 doi:10.20944/preprints202008.0659.v1

Noisy genome data and faulty clade statistics undermine conclusions on SARS-CoV-2 evolution and strain typing in the Brazilian epidemy: A Technical Note

Marcelo R. S. Briones[1], Fernando Antoneli[1], Renata C. Ferreira[1], Isabel M. V. G. Carvalho[2], Luiz M. R. Janini[3]

[1]Setor de Bioinformática, Departamento de Informática em Saúde, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, Brazil.

[2]Laboratório de Parasitologia, Instituto Butantan, São Paulo, Brazil.

[3]Laboratório de Retrovirologia e Departamento de Microbiologia, Imunologia, Parasitologia, Universidade Federal de São Paulo, São Paulo, Brazil.

Correspondence to: marcelo.briones@unifesp.br and mjanini@hotmail.com

Abstract

We show that low quality of all 427 Brazilian SARS-CoV-2 genomes recently published in *Science* (*1*) challenges their phylogenetic inference and may lead to incorrect typing of viral strains in clades with no statistical support. Absence of basecalling quality in genome assemblies and proper phylogeny parameter estimates preclude the assessment of signal-to-noise ratio in the data, downstream analysis and conclusions.

According to our assessment, the recent study on SARS-CoV-2 genomes in Brazil, published in *Science* (*1*), has critical errors that undermine the conclusions drawn from sequence data and analysis. Initially, the inspection of GISAID database using the accession numbers provided in the paper shows that the average number of Ns in the assembled sequences is ≈2% to 3% (Fig. 1). None of the 427 genomes sequenced have less than 1% Ns, an important exclusion criterium in GISAID. Accordingly, in a ≈30,000 bases genome, 1% Ns means ≈300 Ns in the sequence, which is bigger than Gene E with 228 bases. The 427 Brazilian genomes sequenced in this paper have therefore between ≈600 and ≈900 Ns, which are comparable in size with ORF7a (366 bases), Gene M (669 bases) and ORF3a (828 bases). Therefore, more than 1%Ns should be unacceptable by quality standards in genomic and phylogenetic analysis of these viruses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0659.v1

The high number of uncalled bases (Ns>1%), besides undermining the information on those specific positions, affect the quality scores of flanking segments. As observed in sequences downloaded from GISAID the Ns occur in long stretches of ≈50-100 residues (Fig. 1D). Assemblies using Nanopore reads show, in general, a decrease of phred scores in positions flanking the stretches of Ns, which means an increase of uncertainty (noise) in the data. Therefore, the effect of 2% Ns could be, in fact, much higher. Supplementary Data S1 in (*1*) contain information on coverage and number of reads but do not contain information on base quality. Without quality scores of individual positions, it is impossible to assess the signal-to-noise level in this study to determine the impact of sequencing errors and assembly inaccuracy in phylogenetic patterns and dynamics. The noise could be so significant that the tree in Fig. 3A in (*1*) might be interpreted as the temporally structured topology of the data noise. Here, by noise, we mean technical noise, not in the sense of phylogenetic signal-to-noise ratio of macroevolutionary studies (*2*). As a matter of fact, as shown below, sequences with more than 5% up to 15% Ns are included in the analysis.

In the absence of detailed information, we assume that the whole columns containing Ns and gaps were excluded from phylogenetic inference. After the proper exclusion of positions with gaps and Ns how many polymorphic positions and invariant sites remained for the inference? We could not find this information either in the main text or in the Supplementary Material. In other words, if for example a single sequence in the alignment of 1182 genomes contains a gap, or N, in position 15,000, that whole position in the alignment (whole column) should be excluded since neither gaps nor Ns can be treated as 5th state of the character.

No measures of support for phylogeny branch clusters or Clades 1, 2 and 3 (Fig. 3A and Fig. S8 in (*1*)) are presented. The Maximum clade credibility (MCC) summary trees presented do not address the support of individual clades (Supplemental Material pg. 5). Although maximum likelihood and Bayesian trees have been included, bootstrap frequencies and posterior probabilities (*p.p.*) are omitted. Monophyletic clades should be defined with ≥95% bootstrap frequencies (*3*) and *p.p.* should be close to 1 in Bayesian trees (*4*). Also, the *LnL* of the phylogenies are not shown (e.g. Fig. S7 in the Supplementary Material in (*1*)).

A time-resolved tree was presented to mitigate hard polytomies (Fig. 3A in (*1*)). However, temporal data impose an external constraint on branches and topologies which might, in turn, create artifactual clustering of sequences that share very close collection dates. Since ≈10% of positions are polymorphic, minute changes deeply affect splitting and clustering. In the absence of branch support values, it is impossible to determine whether the time-resolved tree significantly improved bootstrap, *p.p.* and likelihood as compared to fully unconstrained trees.

A strict global clock was assumed (Supplementary Material pg. 5 in (*1*)) when data from Fig. 3A in (*1*) (upper right panel) suggests an overdispersed dynamics or possibly a fixed local clock in the Brazilian sequences. It is therefore necessary that the authors, in face of using temporal data to “time-adjust” their trees, test which distributions (e.g. lognormal, exponential) and clock types best fit the data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0659.v1

Typing of individual sequences into clades is based on one or two SNVs for which no phred score or any other measure of support is available. For example, the authors state (pg. 3) that "Clade 1 is characterized by a G25088T SNV in the Spike gene" and that "Clade 2 is ***defined*** by SNVs T27299C and T29148C". For example, if in a given sample the G25088T is present in a position with phred=5 there is a significant chance that this is incorrect. However, if this change is in a position with phred=30 the confidence in the correct classification is 99.9%. Also, how can clades be defined with such certainty in the absence of any statistical support (bootstrap and *p.p.*)? How can samples be placed in a specific cluster when the number of character-state changes required are below the threshold of the methods used for sequencing/assembly? Even if phred≥40 was required for GISAID submission, phred=40 means that there is ≈1 estimated wrong base call every 10,000 bases, or 3 wrongly called bases in the SARS-CoV-2 genome (≈30,000 bases). Generally, the genome sequence is considered finished when the average phred≥40 score for the assembly is achieved. This criterium was used in Brazil for the of *Xylella fastidiosa* genome sequencing (*5*). In viral genomics the phred score threshold is usually around 30 (1 error per 1kb). Nevertheless the abandonment of phred≥40, or phred≥30, for finishing genome sequences was not discussed in Candido *et al.* (*1*).

