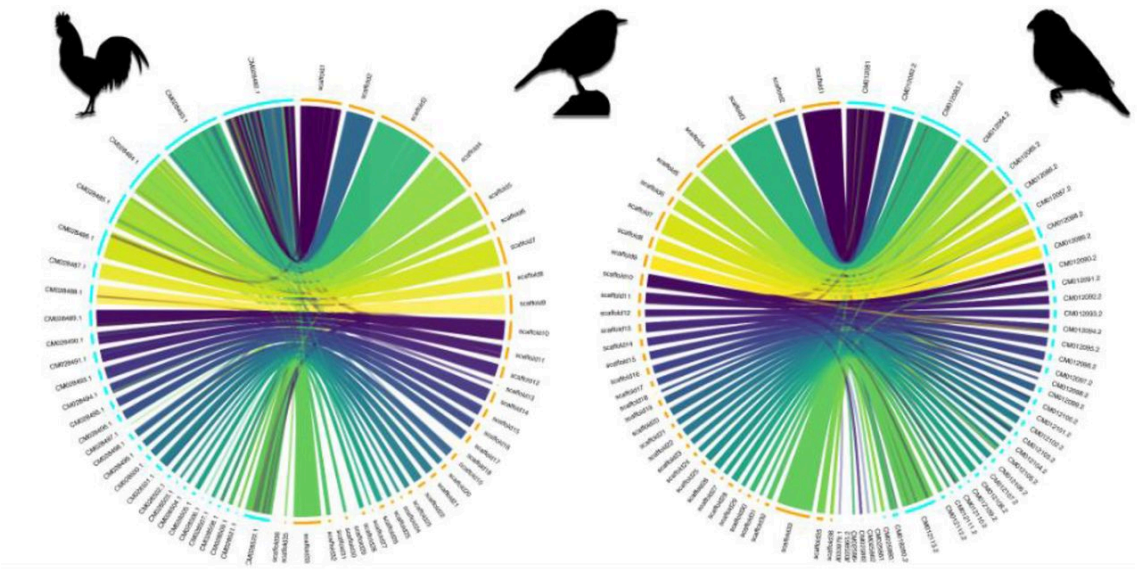

Figure 2. Circos synteny plots plotted using the OmicCircos package showing the comparison between the chromosomes of *Gallus gallus* (left) and *Taeniopygia guttata* (right) with pseudochromosomes of *Sholicola albiventris*. Chromosomes of the compared species are in the cyan-colored hemisphere, and *Sholicola albiventris* is represented in the orange hemisphere. *Gallus gallus and Taeniopygia guttata* illustrations are reproduced from Phylopic (https://www.phylopic.org/).

Genome repeat content and annotation: RepeatMasker identified 11.82 % of the genome as repeat elements, of which 8.64% were interspersed repeats. Retroelements and DNA-transposons comprised 6.96% of repeats, and 1.69% of interspersed repeats were unclassified. Most of the remaining repeat elements were either simple repeats (1.76%) or satellites (0.93%). This is in line with the expected range of transposable elements for Aves[61] and comparable to those reported in other Muscicapidae genomes[56][62]. See Table 2 for a detailed classification of repeats. BRAKER initially found 36815 genes and 39052 mRNAs with a total gene length of 231827394bp. Genes compose 24.4% of the total genome, with a mean gene length of 6297bp. Functional annotation resulted in the names and/or descriptions assigned for 12017 genes and 13572 mRNAs. Additional annotation statistics and annotation files can be found in the Open Science Framework.[63]

	number of elements	length occupied	Percentage of sequences
Retroelements	207427	75147192 bp	6.94%
SINEs:	2173	271604 bp	0.03%
Penelope	0	0 bp	0.00%
LINEs:	132243	35251810 bp	3.25%
CRE/SLACS	0	0 bp	0.00%
L2/CR1/Rex	131762	35093953 bp	3.24%
R1/LOA/Jockey	0	0 bp	0.00%
R2/R4/NeSL	0	0 bp	0.00%
RTE/Bov-B	0	0 bp	0.00%
L1/CIN4	481	157857 bp	0.01%
LTR elements:	73011	39623778 bp	3.66%
BEL/Pao	0	0 bp	0.00%
Ty1/Copia	0	0 bp	0.00%
Gypsy/DIRS1	0	0 bp	0.00%
Retroviral	69498	37393552 bp	3.45%
DNA transposons	1267	166602 bp	0.02%
hobo-Activator	153	18244 bp	0.00%
Tc1-IS630-Pogo	100	15684 bp	0.00%
En-Spm	0	0 bp	0.00%
MuDR-IS905	0	0 bp	0.00%
PiggyBac	0	0 bp	0.00%
Tourist/Harbinger	0	0 bp	0.00%
Other (Mirage, P-element, Transib)	0	0 bp	0.00%
Rolling-circles	2715	1547805 bp	0.14%
Unclassified	50956	18305831 bp	1.69%

	number of elements	length occupied	Percentage of sequences
Total interspersed repeats		93619625 bp	8.64%
Small RNA	440	66018 bp	0.01%
Satellites	4019	10046573 bp	0.93%
Simple repeats	258494	19105522 bp	1.76%
Low complexity	50504	3650726 bp	0.34%

