De-novo assembly of RaTG13 Genome Reveals Inconsistencies further Obscuring SARS-CoV-2 Origins

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

Abstract

An intense scientific debate is ongoing as to the origin of SARS-CoV-2. An oft-cited piece of information in this debate is the genome sequence of a bat coronavirus strain referred to as RaTG13 [1] mentioned in a recent Nature paper [2] showing 96.2% genome homology with SARS-CoV-2. This is discussed as a fossil record of a strain whose current existence is unknown. The said strain is conjectured by many to have been part of the ancestral pool from which SARS-CoV-2 may have evolved [7, 8, 9]. Multiple groups have been discussing the features of the genome sequence of the said strain. In this paper, we report that the currently specified level of details are grossly insufficient to draw inferences about the origin of SARS-CoV-2. De-novo assembly, KRONA analysis for metagenomic and re-examining data quality highlights the key issues with the RaTG13 genome and the need for a dispassionate review of this data. This work is a call to action for the scientific community to better collate scientific evidence about the origins of SARS-CoV-2 so that future incidence of such pandemics may be effectively mitigated.

Introduction

After more than seven months of initial reporting, the world still continues to struggle with the COVID-19 pandemic. The complex spectrum of the disease and the unclear origin of the virus have made it an enigma. The origin of SARS-CoV-2 is still not conclusively proven. Several groups are in heated debates on various similarities or dissimilarities between RaTG13[1] and other coronaviruses. In order to discuss the similarity or dissimilarity between 2 or more sequences, one needs to be sure as to how accurate they are. The sequence of RaTG13 and SARS-CoV-2 were reported by the same group[2]. Whereas adequate experimental details have been provided for the human sample[2], this was found lacking for the RaTG13 which was apparently rediscovered after the COVID-19 outbreak by the same group. Recently multiple other publications have come out which discuss concerns[21,22,23,24] about the RaTG13 sequence and the associated dataset. We analyzed the sequence and found that the data quality issues along with the lack of sufficient experimental details preclude reliable inference of the origins of SARS-CoV-2 .

In this work, we carried out a de-novo assembly from this dataset and found glaring issues that may preclude reliable inferences of the origin of SARS-CoV-2 from this data. Since this has been the single most important piece of evidence about the origins of the SARS-CoV-2, our work highlights the need to examine this data closely prior to basing further scientific studies on it.

Results

RaTG13 genome could not be assembled via de novo assembly of the dataset. A de novo assembly of the RaTG13[1] genome was implemented with Megahit[4] using 3 different settings. Under all the three settings, we did not find any contig larger than 17000 nucleotides in the assembly. We could not reproduce the complete reported sequence[1] 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. There were regions of significant gaps in the coverage of the genome. In order to explore the matching segments, a BLAST[5] search for highly similar sequences using megaBLAST[5] was conducted upon the RaTG13 sequence. Reads from SRR11085797 [3] yielded 1760 hits to the sequence. More than 1% 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 by the authors.

Strong evidence for DNA contamination in the sample that may preclude inferences. We found evidence for DNA contamination in the sequence through multiple analyses listed below:-

1. Presence of complete mitochondrial sequence of Rhinolophus Sinicus. The largest contig which when blasted[5] against the standard nucleotide database shows approximately 98% similarity to mitochondria of Rhinolophus sinicus (KR106992.1). A near complete assembly of mitochondrial sequence is unexpected from an RNA sample which would have been interrupted by stop codons preventing full length assembly.
2. **Presence of non-adapter related repetitive sequences in majority of reads.** We observed highly repeated G-quadruplex 'GGGTTA' and its reverse complement 'TAACCC' units. The motifs [GGGTTGG(R)AACAGGATA(GGGTTA)n]m and its reverse complement [(TAACCC)nTATCCTGTT(Y)CCAACCC]m were found to be present in approximately 60% of the dataset, many a times on the same end. These are unlikely to arise from adapter sequences as we could not find evidence of these

adapter sequences in the TruSeq mRNA kit that was reportedly used. Blasting the motifs against some of the assemblies of Rhinolophus genus available on NCBI did not reveal any strong clues as to their origin. We did not observe these motifs to be present in significant numbers in another dataset deposited by the authors, SRR11085736, which had a similar experimental design, platform and machine ID. Further, since the G-quadruplex GGGTTA is well known to be present in the telomeric region in various genomes, we checked for the presence of microsatellite regions of Rhinolophus affinis[12] in the dataset, but these were not abundantly present. 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. Daoyu Zhang et al.[23] has also pointed out the presence of highly repeated sequences throughout the dataset. We conjecture that the sample has gross contamination issues while handling, with RNA and DNA mixed in the same sample. However, in the absence of details about the experiments we do not rule out any steps that we may have missed, in which case the original studies should make these known.