No justification is provided for selecting the Hasegawa-Kishino-Yano model (HKY) (*6*) with Gamma distribution in the maximum likelihood phylogenies (Supplementary Material pg. 5). The α parameter, indicative of among-site evolutionary rate heterogeneity, is also not shown. For example, if α=100 the gamma distribution is unnecessary. The substitution matrix (model) and parameters should be estimated during the run or by goodness of fit nested testing (*7*). The HKY model and nucleotide frequencies were apparently arbitrarily selected. In the Bayesian inference the critical selection of priors was not detailed and the 10% burn-in removal, instead of the more proper 50% burn-in removal (a point at which the likelihood values stabilize), was not justified.

The authors state that "All data, code, and materials used in the analysis are available on DRYAD (*40*)." This points to Reference (*40*). This reference contains a link supposed to direct to these data (https://doi.org/10.5061/dryad.rxwdbrv5z). However, this link leads to a screen "DOI Not Found" (Fig. 7). Therefore, it is not possible to download the materials as stated in the paper. The authors might be able to promptly correct this problem by issuing an *Errata*.

The raw data of sequencing reads ERR4368102_MN908947.3.fastq from PRJEB39487 as stated in the paper was downloaded from the NCBI website. ERR4368102_MN908947.3.fastq contains 77,468 reads (read lengths from 355 bases to 574 bases and with quality scores). These reads were assembled to reference NC_045512 using minimap2 (the same program used by the authors). The assembled consensus sequence was exported as fastq file and aligned with EPI_ISL_470568, EPI_ISL_470570 and NC_045512. Few positions with Ns were found in the beginning and the end of the alignment (Fig. 8). However, two large blocks of Ns (NB1 and NB2) exist in the coding region. NB1 (N Block 1) is located between positions 20,543 and 20,813 (numbering of reference NC_045512) within ORF1a and affect the encoding sequences for peptides "endo RNAse" and "2'-O-ribose methyltransferase" (Fig. 9). NB2 is located between

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0659.v1

22,325 and 22,542 within the coding region of the Spike protein (Fig. 10). The authors state in the paper (Supplementary Materials pg. 4) that in their assemblies "*Genome regions with a depth of <20-fold were not included in final consensus sequences, and these positions are represented with N characters*". In the case of NB2, replacement of low coverage regions by Ns is acceptable since this region contains low coverage and low quality bases (darker blue indicate quality score below 20 and lighter blue below 30) (Fig. 10). However, the coverage in NB1, as indicated by our assembly, seems to be zero. No reads mapped in this region and that is why it is represented by (?), indicating missing data, not gaps or Ns. Our analysis suggests that this is a real gap in the genome assembly, not a virtual gap. The way this is written in the paper seems misleading. The authors should clarify whether they observe these real gaps in their assemblies. The character N indicates uncertainty in basecalling while (?) indicates missing data. Therefore, the genomes presented and analyzed seem to have real gaps and not only low coverage virtual gaps, as the text of the authors imply. The authors should close these gaps using flanking primers with Nanopore or with orthogonal methods (i.e. Sanger). Nevertheless, these genomes seem to be incomplete and the Ns affect several positions in the Spike gene, that encodes for a very important protein in viral biology and vaccine target strategies and at least two other relevant genes for the virus. If this is the case, we assume that these positions are excluded from all downstream analysis presented in the paper.

In conclusion, if GISAID required the **fastq** files of the consensus of the assemblies instead of **fasta**, most of the problems described above would have been avoided. The corresponding BAM files should also be submitted. We contend that problems at the level of sequencing/assembly quality have precedence over sample bias, sampling errors and tree rooting in SARS-CoV-2 epidemiological studies based on phylogenies and networks (*8*, *9*). In our search of SARS-CoV-2 using GISAID data, we excluded all sequences with >1% Ns, not complete (<29,000 bases) and without patient data. Therefore, among the ≈63,000 GISAID SARS-CoV-2 sequences (as of July 13, 2020) only 3,409 satisfy these criteria, being 64 Brazilian sequences (81 sequences as of July 24, 2020). This stringency preferentially discards noise over data. Also, in GISAID there are 177 Brazilian SARS-CoV-2 complete sequences (>29kb) that have <1% Ns (as of July 24, 2020) but none of these are from Candido *et al*. (*1*). The "big data" proposition that inclusion of more data, regardless of quality, somehow "dilutes" the errors, is false (*10*). As demonstrated by Meng (*10*), inclusion of large amounts of low quality, noisy data will amplify errors and lead to false conclusions. Alignments with GISAID data show global spurious bases (Fig. 2A) and possible artifactual deletions (Fig. 2B). These exemplify the inaccuracies in these data when not properly curated. As a final remark, a genome sequence deposited in a database is an inference not data. The data are the individual reads outputted by the sequencer. These reads are used to assemble, or infer, the genome. The deposited genome is therefore a hypothesis and as such, requires a statistical support measure to be valid for additional analysis.

