Review of: "A chromosome-scale hybrid genome assembly of the extinct Tasmanian tiger (Thylacinus cynocephalus)"

Zijun Xiong[1]

1 Beijing Genomics Institute

Potential competing interests: The author(s) declared that no potential competing interests exist.

This manuscript by Feigin and his colleagues reported an update genome assembly of the extinct marsupial carnivore (*Thylacinus cynocephalus*) native to Australia. The authors utilized short-reads data for the genome assembly using museum thylacine sample. They then applied reference-guided scaffolding to link the contigs to scaffolds, and obtained long enough genome sequence that represents each number of chromosome. The authors improved the genome assembly by increasing the contig N50 from 3kb to 15kb. The thylacine genome sequences could provide valuable resources for studying on marsupial carnivores, in particular convergent evolution with Canida.

General comments:
The negative consequence of DNA fragmentation from thylacine sample is contamination. If relatively intact exogenous DNA is co-extracted with the fragmented DNA from the thylacine sample, the exogenous DNA may be amplified during PCR amplification, even if it is present at low concentration. To make sure the overall data quality, I would recommend the authors map the short-reads or genome assembly to NT/bacteria database. The contamination of the sample might be a concern.

Minor comments:
Line 70: Do you assemble the Y chromosome? Is the sample female or male?
Line 72: Fig. 1d. 3014 Gb should be 3.014Gb. How many high-quality data do you use in this study? What is the sequencing coverage?
Line 77: I don't think the high gap percentage (~10%) is caused by 205 Mbp remained unplaced scaffolds. Do you have a stats for the gap percentage of each assembled chromosome. I suppose the high percentage was induced by the reference-guided scaffolding. In addition, the assembled sequences are pseudo-chromosomes. This should be described clearly.
Line 80: The 12.7% missing BUSCO suggest that there were still regions that were not assembled. It could also be caused by the erroneous sequences. It would be good if you can obtain per-base quality values by calculating Phred quality score. This will confirm the sequence base was of high accuracy. As thylacine is also a good candidate for de-extinction, I suggest to estimate the actual genome size using Illumina short reads, to confirm what percentage of the genome was still missing.

Line 103: the author used only homology-based annotation to predict gene models. This is not enough. Lineage-specific genes can be missing when only used homology-based methods. To gain a comprehensive gene annotations, you need to add de novo prediction (e.g. Augustus).