Data Quality issues in the sample

1. **Coverage.** We calculated the average coverage by using a splice-aware aligner[6] to map reads to the RaTG13 sequence and found it to be 9.73. This low coverage may be responsible for assembly of only partial segments of the RaTG13 sequence[1]. Further, we observed that roughly 3000 bases have coverage less than or equal to 2 (Figure 1(a)), thus base calling errors in this region could impact the accuracy of the assembled sequence. A large proportion of second end reads have Phred score in the twenties (Figure 1(b)), while multiple ambiguous bases in the tail of the first end (Figure 2(a)) could have interfered with the de novo assembly.
2. **Non random errors in the sequenced reads.** We observed that the distribution of sequence lengths in the first end on tiles 1104 to 1128 is significantly different from the remainder of the dataset. Lack of sequences having a read length either 151 or 149 or anything between 18 - 39 was observed for these tiles (Figure 3). The second end does not seem to have this peculiarity. We could not rule out post generation processing or trimming of sequences, but it isn't clear to us if any post-processing should create an unusual distribution of read lengths. Further, an unusual tile wise distribution of ambiguous base call (N) was also observed in end 1. Figure 4 shows such an occurrence at position 140 where reads from the tile 1201 to tile 1228 reads have an ambiguous base call. Similarly from the tile 1201 to tile 2128, reads have an ambiguous base call at position 141. Positions 132, 137, 138 and 142 also show such tile specificity.
3. **Tentative presence of 18S rRNA**. We found 25,744 instances of a 150 bases long sequence in end 2:

"CGACGACCCATTCGAACGTCTGCCCTATCAACTTTCGATGGTAGTCGCCGTG CCTACCATGGTGACCACGGGTGACGGGGAATCAGGGTTCGATTCCGGAGAGG GAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGC". This sequence is identical to a segment from 18S rRNA present in many species. There are a total of 4,280 151 bp long sequences with the first 150 bp identical to the above-mentioned sequence. Surprisingly, a 100 percent of the 4,280 occurrences of this sequence had a base calling error in the 151st position making it highly unlikely to be a random error. Another example which we found for such an occurrence was with read number 11603657 on end 2.