References

1. D. S. Candido, I. M. Claro, J. G. de Jesus, W. M. Souza, F. R. R. Moreira, S. Dellicour, T. A. Mellan, L. du Plessis, R. H. M. Pereira, F. C. S. Sales, E. R. Manuli, J. Thézé, L. Almeida, M. T. Menezes, C. M. Voloch, M. J. Fumagalli, T. M. Coletti, C. A. M. da Silva, M. S. Ramundo, M. R. Amorim, H. H. Hoeltgebaum, S. Mishra, M. S. Gill, L. M. Carvalho, L. F. Buss, C. A. Prete, J. Ashworth, H. I. Nakaya, P. S. Peixoto, O. J. Brady, S. M. Nicholls, A. Tanuri, Á. D. Rossi, C. K. V. Braga, A. L. Gerber, A. P. de C. Guimarães, N. Gaburo, C. S. Alencar, A. C. S. Ferreira, C. X. Lima, J. E. Levi, C. Granato, G. M. Ferreira, R. S. Francisco, F. Granja, M. T. Garcia, M. L. Moretti, M. W. Perroud, T. M. P. P. Castiñeiras, C. S. Lazari, S. C. Hill, A. A. de S. Santos, C. L. Simeoni, J. Forato, A. C. Sposito, A. Z. Schreiber, M. N. N. Santos, C. Z. de Sá, R. P. Souza, L. C. Resende-Moreira, M. M. Teixeira, J. Hubner, P. A. F. Leme, R. G. Moreira, M. L. Nogueira, D. Brazil-UK Centre for Arbovirus Discovery, N. M. Ferguson, S. F. Costa, J. L. Proenca-Modena, A. T. R. Vasconcelos, S. Bhatt, P. Lemey, C.-H. Wu, A. Rambaut, N. J. Loman, R. S. Aguiar, O. G. Pybus, E. C. Sabino, N. R. Faria, Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* (2020), doi:10.1126/science.abd2161.

2. D. M. Hillis, J. P. Huelsenbeck, Signal, Noise, and Reliability in Molecular Phylogenetic Analyses. *J Hered*. **83**, 189–195 (1992).

3. J. Felsenstein, CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution*. **39**, 783–791 (1985).

4. M. E. Alfaro, S. Zoller, F. Lutzoni, Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence. *Mol Biol Evol*. **20**, 255–266 (2003).

5. A. J. G. Simpson, F. C. Reinach, P. Arruda, F. A. Abreu, M. Acencio, R. Alvarenga, L. M. C. Alves, J. E. Araya, G. S. Baia, C. S. Baptista, M. H. Barros, E. D. Bonaccorsi, S. Bordin, J. M. Bové, M. R. S. Briones, M. R. P. Bueno, A. A. Camargo, L. E. A. Camargo, D. M. Carraro, H. Carrer, N. B. Colauto, C. Colombo, F. F. Costa, M. C. R. Costa, C. M. Costa-Neto, L. L. Coutinho, M. Cristofani, E. Dias-Neto, C. Docena, H. El-Dorry, A. P. Facincani, A. J. S. Ferreira, V. C. A. Ferreira, J. A. Ferro, J. S. Fraga, S. C. França, M. C. Franco, M. Frohme, L. R. Furlan, M. Garnier, G. H. Goldman, M. H. S. Goldman, S. L. Gomes, A. Gruber, P. L. Ho, J. D. Hoheisel, M. L. Junqueira, E. L. Kemper, J. P. Kitajima, J. E. Krieger, E. E. Kuramae, F. Laigret, M. R. Lambais, L. C. C. Leite, E. G. M. Lemos, M. V. F. Lemos, S. A. Lopes, C. R. Lopes, J. A. Machado, M. A. Machado, A. M. B. N. Madeira, H. M. F. Madeira, C. L. Marino, M. V. Marques, E. a. L. Martins, E. M. F. Martins, A. Y. Matsukuma, C. F. M. Menck, E. C. Miracca, C. Y. Miyaki, C. B. Monteiro-Vitorello, D. H. Moon, M. A. Nagai, A. L. T. O. Nascimento, L. E. S. Netto, A. Nhani, F. G. Nobrega, L. R. Nunes, M. A. Oliveira, M. C. de Oliveira, R. C. de Oliveira, D. A. Palmieri, A. Paris, B. R. Peixoto, G. a. G. Pereira, H. A. Pereira, J. B. Pesquero, R. B. Quaggio, P. G. Roberto, V. Rodrigues, A. J. de M. Rosa, V. E. de Rosa, R. G. de Sá, R. V. Santelli, H. E. Sawasaki, A. C. R. da Silva, A. M. da Silva, F. R. da Silva, W. A. Silva, J. F. da Silveira, M. L. Z. Silvestri, W. J. Siqueira, A. A. de Souza, A. P.

de Souza, M. F. Terenzi, D. Truffi, S. M. Tsai, M. H. Tsuhako, H. Vallada, M. A. Van Sluys, S. Verjovski-Almeida, A. L. Vettore, M. A. Zago, M. Zatz, J. Meidanis, J. C. Setubal, The genome sequence of the plant pathogen Xylella fastidiosa. *Nature*. **406**, 151–157 (2000).

6. M. Hasegawa, H. Kishino, T. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. **22**, 160–174 (1985).

7. D. Posada, K. A. Crandall, MODELTEST: testing the model of DNA substitution. *Bioinformatics*. **14**, 817–818 (1998).