Table 2. Different classes of identified repeats within the bShoAlb1.1 genome.

Here, we present the first de novo assembled highly contiguous genome for the genus *Sholicola*, using a combination of Oxford Nanopore long reads and Illumina short read sequencing technologies. We believe this will serve as an essential resource for the investigations into landscape connectivity, sub-population genetics, local adaptation, and conservation genetics of this high-elevation, range-restricted endemic *Sholicola albiventris* and enhance our understanding of the genetic and evolutionary mechanisms underlying the unique characteristics and contribute towards the deeper understanding of the evolutionary trajectory of the avian genomes and ever-growing repository of avian reference genomes.

Statements and Declarations

Data Availability Statement

This Whole Genome Shotgun project has been deposited at GenBank under the accession JBDGPF000000000. The version described in this paper is version JBDGPF010000000. Raw reads with accession numbers SRR28564530 and SRR28558515 under the BioProject PRJNA1096119 are available from NCBI. Additional supporting data are available from the Open Science Framework.[63] Associated scripts can be found in the GitHub repository (github.com/stachyris/ShoAlb_Ref_Genome).

Conflict of Interest

Authors declare no conflict of interest.

Funding

Science And Engineering Research Board, New Delhi No. CRG/2022/001182

Acknowledgments

We thank the Bird Lab field team at IISER Tirupati for collecting the sample. The sample was collected with permits from the Tamil Nadu Forest Department (permit no WL5(A)/43781/2017). We are grateful to the Scientific Computing Facility at IISER Tirupati and members of the IT Department for HPC access. We thank Brant Faircloth for his valuable input during the project design and for providing computational resources. Portions of this research were conducted with high-performance computational resources provided by Louisiana State University (http://www.hpc.lsu.edu).

References

1. ^a, ^b*Robin VV, Vishnudas CK, Gupta P, Rheindt FE, Hooper DM, Ramakrishnan U, Reddy S (2017). "Two new genera of songbirds represent endemic radiations from the Shola Sky Islands of the Western Ghats, India". BMC Evolutionary Biology. 17 (1): 31. doi:10.1186/s12862-017-0882-6.*
2. ^*BirdLife International (2024) Species factsheet: White-bellied Sholakili Sholicola albiventris. Downloaded from https://datazone.birdlife.org/species/factsheet/white-bellied-sholakili-sholicola-albiventris on 12/05/2024.*
3. ^*Rasmussen P, Anderton JC. "Birds of South Asia: The Ripley Guide. 2nd Edition. 2 vols.". Barcelona: Lynx Edicions; 2012.*
4. ^*Dickinson EC, Christidis L Dickinson EC, Christidis L. The Howard and Moore Complete Checklist of the Birds of the World Fourth Edition, Volume 2: Passerines. Dickinson EC, Christidis L, editors. Eastbourne, UK: Aves Press; 2014.*
5. ^*Rasmussen P. "Biogeographic and conservation implications of revised species limits and distributions of South Asian birds". Zool Med Leiden. 2005;79-3:137–46.*
6. ^*Sawant S, Arvind C, Joshi V, Robin VV (2022). "Spectrogram cross-correlation can be used to measure the complexity of bird vocalizations". Methods in Ecology and Evolution. 13 (2): 459–472. doi:10.1111/2041-210x.13765.*
7. ^*Robin VV, Nandini R (2012). "Shola habitats on sky islands: status of research on montane forests and grasslands in southern India". Current Science. 103 (12): 1427–1437. http://www.jstor.org/stable/24089*

350.

