# The Curious Case of the RaTG13 Genome: Time to Revise the Sequence Reporting Standards for Pathogens

Mohit Singla[1], Saad Ahmad[2], Chandan Gupta[2], Tavpritesh Sethi[2]

[1]All India Institute of Medical Sciences, New Delhi, India

[2]Indraprastha Institute of Information Technology Delhi, India

**Mohit Singla** was a clinician researcher at All India Institute of Medical Sciences, New Delhi, India.

**Saad Ahmad** is an undergraduate student in the Computer Science and Biosciences program at Indraprastha Institute of Information Technology Delhi, India. His research interest is application of machine learning algorithms to bio medicinal data.

**Chandan Gupta** is an undergraduate student in the Computer Science and Biosciences program at Indraprastha Institute of Information Technology Delhi, India. His research interests are computational genomics and machine learning applications in biomedicine.

**Tavpritesh Sethi** is an Associate Professor in the Department of Computational Biology at Indraprastha Institute of Information Technology Delhi, India. His research interests are at the intersection of machine learning, artificial intelligence, genomics and medicine.

**Corresponding author.** Tavpritesh Sethi, Departmental of Computational Biology, Indraprastha Institute of Information Technology, Delhi. Telephone Number: +91-11-26907533 Email: tavpriteshsethi@iiitd.ac.in

## Abstract

The origin of SARS-CoV-2 is debated, even after 18 months into the COVID-19 pandemic and a special investigation conducted by the World Health Organization. The RaTG13 sequence has been a highlight of discussions surrounding the origin of SARS-CoV-2. Here we express our opinion about the need for better reporting standards for information about sequencing data, especially for pathogens, citing our findings with the reported RaTG13 genome.

Scientists have asked for a deeper investigation into the lab-leak theory[1] for insights into the origin of SARS-CoV-2, with many citing inconsistencies in the original paper[2] that reported the sequence of RaTG13 and SARS-CoV-2[3–6]. We carried out a de-novo assembly from this dataset and found a glaring lack of information, thus precluding reliable inferences of the origin of SARS-CoV-2 from this data. Our de novo assembly of the RaTG13 genome was implemented with Megahit using three different settings. We did not find any contig larger than 17000 nucleotides in the assembly and could not reproduce the complete reported sequence as a single contig. We did, however, obtain several matching segments, the largest of which was less than 20% of the length of the reported sequence. Reads from SRR11085797 yielded 1760 hits to the sequence upon conducting a BLAST search. More than 1% of the region was uncovered and we observed zero coverage for a region spanning positions 13182 to 13293, which raises questions about the veracity of reporting of this region. Further, we see strong evidence for DNA contamination supported by (i) the presence of 98% match with the mitochondrial sequence of Rhinolophus sinicus (KR106992.1) which is surprising because a near-complete assembly of a mitochondrial sequence is unexpected from an RNA sample, (ii) the presence of non-adapter related repetitive motifs, [GGGTTGG(R)AACAGGATA(GGGTTA)n]m and its reverse complement

[(TAACCC)nTATCCTGTT(Y)CCAACCC]m in a majority of reads. The latter was found in approximately 60% of the dataset, usually on the same end, despite no evidence for these in the TruSeq mRNA kit that was reportedly used. Further, since the G-quadruplex GGGTTA is well known to be present in the telomeric region of many genomes, we checked for the presence of microsatellite regions of Rhinolophus affinis[7] in the dataset, but these were not significantly seen. Hence, we were unable to explain the abundant presence of this segment and its reverse complement on the same end of the "RNA" sample, where only one strand should have dominated. We conjecture that the sample has gross contamination issues with RNA and DNA mixed in the same sample and any such possibility needs to be reported clearly in sequencing experiments. Additionally, we request better standards of reporting for data-quality metrics. These should not just be included but clearly elaborated as a routine feature. In the RaTG13 reported genome, we calculated the average coverage as 9.73, which is much lower than what is expected for a sequence of such import. This metric was missing in the report, but more surprisingly, we observed that roughly 3000 bases had coverage less than or equal to 2, and a large proportion of second end reads have Phred score < 30, a standard cut-off used in genomics experiments. While it is expected for both ends in a paired-end sequencing to have similar features, the distribution of sequence lengths and ambiguous base calls (N) on the first end to be significantly different from the second end. Only 150bp length reads were present in the tiles 1104-1128; 149, 151, 18-39 bp reads are absent in this region. Reads from various tiles show specificity for ambiguous base calls at positions 140, 141, 132, 137, 138 and 142. We do not discount post-processing which may have introduced these differences, but the protocols followed should be clearly reported with sequences of such significance.