8. C. Mavian, S. K. Pond, S. Marini, B. R. Magalis, A.-M. Vandamme, S. Dellicour, S. V. Scarpino, C. Houldcroft, J. Villabona-Arenas, T. K. Paisie, N. S. Trovão, C. Boucher, Y. Zhang, R. H. Scheuermann, O. Gascuel, T. T.-Y. Lam, M. A. Suchard, A. Abecasis, E. Wilkinson, T. de Oliveira, A. I. Bento, H. A. Schmidt, D. Martin, J. Hadfield, N. Faria, N. D. Grubaugh, R. A. Neher, G. Baele, P. Lemey, T. Stadler, J. Albert, K. A. Crandall, T. Leitner, A. Stamatakis, M. Prosperi, M. Salemi, Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. *Proc Natl Acad Sci USA*. **117**, 12522–12523 (2020).

9. C. Mavian, S. Marini, M. Prosperi, M. Salemi, A Snapshot of SARS-CoV-2 Genome Availability up to April 2020 and its Implications: Data Analysis. *JMIR Public Health Surveill*. **6**, e19170 (2020).

10. X.-L. Meng, Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12**, 685–726 (2018).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0659.v1

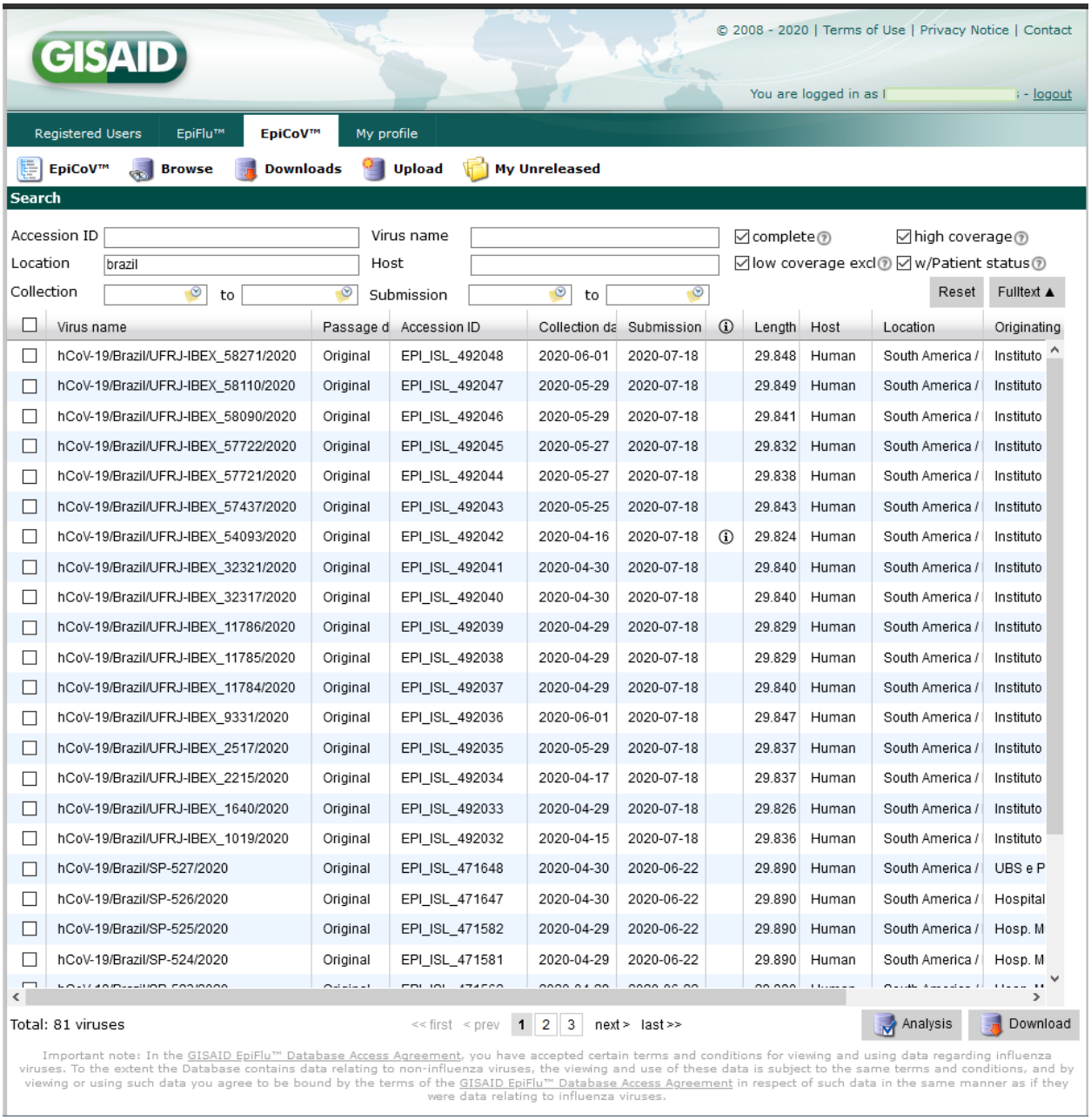


Figure 1. GISAID browse. Checking the four boxes for maximal quality (upper right) exclude all 427 sequences of the Candido et al. paper on *Science*. As of July 13, 2020, there were 64 Brazilian good quality sequences. By July 24 it increased to 81. None sequenced in the Candido *et al*. study (*1*) .