8. ^ *Arasumani M, Khan D, Das A, Lockwood I, Stewart R, Kiran RA, Muthukumar M, Bunyan M, Robin VV (2018). "Not seeing the grass for the trees: Timber plantations and agriculture shrink tropical montane grassland by two-thirds over four decades in the Palani Hills, a Western Ghats Sky Island." PloS One. 13 (1): e0190003. doi:10.1371/journal.pone.0190003.*
9. ^ *Robin VV, Gupta P, Thatte P, Ramakrishnan U (2015). "Islands within islands: two montane palaeo-endemic birds impacted by recent anthropogenic fragmentation". Molecular Ecology. 24 (14): 3572–3584. doi:10.1111/mec.13266.*
10. ^ *Pavlova A, Beheregaray LB, Coleman R, Gilligan D, Harrisson KA, Ingram BA, Kearns J, Lamb AM, Lintermans M, Lyon J, Nguyen TTT, Sasaki M, Tonkin Z, Yen JDL, Sunnucks P (2017). "Severe consequences of habitat fragmentation on genetic diversity of an endangered Australian freshwater fish: A call for assisted gene flow". Evolutionary Applications. 10 (6): 531–550. doi:10.1111/eva.12484.*
11. ^ *Robin VV, Sinha A, Ramakrishnan U (2010). "Ancient geographical gaps and paleoclimate shape the phylogeography of an endemic bird in the sky islands of southern India". PLoS ONE. 5 (10): e13321. doi:10.1371/journal.pone.0013321.*
12. ^ a, b *De Coster W, Rademakers R (2023). "NanoPack2: population-scale evaluation of long-read sequencing data." Bioinformatics. 39 (5). doi:10.1093/bioinformatics/btad311.*
13. ^ *Wick RR, Judd LM, Gorrie CL, Holt KE. "Completing bacterial genome assemblies with multiplex MinION sequencing". Microb Genom. 2017; 3(10): e000132. doi:10.1099/mgen.0.000132.*
14. ^ *Rhie A, Walenz BP, Koren S, Phillippy AM (2020). "Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies". Genome Biology. 21 (1): 245. doi:10.1186/s13059-020-02134-9.*
15. ^ a, b *Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020). "GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes". Nature Communications. 11 (1): 1432. doi:10.1038/s41467-020-14998-3.*
16. ^ *Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu. "fastp: an ultra-fast all-in-one FASTQ preprocessor". Bioinformatics. 34 (17): i884–i890. doi:10.1093/bioinformatics/bty560.*
17. ^ *Bolger AM, Lohse M, Usadel B (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics. 30 (15): 2114–2120. doi:10.1093/bioinformatics/btu170.*
18. ^ *Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019). "Assembly of long, error-prone reads using repeat graphs". Nature Biotechnology. 37 (5): 540–546. doi:10.1038/s41587-019-0072-8.*

19. ^Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, Mozes E, Strope PK, Sylla PM, Wagner L, Bidwell SL, Brown LC, Clark K, Davis EW, Smith-White B, Hlavina W, Pruitt KD, Schneider VA, Murphy TD (2024). "Rapid and sensitive detection of genome contamination at scale with FCS-GX." Genome Biology. 25 (1): 60. doi:10.1186/s13059-024-03198-7.

20. ^Vaser R, Sović I, Nagarajan N, Šikić M (2017). "Fast and accurate de novo genome assembly from long uncorrected reads". Genome Research. 27 (5): 737–746. doi:10.1101/gr.214270.116.

21. ^Zimin AV, Salzberg SL. "The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies". PLOS Computational Biology. 2020; 16(6): e1007981. doi:10.1371/journal.pcbi.1007981.

22. ^Roach MJ, Schmidt SA, Borneman AR (2018). "Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies". BMC Bioinformatics. 19 (1): 460. doi:10.1186/s12859-018-2485-7.

23. ^Coombe L, Li JX, Lo T, Wong J, Nikolic V, Warren RL, Birol I (2021). "LongStitch: high-quality genome assembly correction and scaffolding using long reads." BMC Bioinformatics. 22 (1): 534. doi:10.1186/s12859-021-04451-7.

24. ^Coombe L, Warren RL, Wong J, Nikolic V, Birol I (2023). "ntLink: A Toolkit for De Novo Genome Assembly Scaffolding and Mapping Using Long Reads." Current Protocols. 3 (4): e733. doi:10.1002/cpz1.733.

25. ^Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S (2022). "Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing." Genome Biology. 23 (1): 258. doi:10.1186/s13059-022-02823-7.

26. ^Formenti G, Abueg L, Brajuka A, Brajuka N, Gallardo-Alba C, Giani A, Fedrigo O, Jarvis ED (2022). "Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs." Bioinformatics. 38 (17): 4214–4216. doi:10.1093/bioinformatics/btac460.

27. ^Huang N, Li H (2023). "compleasm: a faster and more accurate reimplementation of BUSCO." Bioinformatics. 39 (10). doi:10.1093/bioinformatics/btad595.