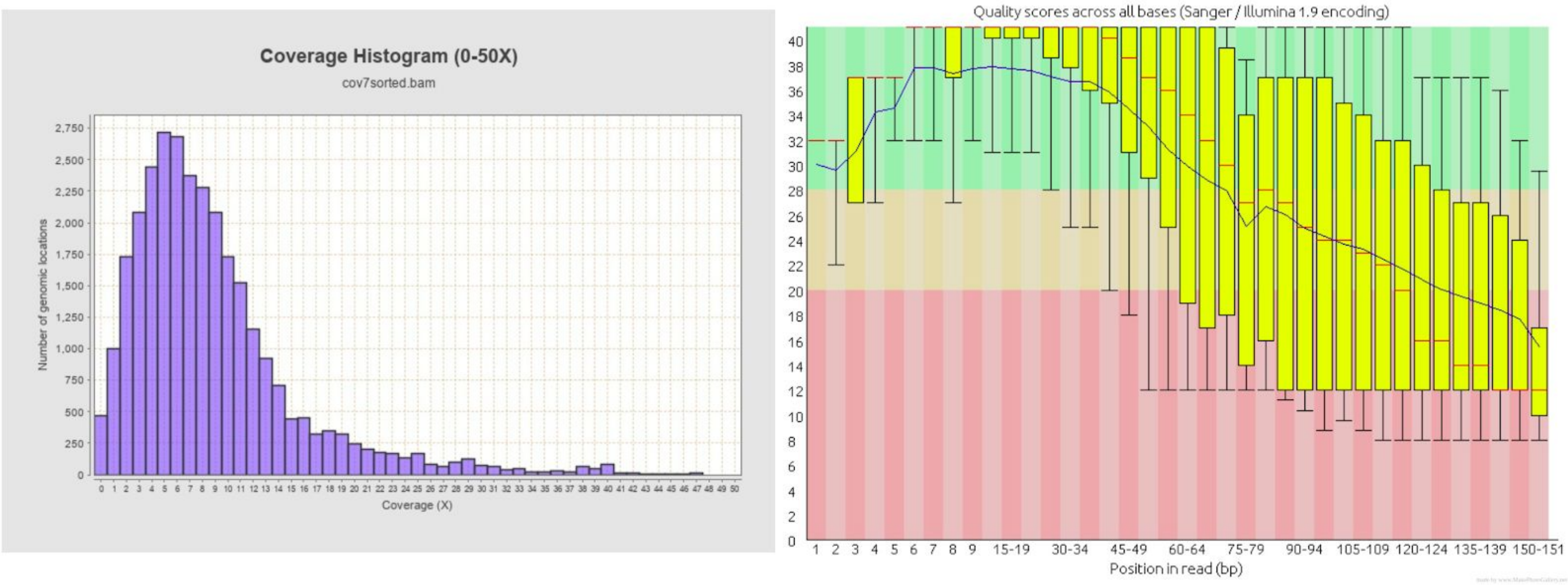

Figure 1. (a) Coverage of bases of RaTG13 genome. (b) Quality scores across for all bases in end2

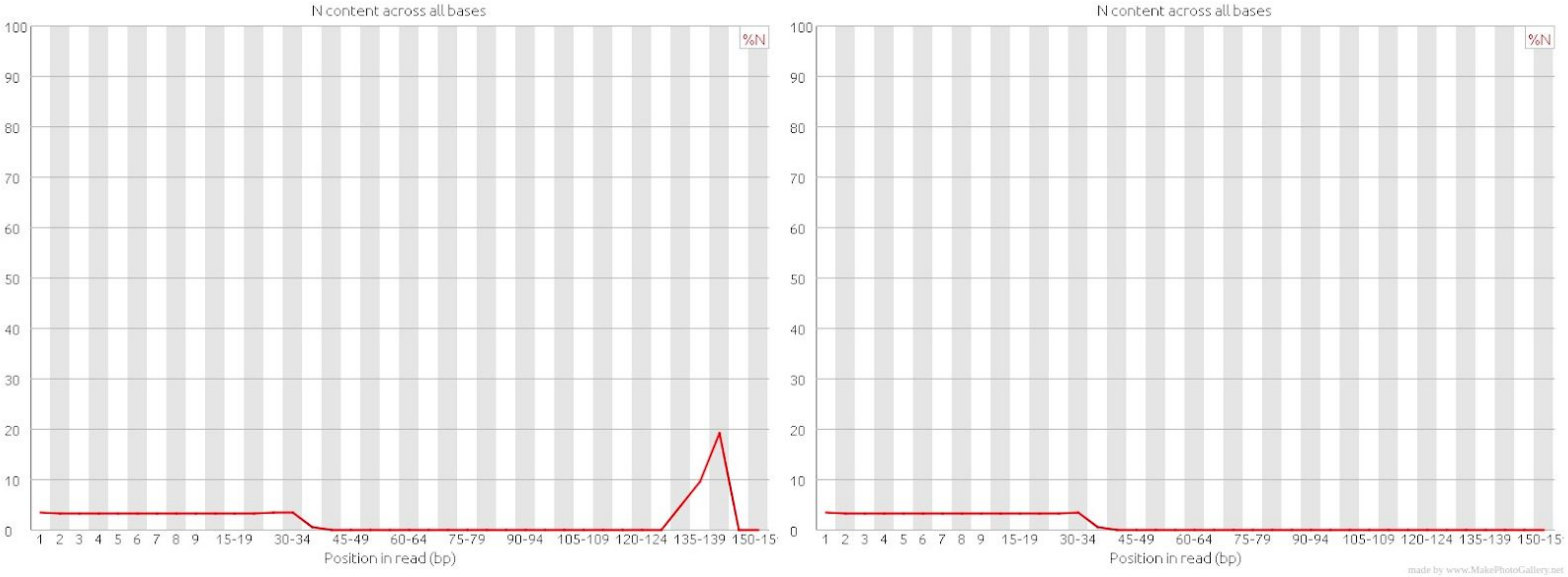

Figure 2. N content per base in both reads. (a) Left plot depicts end 1, (b) Right plot depicts end 2