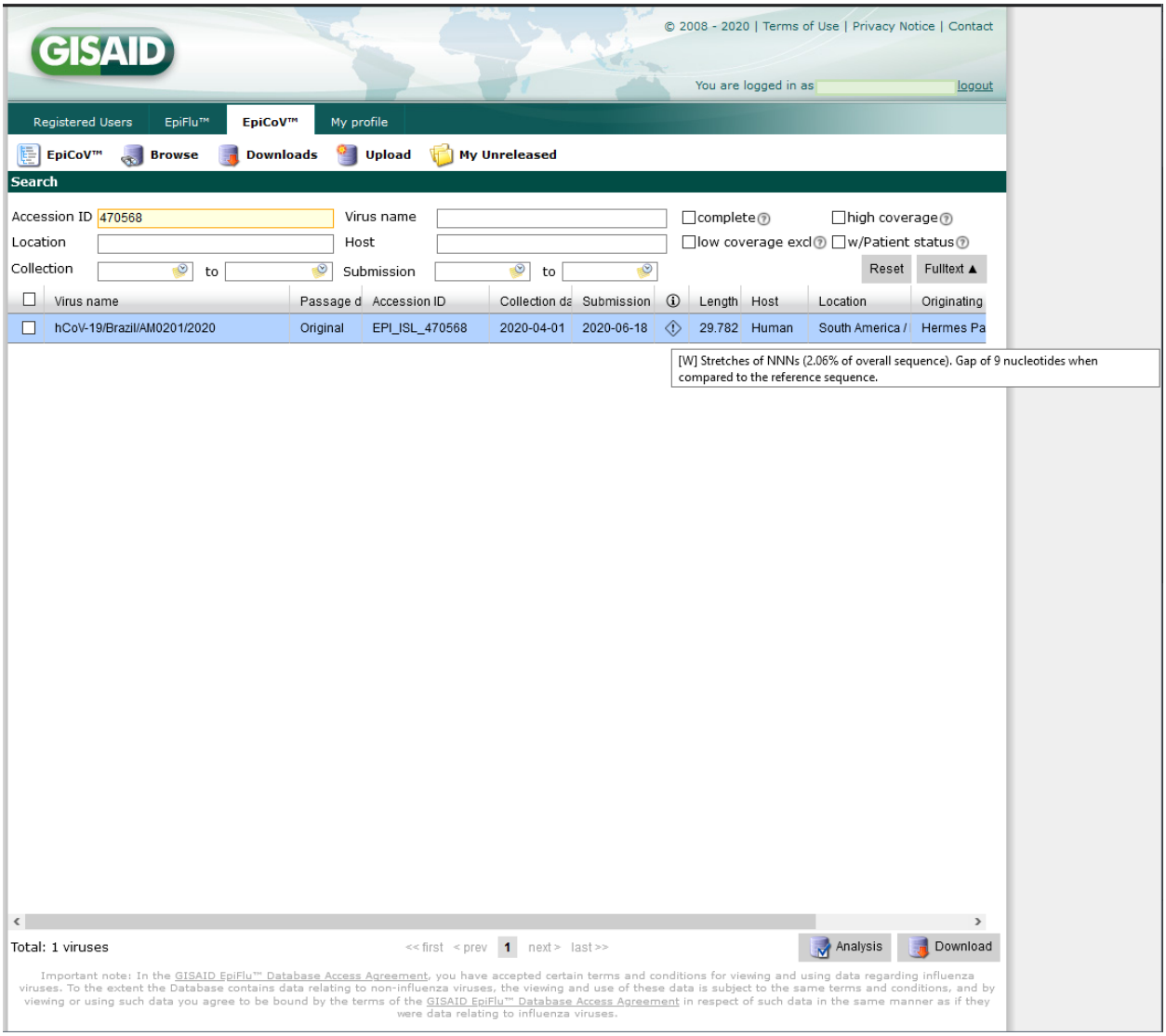


Figure 2. Sequence EPI_ISL_470468, the first in the list of sequences made by Candido *et al*. and listed in pg. 8 of the paper (*1*).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0659.v1

Figure 3. Another example of low quality (>1% Ns) in sequence EPI_ISL_470570 also listed in pg. 8 of the paper. Again, 2% Ns, which corresponds to 600 positions in the coding region.