Further, the SRR11085797 dataset, from which RaTG13 is reported to have been derived, was found to have a very diverse and possibly contaminated biological composition. The NCBI KRONA analysis revealed that (i) SRR11085797 shows a significant difference in percentage composition when compared to another sample submitted by the same group from a similar host (Rhinolophus affinis), body site (anus), platform (ST-J00123) and temporally close in terms of the sequencing run (84 and 89) on the same sequencing machine. Whereas the SRR11085797 sample has only 0.65% bacterial content, which is unexpected from a fecal sample. It has 67.91% eukaryotic content and 29.38% unidentified content. In contrast, the other similar sample submitted by the same group, SRR11085736, has 91.07% bacterial content and only 4.36% eukaryotic content. Further, 0.2% of SRR11085797 hits Manis javanica, which has been identified as a potential intermediate host of SARS-CoV-2[8]. 0.1% of SRR11085797 hits embryophytes, an odd occurrence for a bat sample; the ST-J00123 sequencing platform has been earlier used for sequencing maize as evidenced by BioProjects PRJNA338015 & PRJNA383075. In the absence of relevant information, we conjecture that the sample suffers from contamination by index hopping since the other similar sample, SRR11085736, that is claimed to have been guarded against index hopping by using separate lanes[9], has sequences resembling viruses found in other datasets (NCBI accession MN611520.1 apart from the virus identified from its own accession, i.e., MN611522.1). Thus, the presence of cross-contamination in SRR11085736 with specific guarding against index hopping raises a distinct possibility of samples in earlier runs, i.e., SRR11085797, not being guarded against index hopping. This could explain the diversity of contaminating sequences in SRR11085797 if the experiment was not done in separate lanes of a multiplexed run. Therefore, our observations convey the need for pathogen sequencing data to be mandatorily accompanied by a report delineating the details of sequencing protocols and the biological

influences on a sample in order to prevent the propagation of misinformation and accurate tracing of emerging pathogens.

The RaTG13 datasets have listed down most of the details as required by the MIGS specification[10]; however, our experiments show glaring inconsistencies and a need for more information. To mitigate the risk of misinformation, in addition to being MIGS compliant, **we propose the following checklist of meta-data details to be reported with a pathogen sequence to ensure reproducibility and accurate surveillance of pathogens:**

1. All possible biological influences on the sample during handling or during sequencing.
2. Source, purity, format and amount of sample before the sequencing experiment.
3. Complete details of the sequencing run, platform and the library preparation detail.
4. The spike-in controls used in the experiment. The changes (if any) in the controls should also be reported.
5. Complete Quality Check (QC) details, including the coverage for each target position, percentage of bases that meet the required minimum coverage requirement, percentage of aligned reads, percentage of bases corresponding to the target sequence, percentage of target bases with no coverage, percentage of ambiguous base calls, etc.
6. Contamination checks at several stages in the sequencing workflow.
7. Details on all post-processing protocols.
8. Any other investigations performed on the sample.

**Key Points:**

- Minimum standards for meta-data information about pathogen sequences need to be revised beyond the MIGS specification. This is required for accurate surveillance of pathogens around the globe.
- The case in point of this opinion article is the RaTG13 sequence which has emerged to be a key evidence about the origin of SARS-CoV-2.
- De-novo assembly and quality analysis of RaTG13 highlights glaring inconsistencies and contamination issues which cannot be resolved in the absence of meta-data and optimum reporting standards.

## Contributions

Dr Mohit Singla - Devising the workflow, drawing novel inferences, mentoring and guiding the project.

Saad Ahmad: Worked on De Novo Assembly, coverage calculation and data analysis.

Chandan Gupta: Worked on sequence similarity, data preprocessing and visualizations.

Dr Tavpritesh Sethi - Mentoring and guiding the project and publishing the paper.

## Acknowledgements

This work was partially supported by the Wellcome Trust/DBT India Alliance Fellowship IA/CPHE/14/1/501504 awarded to Tavpritesh Sethi and the Center of Excellence in Healthcare at IIIT-Delhi.

## Competing Interests

Authors declare that they have no competing interests.

## References

1. Bloom JD, Chan YA, Baric RS, et al. Investigate the origins of COVID-19. Science 2021; 372:694
2. Zhou P, Yang X lou, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020; 579:270–273
3. Zhang D. Anomalies in BatCoV/RaTG13 sequencing and provenance. 2020;
4. Deigin Y, Segreto R. SARS-CoV-2′s claimed natural origin is undermined by issues with genome sequences of its relative strains. BioEssays 2021; 43:2100015
5. Singla M, Ahmad S, Gupta C, et al. De-novo assembly of RaTG13 Genome Reveals Inconsistencies further Obscuring SARS-CoV-2 Origins. 2020;
6. Seyran M, Hassan SS, Uversky VN, et al. Urgent need for field surveys of coronaviruses in southeast asia to understand the sars-cov-2 phylogeny and risk assessment for future outbreaks. Biomolecules 2021; 11:1–7
7. Mao X, Liu Y, Zhou Y, et al. Development of 19 polymorphic microsatellite loci for the intermediate horseshoe bat, Rhinolophus affinis (Rhinolophidae, Chiroptera). Conservation Genetics 2009; 10:709–711
8. Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Current Biology 2020; 30:1346-1351.e2
9. Li B, Si H-R, Zhu Y, et al. Discovery of Bat Coronaviruses through Surveillance and Probe Capture-Based Next-Generation Sequencing. mSphere 2020; 5:
10. Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. Nature Biotechnology 2008; 26:541–547