28. ^Faircloth BC (2016). "PHYLUCE is a software package for the analysis of conserved genomic loci." Bioinformatics. 32 (5): 786–788. doi:10.1093/bioinformatics/btv646.

29. ^Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, Darwin Tree of Life Consortium, Formenti G, Abueg L, Torrance J, Myers EW, Durbin R, Blaxter M, McCarthy SA (2023). "MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads". BMC Bioinformatics. 24 (1): 288. doi:10.1186/s12859-023-05385-y.

30. ^Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M, Nishida M (2013). "MitoFish and MitoAnnotator: a mitochondrial genome database of fish

with an accurate and automatic annotation pipeline." Molecular Biology and Evolution. 30 (11): 2531–2540. doi:10.1093/molbev/mst141.

31. ^*Sato Y, Miya M, Fukunaga T, Sado T, Iwasaki W (2018). "MitoFish and MiFish Pipeline: A Mitochondrial Genome Database of Fish with an Analysis Pipeline for Environmental DNA Metabarcoding". Molecular Biology and Evolution. 35 (6): 1553–1555. doi:10.1093/molbev/msy074.*
32. ^*Zhu T, Sato Y, Sado T, Miya M, Iwasaki W. "MitoFish, MitoAnnotator, and MiFish Pipeline: Updates in 10 Years". Molecular Biology and Evolution. 2023; 40(3). doi:10.1093/molbev/msad035.*
33. ^*Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A (2018). "MUMmer4: A fast and versatile genome alignment system". PLoS Computational Biology. 14 (1): e1005944.*
34. ^*Hu Y, Yan C, Hsu C-H, et al. "OmicCircos: A Simple-to-Use R Package for the Circular Visualization of Multidimensional Omics Data." Cancer Informatics. 13. doi:10.4137/CIN.S13495.*
35. ^*R Core Team (2021). "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*
36. ^*Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF (2020). "RepeatModeler2 for automated genomic discovery of transposable element families." Proceedings of the National Academy of Sciences of the United States of America. 117 (17): 9451–9457. doi:10.1073/pnas.1921046117.*
37. ^*TETools: Dfam transposable element tools Docker container [Internet]. Github; [date unknown]. Available from: https://github.com/Dfam-consortium/TETools. Accessed 2024 Apr 11.*
38. ^*Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C (2017). "TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes." Nucleic Acids Research. 45 (4): e17. doi:10.1093/nar/gkw953.*
39. ^*Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF (2021). "The Dfam community resource of transposable element families, sequence models, and genome annotations". Mobile DNA. 12 (1): 2. doi:10.1186/s13100-020-00230-y.*
40. ^*Bao W, Kojima KK, Kohany O (2015). "Repbase Update, a database of repetitive elements in eukaryotic genomes." Mobile DNA. 6: 11. doi:10.1186/s13100-015-0041-9.*
41. ^*Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005). "Repbase Update, a database of eukaryotic repetitive elements". Cytogenetic and Genome Research. 110 (1-4): 462–467. doi:10.1159/000084979.*
42. ^*Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M (2021). "BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database." NAR Genomics a*

nd Bioinformatics. 3 (1): lqaa108. doi:10.1093/nargab/lqaa108.

43. ^ *Brůna T, Lomsadze A, Borodovsky M (2020). "GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins." NAR Genomics and Bioinformatics. 2 (2): lqaa026. doi:10.1093/nargab/lqaa026.*
44. ^ *Buchfink B, Xie C, Huson DH (2015). "Fast and sensitive protein alignment using DIAMOND." Nature Methods. 12 (1): 59–60. doi:10.1038/nmeth.3176.*
45. ^ *Gotoh O (2008). "A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence." Nucleic Acids Research. 36 (8): 2630–2638. doi:10.1093/nar/gkn105.*
46. ^ *Hoff KJ, Lomsadze A, Borodovsky M, Stanke M (2019). "Whole-Genome Annotation with BRAKER." Methods in Molecular Biology. 1962: 65–95. doi:10.1007/978-1-4939-9173-0_5.*
47. ^ *Iwata H, Gotoh O (2012). "Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features." Nucleic Acids Research. 40 (20): e161. doi:10.1093/nar/gks708.*
48. ^ *Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005). "Gene identification in novel eukaryotic genomes by self-training algorithm". Nucleic Acids Research. 33 (20): 6494–6506. doi:10.1093/nar/gki937.*
49. ^ *Stanke M, Schöffmann O, Morgenstern B, Waack S (2006). "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources". BMC Bioinformatics. 7: 62. doi:10.1186/1471-2105-7-62.*
50. ^ *Stanke M, Diekhans M, Baertsch R, Haussler D (2008). "Using native and syntenically mapped cDNA alignments to improve de novo gene finding". Bioinformatics. 24 (5): 637–644. doi:10.1093/bioinformatics/btn013.*
51. ^ *Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S (2014). "InterProScan 5: genome-scale protein function classification." Bioinformatics. 30 (9): 1236–1240. doi:10.1093/bioinformatics/btu031.*
52. ^ *Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021). "eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." Molecular Biology and Evolution. 38 (12): 5825–5829. doi:10.1093/molbev/msab293.*
53. ^ *Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P (2019). "eggNOG 5.0: a hierarchical, functionally and phylog*