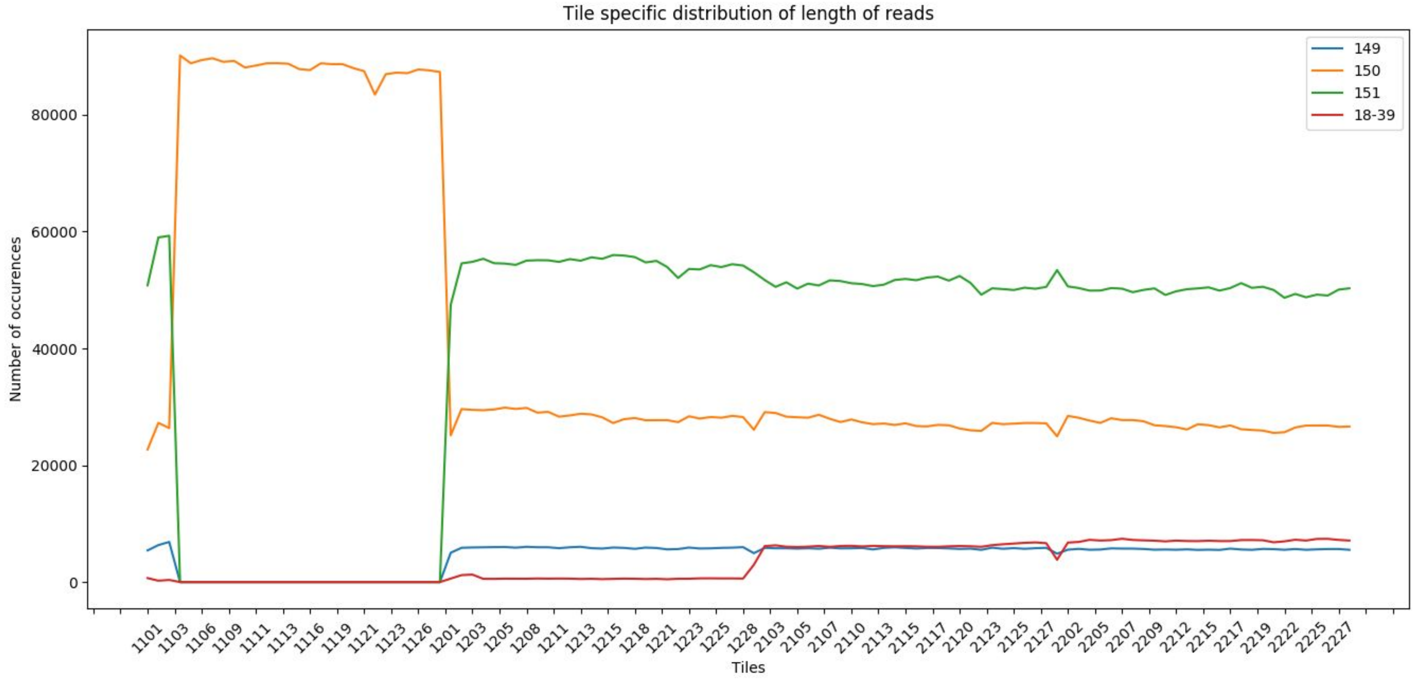

Figure 3. Distribution of read lengths across tiles in end 1

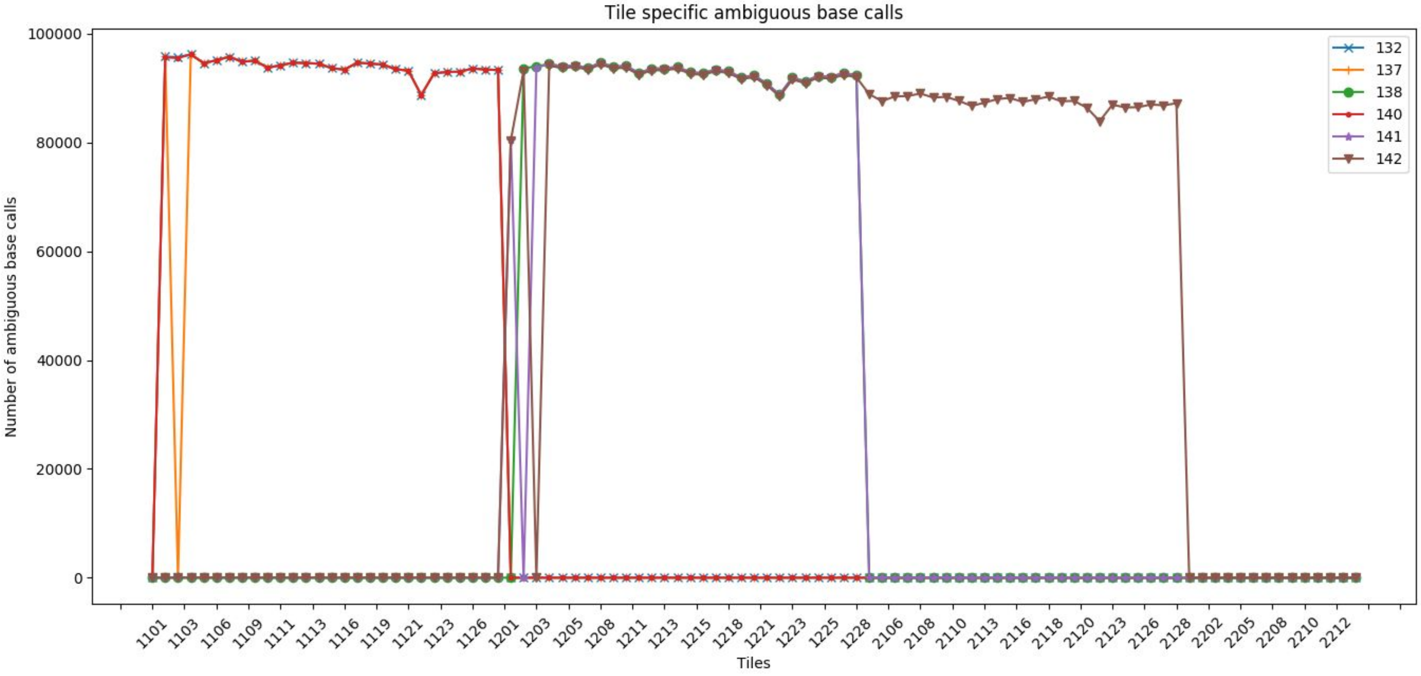

Figure 4. Distribution of ambiguous base calls across tiles in end 1

NCBI KRONA analysis of SRR11085797 raises concerns about the experimental procedure. The datasets (SRR11085797 and SRR11085736) were expected to have similar large amounts of bacterial content[14] given both have similar origins (Fecal and Anal). However, the KRONA[10] analysis of both the samples shows a significant difference in that 0.65% of SRR11085797 is bacteria with 67.91% of it being Eukaryota while SRR11085736 is 91.07% bacteria and 4.36% Eukaryota. Additionally, 29.38% of SRR11085797 is unidentified which is quite different from SRR11085736. Similar observations have been pointed out in other publications[21,23] as well. Further, 0.1% of SRR11085797 hits embryophytes. We find 0.05% Oryza sativa Japonica (Rice), and 0.007% Zea mays (Maize). This was certainly a very odd finding for a bat sample (Rhinolophus Affinis). Therefore we conjecture that the data suffers from contamination by index hopping. The same sequencing platform seems to have been used earlier for sequencing maize as evidenced by data from another BioProject, specifically, the accessions PRJNA338015 & PRJNA383075 (from Huazhong Agriculture university) bearing the same platform "ST-J00123" and having a lower run number. The occurrence of maize like reads could be explained by index hopping if a maize sample was being sequenced in a multiplex lane simultaneously. We also found that 0.2% of the dataset hit Manis javanica (which is not known to reside in mines and does not even have the same order as Rhinolophus Affinis in the hierarchy of organisms). This is important because there have been reports that 2019-nCoV may have originated from pangolins[11]. Although there is a small likelihood, it cannot be dismissed that some fragments of RaTG13 may have index hopped from pangolins if the experiment was not done in separate lanes and may be sparking a misled debate and erroneous scientific inferences about the origins of SARS-CoV-2. And one might also infer, seeing the various hits to unrelated organisms, that the portions of Bat Coronavirus RaTG13 (approximately 0.008%) in the sample might be from a contaminant.