 doi:10.20944/preprints202008.0659.v1

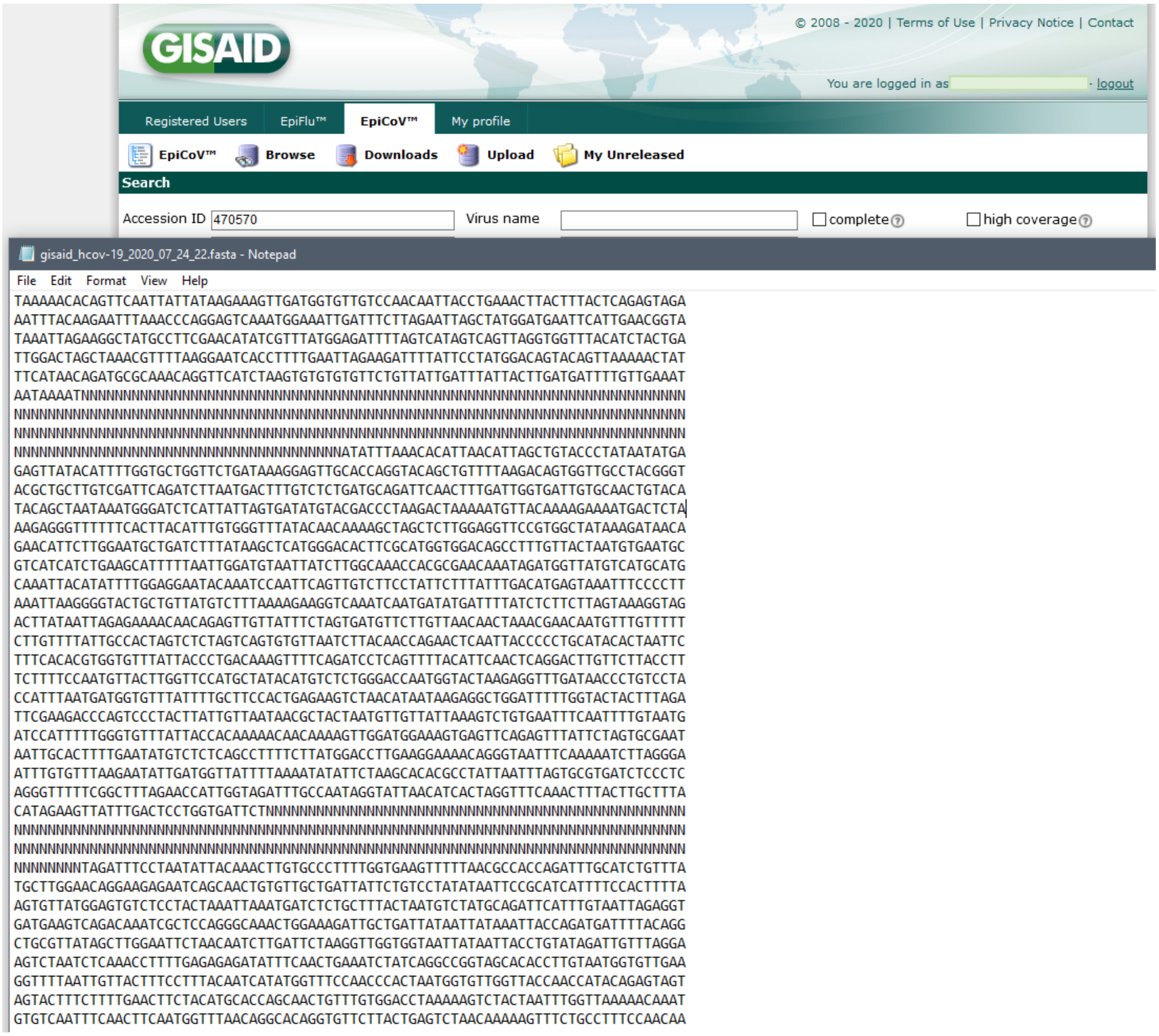


Figure 4. Stretches of Ns in the coding region of Brazilian sequence EPI_ISL_470570. This cannot be considered a “finished” genome by current standards.

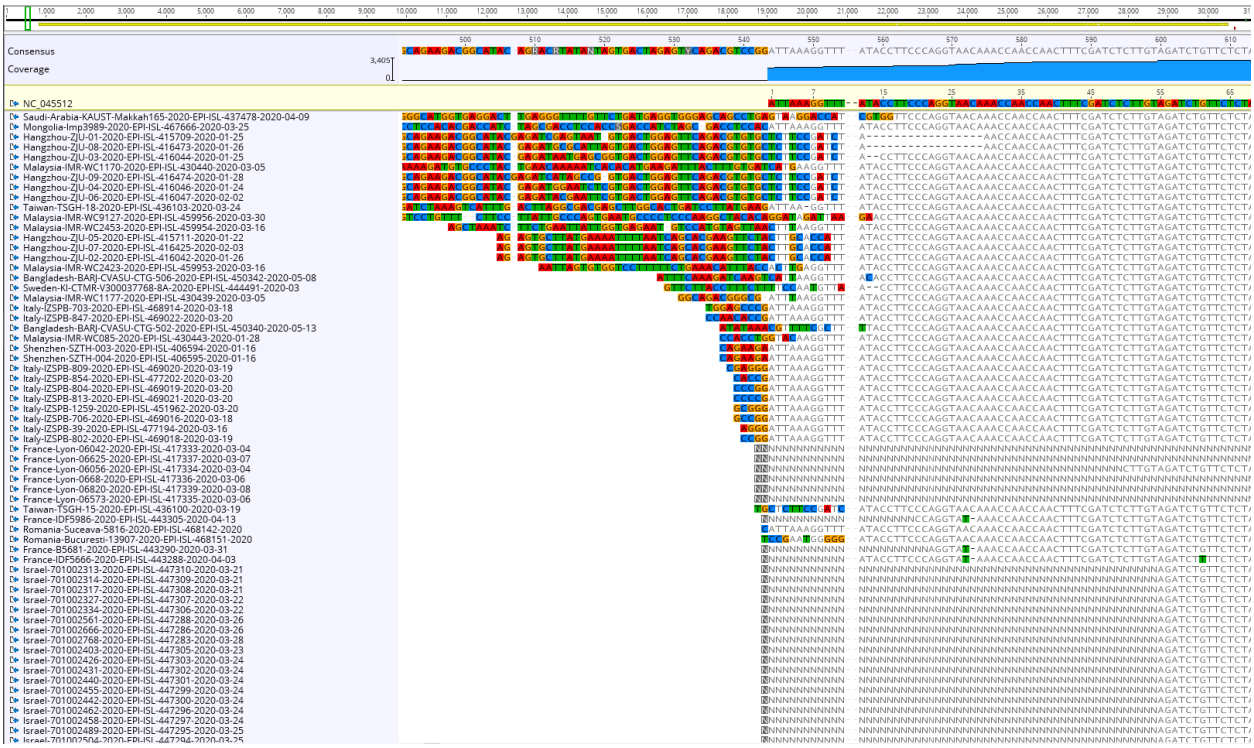


Figure 5. Artifacts and other issues as revealed by the alignment of SARS-CoV-2 genomes from GISAID filtered by checking the four quality boxes. Even with these filters the qualities of the sequences are clearly undermined. Phred quality scores of the assemblies are absolutely necessary to resolve ambiguities.