enetically annotated orthology resource based on 5090 organisms and 2502 viruses." Nucleic Acids Rese arch. 47 (D1): D309–D314. doi:10.1093/nar/gky1085.

54. ^*Dainat J, Hereñú D, Pucholt P (2020). "AGAT: Another Gff Analysis Toolkit to handle annotations in an y GTF/GFF format." (Version v1.0.0) Zenodo. doi:10.5281/zenodo.3552717.*
55. ^*Challis R, Richards E, Rajan J, Cochrane G, Blaxter M (2020). "BlobToolKit - Interactive Quality Assess ment of Genome Assemblies." G3 (Bethesda, Md.). 10 (4): 1361–1374. doi:10.1534/g3.119.400908.*
56. a, b, c*Baudrin G, Pons J-M, Bed'Hom B, Gil L, Boyer R, Dusabyinema Y, Jiguet F, Fuchs J (2023). "A Refer ence Genome Assembly for the Spotted Flycatcher (Muscicapa striata)." Genome Biology and Evolution. 15 (8). doi:10.1093/gbe/evad140.*
57. ^*Benham PM, Cicero C, Escalona M, Beraut E, Marimuthu MPA, Nguyen O, Nachman MW, Bowie RCK (2023). "A highly contiguous genome assembly for the California quail (Callipepla californica)." The Jo urnal of Heredity. 114 (4): 418–427. doi:10.1093/jhered/esad008.*
58. ^*Lan G, Yu J, Liu J, Zhang Y, Ma R, Zhou Y, Zhu B, Wei W, Liu J, Qi G (2024). "Complete Mitochondrial Ge nome and Phylogenetic Analysis of Tarsiger indicus (Aves: Passeriformes: Muscicapidae)". Genes. 15 (1): 90. doi:10.3390/genes15010090.*
59. ^*Lu CH, Sun CH, Hou SL, Huang YL, Lu CH (2019). "The complete mitochondrial genome of dark-sided flycatcher Muscicapa sibirica (Passeriformes: Muscicapidae)". Mitochondrial DNA Part B. 4 (2): 2675–2 676. doi:10.1080/23802359.2019.1644240.*
60. ^*Stiller J, Feng S, Chowdhury AA, Rivas-González I, Duchêne DA, Fang Q, Deng Y, Kozlov A, Stamatakis A, Claramunt S, Nguyen JMT, Ho SYW, Faircloth BC, Haag J, Houde P, Cracraft J, Balaban M, Mai U, Che n G, ... Zhang G (2024). "Complexity of avian evolution revealed by family-level genomes". Nature. doi: 10.1038/s41586-024-07323-1.*
61. ^*Sotero-Caio CG, Platt RN II, Suh A, Ray DA (2017). "Evolution and Diversity of Transposable Elements i n Vertebrate Genomes". Genome Biology and Evolution. 9 (1): 161–177. doi:10.1093/gbe/evw264.*
62. ^*Peona V, Palacios-Gimenez OM, Lutgen D, Olsen RA, Alaei Kakhki N, Andriopoulos P, Bontzorlos V, Sch weizer M, Suh A, Burri R (2023). "An annotated chromosome-scale reference genome for Eastern black -eared wheatear (Oenanthe melanoleuca)". G3: Genes, Genomes, Genetics. 13 (6). doi:10.1093/g3journa l/jkad088.*
63. a, b*Vinay KL, Arvind C, Goyal N, Robin VV (2024). "A high-quality genome assembly for an endemic vul nerable bird, the White-bellied Sholakili (Muscicapidae: Blanford, 1868)". OSF. doi:10.17605/OSF.IO/M9 5Q7.*

Supplementary data: available at https://doi.org/10.32388/OFCM3I

Declarations

Funding: Science And Engineering Research Board, New Delhi No. CRG/2022/001182

Potential competing interests: No potential competing interests to declare.