Review of: "Putative host-derived insertions in the genome of circulating SARS-CoV-2 variants"

Brianna Chrisman[1]

1 Stanford University

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

The authors seek to answer the ongoing question of whether recent insertions in SARS-CoV-2 evolution could have human origin and resulted from virus→human RdRp template-switching events. I thought the authors provided novel evidence in favor of this hypotheses (direct RNA-seq data showing clear virus-human genome chimeric events + insertions in a handful of recent SARS-CoV-2 sequences that are clearly derived from another organism) though I still have a few lingering questions that I think the authors can address given the data and experiments accessible to them.

1. Via direct RNA-seq, the authors showed solid evidence of chimeric transcripts containing both human and viral RNA. I'm not an expert in direct RNA-seq, but is there definitely nothing in the storage/library prep pipeline that might prevent removal of polyA sites + ligation of two RNAs? Assuming not, then RdRp template-switching seems a reasonable hypothesis for how these chimeric transcripts were formed. However, if human RNA were to make it into a SARS-CoV-2 genome, it seems like RdRp would have to move from SARS-CoV-2 to human back to the same spot on SARS-CoV-2. Did the authors see any double chimeric transcripts where this could have happened? Are there any theories for insertions being derived from strand breakage + repair being discussed?
2. Although the authors mention that the GISAID sequences with human-derived insertions were from different laboratories and therefore not likely the result of artifacts, is there anyway the insertions in the GISAID references have come from mispriming or chimeric reads during sequencing & assembly? Did the authors filter to high coverage sequence or long vs short read sequences? The authors should double check that these insertions don’t just originate from long reads as there are some systematic errors with long read sequences. Though it is highly doubtful a systematic error would produce such a long insertion similar to human sequence. Unless that same systematic error has made its way into the human reference genome assembly too… (which is not impossible!).
3. What are the other 98 organisms the likely human-derived insertions are found? Any chance the insertion is actually bacteria- or virus- derived rather than human-derived?
4. There is a lot of raw data deposited on NCBI's SRA that was used to construct GISAID sequences (and I think a lot of U of Washington where some of the insertions were found?) and many GISAID contributors are willing to share their raw data as well. Did the authors double check there is no way to get access to

the raw reads for the samples with human-derived insertions? I think looking at the heterozygosity of the insertions could be extremely useful!

5. Minor point: the enrichment of lncRNA in chimeric transcript has a significant but pretty weak p-value (.03). Is this uncorrected p-value? Was this the only hypthesis tested? If the authors tested the enrichment of different types of RNAs then this multiple hypothesis testing correction might invalidate this.