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

Kathleen Collier[1,2]

1 Conservation Genetics, Reneco (United Arab Emirates), Abu Dhabi, United Arab Emirates

2 University of Colorado System, Boulder, United States

Potential competing interests: No potential competing interests to declare.

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

K.Collier

Keilercollier@icloud.com

General comments

First, congratulations on the assembly – I'm genuinely happy we're at a point where high-quality draft genomes for range-restricted endemics are now feasible to generate. I will refrain from any commentary on the ecology of the species in favor of methods.

Sample collection and DNA extraction

"and stored in the Queen's lysis buffer"

I am unsure what this is. I trust that it works (otherwise, downstream extraction would've failed), but I've never heard of it, and no information or specifications are linked.

"K-mer counts estimated using Mery v1.4.1[14] with k=21, and GenomeScope2 was utilized for visualizing the generated histogram to determine genome size and heterozygosity"

Were the kmer counts performed with the ONT or the Illumina data? If ONT, how did you account for the relatively high (and nonrandom) error profile?

"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."

Did you calculate these coverage cutoffs before or after removing mtDNA-associated contigs? In my experience, the mtDNA-associated contigs tend to have substantially higher depth than genomic ones and can potentially upset the calculations done by both purge_dups and purge_haplotigs.

“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.”

Scaffolding to the Zebra Finch assembly makes sense to me, but I question the utility of inferences based on the Chicken. Synteny is prevalent but not universal throughout birds, and the Zebra Finch is also a model organism with a chromosomal assembly and much closer status taxonomically.
Both the Collared Flycatcher and the Eastern Black-eared Wheatear also have chromosomal assemblies (GCF_000247815.1, GCF_029582105.1) and are representatives of the Muscicapidae, which could be expected to share closer synteny with the White-bellied Sholakili. It would be interesting to see one of them run as well.

“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.”

It is my understanding that the OrthoDB sets for a given taxon are equivalent to the BUSCO/CompleASM sets for that taxon. Is there a reason why you picked the Vertebrata dataset (3354 genes) rather than the Aves set (8338 genes)?

“We ran InterProScan-5.66-98.0 … and eggNOG-mapper v2 (with eggNOG DB v5.0.2) on the protein domains identified by BRAKER”

That is a very long list of gene predictors and annotators. How did you deal with conflicting output between all of these algorithms? BRAKER3 and TSEBRA integrate information from multiple sources, but, accepting the reality of false positives, they have reasonably strict consensus requirements for doing so.