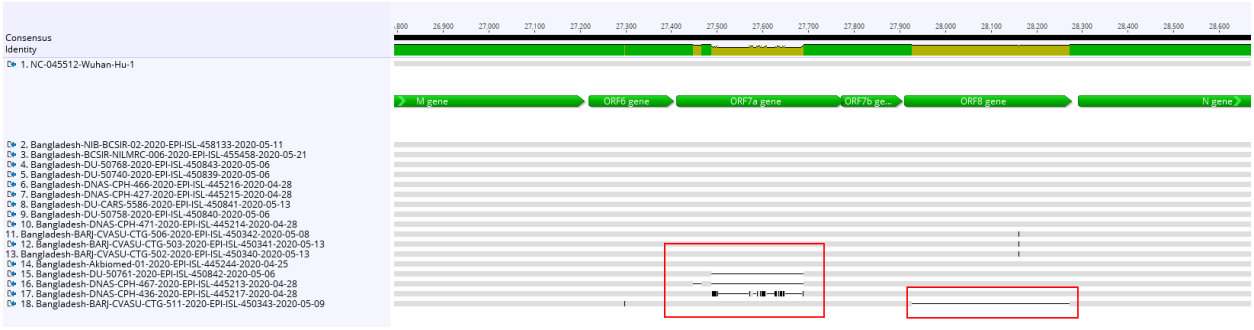


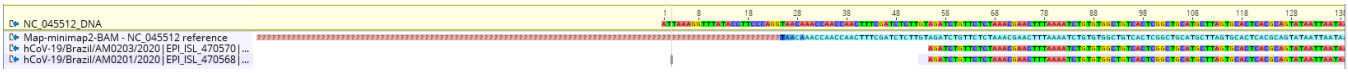
Figure 6. Potential artifactual deletions in four sequences of Bangladesh that passed the four boxes filters of GISAID. Are these deletions real? In sequence 450343 the whole ORF8 is deleted. This could be significant but without base quality data is just noise.

The link in the paper:

Granja, M. T. Garcia, M. L. Moretti, M. W. Perroud Jr., T. M. P. P. Castiñeiras, C. S. Lazari, S. C. Hill, A. A. de Souza Santos, C. L. Simeoni, J. Forato, A. C. Sposito, A. Z. Schreiber, M. N. N. Santos, C. Zolini de Sá, R. P. Souza, L. C. Resende-Moreira, M. M. Teixeira, J. Hubner, P. A. F. Leme, R. G. Moreira, M. L. Nogueira, CADDE-Genomic-Network, N. M. Ferguson, S. F. Costa, J. L. Proenca-Modena, A. T. R. Vasconcelos, S. Bhatt, P. Lemey, C.-H. Wu, A. Rambaut, N. J. Loman, R. S. Aguiar, O. G. Pybus, E. C. Sabino, N. Rodrigues Faria, Evolution and epidemic spread of SARS-CoV-2 in Brazil, Dryad (2020); https://doi.org/10.5061/dryad.rxwdbrv5z.

41. A. Aktay, S. Bavadekar, G. Cossoul, J. Davis, D. Desfontaines, A. Fabrikant, E. Gabrilovich, K. Gadepalli, B. Gipson, M. Guevara, C. Kamath, M. Kansal, A. Lange, C. Mandayam, A. Oplinger, C. Pluntke, T. Roessler, A. Schlosberg, T. Shekel, S. Vispute, M. Vu, G. Wellenius, B. Williams, R. J. Wilson, Google COVID-19 community mobility reports: Anonymization process description (version 1.0).

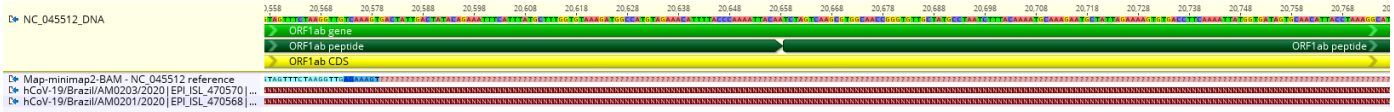
Figure 7. Missing download link for paper data and materials as of 3:04 PM 7/26/2020 (Red Boxes).



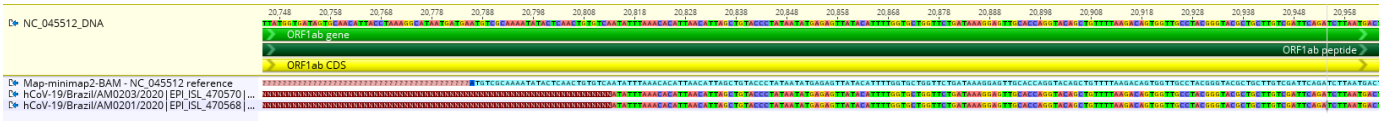
(A) 5’ end of the alignment.

(B) 3’ end of the alignment.

Figure 8. Terminal portions of the alignment of the ERR4368102_MN908947.3.fastq assembled consensus sequence was exported as fastq file and aligned with EPI_ISL_470568, EPI_ISL_470570 and NC_045512. In (A) the 5’ end and in (B) the 3’ end.

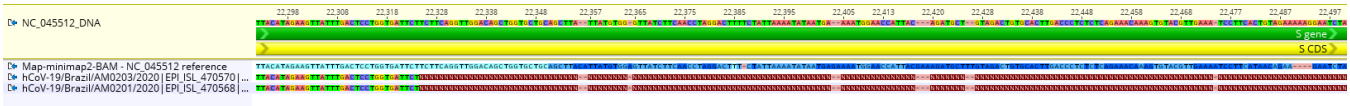


(A) NB1 5’ end.