It was also observed that SRR11085736 has sequences resembling another virus (accession MN611520.1 apart from the virus identified from its own accession i.e. MN611522.1). SRR11085736 is claimed to have been derived from a separate lane which strongly suggests that index hopping does not play a role here. Interestingly enough the reads present under SRR11085736 have a run number of 89, while the reads under SRR11085797 have a run number 84. Thus, the presence of cross-contamination of samples in a later run with specific guarding against index hopping raises a distinct possibility that sample from previous runs might not have been guarded against either index hopping or cross-contamination and which might explain the diverse contaminations we find in the KRONA analysis of SRR11085797.

We also found several sequences that bear resemblance to retroviruses. However, we could not conclusively attribute this to index hopping. For example, fragments of a Rhinolophus

doi:10.20944/preprints202008.0595.v1

ferrumequinum retrovirus were found in the sample but we were unsuccessful in assembling a complete retrovirus.

Discussion

We report inability to reproduce the 29855 base long RaTG13 sequence from the supporting dataset as well as the observations from the experiments we performed on the dataset. The said dataset is unique and has much more information apart from scattered fragments of coronavirus. However, details about its generation are vague and critical to infer the origin of the SARS-CoV-2. The scientific community has focused on the human isolate, however we show that the RaTG13 sequence needs equal attention in order to draw valid conclusions about the origins of SARS-CoV-2. Our work, in addition to works by other authors[21,22,23,24], is an attempt to stimulate a dispassionate review of the dataset and to determine the (complete/partial) sequence(s) that emanates from it. We also argue that it may be prudent to withhold conjectures and await details as to the methods adopted that led to the generation of this unique dataset and the genome sequence. The authors reporting the existence of RaTG13 [1] thus need to expeditiously detail the experimental procedure adopted while the scientific community is called upon to refrain from citing the genome of RaTG13 [1] till the entirety of the said genome is actually established as being fully supported by the data by independent reviews. Therefore, our attempt is a call to action for the scientific community for collating evidence about the origins of SARS-CoV-2 so that the risk of future pandemics may be mitigated.

Materials and Methods

Dataset. The dataset for the RaTG13 sequence[3] , isolated from a Fecal swab of a *Rhinolophus affinis* host was made publicly available both from the National Genomics Data Center (NGDC[1], BioSample accession SAMC133252, Run accession CRR122287) as well from NCBI (SRA accession SRR11085797). NDGC[1] provides some details about the sample and sequence generated. This is supposedly a metagenomic RNA dataset generated through sequencing on Illumina HiSeq 3000 platform. It has a paired layout with 11,604,666 reads on one end. The planned read lengths for both the ends were 151 (bp) with Insert size of 280 (bp) and the authors have used CLC Genomics Workbench v12.0 for assembly. Importantly, nine other datasets were deposited by the same research group which were run on the same machine id ST-J00123. These

[1] "BIGD - GSA - National Genomics Data Center." 13 Mar. 2020, https://bigd.big.ac.cn/gsa/browse/CRA002424/CRX097481. Accessed 10 Aug. 2020.

are available on NCBI under the BioProject accession PRJNA606159 and were used to draw comparisons with the RaTG13 dataset in question.

Fastq files and Quality of reads. We downloaded the paired fastq files from NGDC[1] (CRR122287_f1.fastq.gz & CRR122287_r2.fastq.gz). Seq.IO from the Biopython package was used to iterate over the reads and find reads with desirable sequences. Additionally, the information about run number and machine number, tile numbers, read lengths was also extracted from read headers using the Biopython package. FastQC[15] software was used to find per base sequence quality, percentage of overrepresented sequences, per base N content, etc.

Sequence similarity. We used megaBLAST[5] upon the dataset submitted to NCBI (under the run number SRR11085797) against various sequences to get the number of reads hitting the sequence and coverage of the sequence. Furthermore, we used the 'Reads' tab in NCBI run browser to search for reads having specific sequences.

De novo assembly. Megahit[4] with default settings[20] and two other custom settings were used for the de novo assembly of the RaTG13 genome. The custom settings had the following modifications- (i) reduced maximum k-mer size to 79, looking at both k 79 contigs and the final contigs[4] and, (ii) an adoption of settings recommended with k-step = 10 and the --no-mercy option.

Coverage calculation. bbmap.sh script was used from the BBMap[6] package (version 35.34) to calculate average coverage and get the corresponding BAM file. BBMap[6] is a splice-aware global aligner for DNA/RNA reads which uses a multi-kmer-seed-and-extend approach [6]. BBMap[6] aligner was used because it has no upper limit on the genome size or number of reads and based on multiple comparative studies [18,19], it was found to be fairly robust.

