

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

Ole Kristian Torresen¹

¹ University of Oslo, Norway

Potential competing interests: No potential competing interests to declare.

General comments:

The authors present a genome assembly of an endemic and threatened species. On general principles, it is always good to have more genome assemblies out there for different comparative studies. This particular study seems to be well-planned and does create a genome assembly of good quality, while things could have been done differently and maybe better.

How can you tell if something is "high-quality" or not? Finding almost all BUSCO genes is a pretty good indicator, but there are other measures of quality you have not used. For instance, k-mer occurrence from Merqury (compare the k-mers in your assembly to the k-mers found in the Illumina reads [and maybe also the ONT reads]). When you collapse your assembly like you do (that is, not creating a haplotype-resolved assembly), you do lose quite a bit of the information inherent in a diploid genome (it is hard to be sure to capture everything that actually does exist in a genome in a cell without spending vast amounts of money anyway). In general, creating collapsed assemblies when we have the opportunity to create haplotype-resolved assemblies is not a good thing. With both ONT data (and especially) PacBio, it is straightforward to create haplotype-resolved assemblies, and thereby showing and capturing more of the information found in a diploid species.

In general, I am highly skeptical of anchoring a species' contigs based on the chromosomes of another. There are usually quite a bit of differences between two species, as you also mention when comparing your assembly to the chicken assembly. I am afraid that you have "zebrafinched" your assembly. It would have been better to just keep your assembly as contigs, or ideally, created a Hi-C library from the same individual and scaffolded it yourself.

It looks like a couple of the figures are of poor quality. Figures 1b and 2 are impossible to read, in both the HTML and the PDF versions.

Specific comments:

Please use lowercase 'k' as a thousand denominator instead of capital 'K' as in 428.013Kbp, for instance (should be 428.013kbp).

What is the unit in 28.45GB? Capital B is usually meant to be bytes, as in gigabytes. I think you might mean base pairs here. Then you should use Gbp (or Gb). Same with 1.16GB and maybe other places. You have used Gbp many other places.

Why did you lose contigs when polishing? I did not know that any of the tools usually used in polishing remove any contigs (but I am not specifically familiar with what you have used). What would be the characteristics of contigs removed? Little coverage?