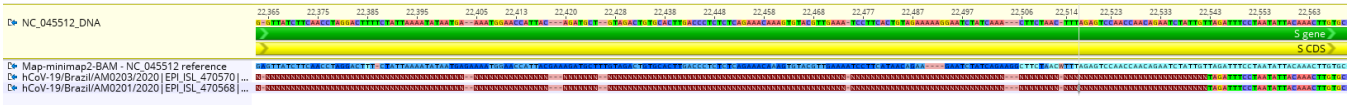


(B) NB1 3’ end.

Figure 9. Terminal portions of Ns Block 1 (NB1) of the alignment of the ERR4368102_MN908947.3.fastq assembled consensus sequence was exported as fastq file and aligned with EPI_ISL_470568, EPI_ISL_470570 and NC_045512. In (A) the 5’ end and in (B) the 3’ end. NB1 is located at 20,543 and 20,813 (numbering of reference NC_045512).



(A) NB2 5’ end.



(B) NB2 3’ end.

Figure 10. Terminal portions of Ns Block 2 (NB2) of the alignment of the ERR4368102_MN908947.3.fastq assembled consensus sequence was exported as fastq file and aligned with EPI_ISL_470568, EPI_ISL_470570 and NC_045512. In (A) the 5’ end and in (B) the 3’ end. NB2 is located between 22,325 and 22,542 (NC 045512 reference numbering).

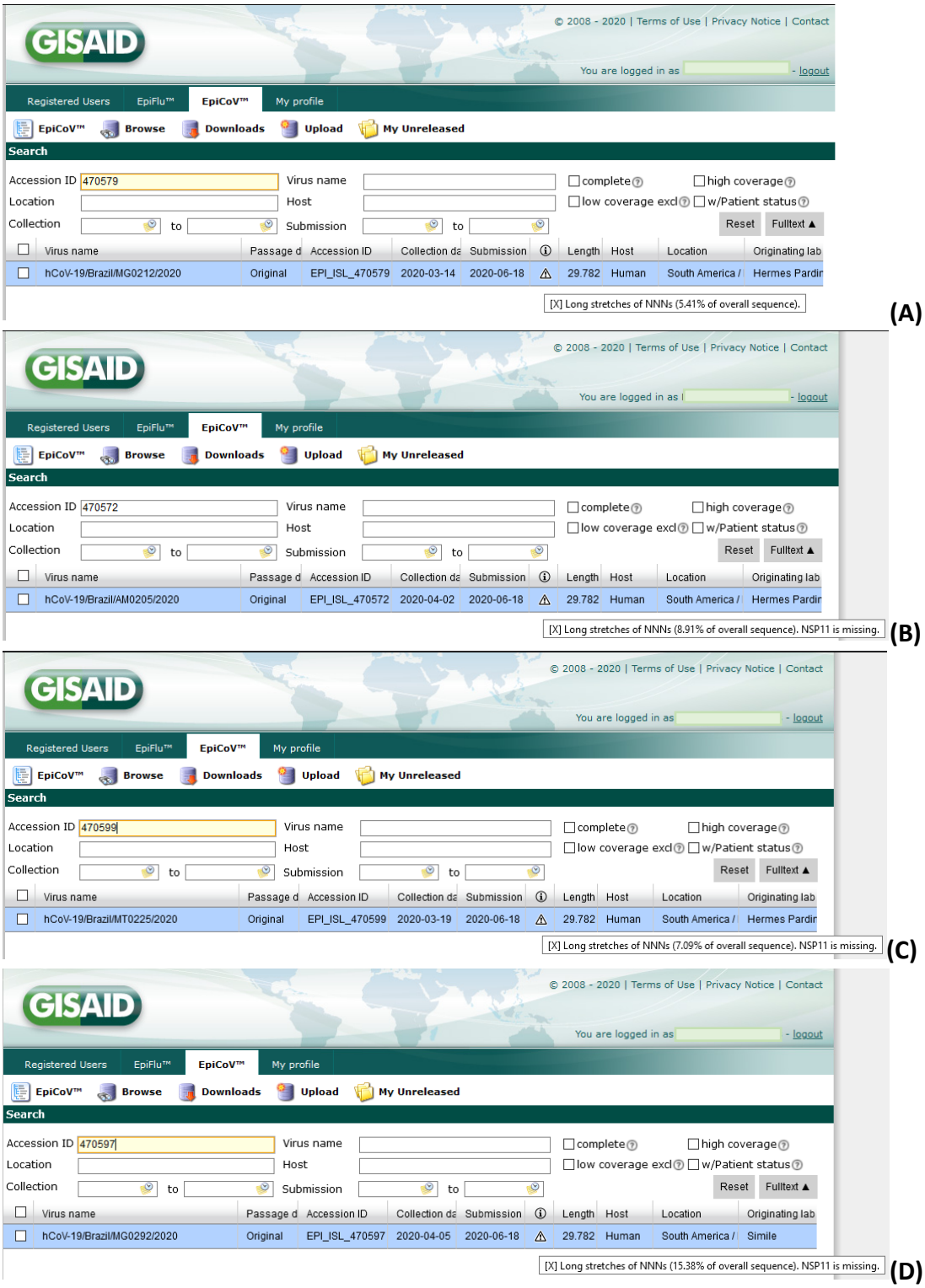


Figure 11. GISAID site showing examples of four sequences with very high contents of Ns used in the study. From 5.41% to 15.38% and missing gene NSP11 (e.g. EPI_ISL_470579, 470572, 470599 and 470597).