Quality analysis and visualization. Quality indices were visualized using Qualimap[13] package (v 2.2.1) on the alignment BAM file, FastQC[15] report and the Matplotlib[17] library in Python. Biopython[16] was used to find read lengths and ambiguous base calls across tiles.

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.

References

1) Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. MN996532, Bat coronavirus RaTG13, complete genome; [cited 2020 June 16]. Available from: https://www.ncbi.nlm.nih.gov/nuccore/MN996532
2) Zhou, P., Yang, X., Wang, X. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7
3) SRA [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. SRR11085797, RNA-Seq of Rhinolophus affinis:Fecal swab; [cited 2020 June 16]. Available from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR11085797
4) Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;31(10):1674-1676. doi:10.1093/bioinformatics/btv033
5) Blast [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 June 16]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi
6) Bushnell, Brian. Mon . "BBMap: A Fast, Accurate, Splice-Aware Aligner". United States. https://www.osti.gov/servlets/purl/1241166.
7) Jaimes JA, André NM, Chappie JS, Millet JK, Whittaker GR. Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop. *J Mol Biol*. 2020;432(10):3309-3325. doi:10.1016/j.jmb.2020.04.009
8) Huang, Jiao-Mei & Jan, Syed & Wei, Xiaobin & Wan, Yi & Ouyang, Songying. (2020). Evidence of the Recombinant Origin and Ongoing Mutations in Severe Acute Respiratory Syndrome 2 (SARS-COV-2). 10.1101/2020.03.16.993816.
9) Zhang, Tao et al. "Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak." *Current biology : CB* vol. 30,7 (2020): 1346-1351.e2. doi:10.1016/j.cub.2020.03.022
10) Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*. 2011;12:385. Published 2011 Sep 30. doi:10.1186/1471-2105-12-385
11) Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak [published correction appears in Curr Biol. 2020 Apr 20;30(8):1578]. *Curr Biol*. 2020;30(7):1346-1351.e2. doi:10.1016/j.cub.2020.03.022

12) Mao, X., Liu, Y., Zhou, Y. *et al.* Development of 19 polymorphic microsatellite loci for the intermediate horseshoe bat, *Rhinolophus affinis* (Rhinolophidae, Chiroptera). *Conserv Genet* 10, 709–711 (2009). https://doi.org/10.1007/s10592-008-9625-y

13) Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." Bioinformatics, btv566

14) Rose C, Parker A, Jefferson B, Cartmell E. The Characterization of Feces and Urine: A Review of the Literature to Inform Advanced Treatment Technology. Crit Rev Environ Sci Technol.2015; 45(17):1827-1879. doi:10.1080/10643389.2014.1000761

15) Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

16) Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25, 1422-1423

17) J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

18) Smith, H. E., & Yun, S. (2017). Evaluating alignment and variant-calling software for mutation identification in C. elegans by whole-genome sequencing. *PloS one*, *12*(3), e0174446. https://doi.org/10.1371/journal.pone.0174446

19) https://www.ecseq.com/support/ngs/best-RNA-seq-aligner-comparison-of-mapping-tools

20) Github - Megahit, https://github.com/voutcn/megahit. Accessed 10 Aug. 2020.

21) Rahalkar, M.; Bahulikar, R. The Abnormal Nature of the Fecal Swab Sample used for NGS Analysis of RaTG13 Genome Sequence Imposes a Question on the Correctness of the RaTG13 Sequence . Preprints 2020, 2020080205 (doi: 10.20944/preprints202008.0205.v1).

22) lin, X.; Chen, S. Major Concerns on the Identification of Bat Coronavirus Strain RaTG13 and Quality of Related Nature Paper. Preprints 2020, 2020060044 (doi: 10.20944/preprints202006.0044.v1).

23) Daoyu Zhang. (2020, August 1). Anomalies in BatCoV/RaTG13 sequencing and provenance. Zenodo. http://doi.org/10.5281/zenodo.3987503

24) Rahalkar, M.C.; Bahulikar, R.A. Understanding the Origin of 'BatCoVRaTG13', a Virus Closest to SARS-CoV-2. Preprints 2020, 2020050322 (doi: 10.20944/preprints202005.0322.v1).