Original Paper

The Easy-to-Use SARS-CoV-2 Assembler for Genome Sequencing: Development Study

Martina Rueca[1], MSc; Emanuela Giombini[1], PhD; Francesco Messina[2], PhD; Barbara Bartolini[2], PhD; Antonino Di Caro[2,3], MSc; Maria Rosaria Capobianchi[1,3], MSc; Cesare EM Gruber[1], PhD

[1]Laboratory of Virology and Biosafety Laboratories, National Institute for Infectious Diseases "Lazzaro Spallanzani", Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy

[2]Laboratory of Microbiology and Biological Bank, National Institute for Infectious Diseases "Lazzaro Spallanzani", Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy

[3]UniCamillus - Saint Camillus International University of Health Sciences, Roma, Italy

Corresponding Author:
Francesco Messina, PhD
Laboratory of Microbiology and Biological Bank
National Institute for Infectious Diseases "Lazzaro Spallanzani"
Istituto di Ricovero e Cura a Carattere Scientifico
Via Portuense 292
Rome, 00149
Italy
Phone: 39 0655170668
Email: francesco.messina@inmi.it

Abstract

Background: Early sequencing and quick analysis of the SARS-CoV-2 genome have contributed to the understanding of the dynamics of COVID-19 epidemics and in designing countermeasures at a global level.

Objective: Amplicon-based next-generation sequencing (NGS) methods are widely used to sequence the SARS-CoV-2 genome and to identify novel variants that are emerging in rapid succession as well as harboring multiple deletions and amino acid–changing mutations.

Methods: To facilitate the analysis of NGS sequencing data obtained from amplicon-based sequencing methods, here, we propose an easy-to-use SARS-CoV-2 genome assembler: the Easy-to-use SARS-CoV-2 Assembler (ESCA) pipeline.

Results: Our results have shown that ESCA could perform high-quality genome assembly from Ion Torrent and Illumina raw data and help the user in easily correct low-coverage regions. Moreover, ESCA includes the possibility of comparing assembled genomes of multisample runs through an easy table format.

Conclusions: In conclusion, ESCA automatically furnished a variant table output file, fundamental to rapidly recognizing variants of interest. Our pipeline could be a useful method for obtaining a complete, rapid, and accurate analysis even with minimal knowledge in bioinformatics.

(JMIR Bioinform Biotech 2022;3(1):e31536) doi: 10.2196/31536

KEYWORDS

SARS-CoV-2 genome; bioinformatics tool; NGS data analysis; COVID-19; genome; health informatics; bioinformatic; digital tools; algorithms

Introduction

Next-generation sequencing (NGS) has reached a pivotal role in the field of emerging infectious diseases by enhancing the development capacity of new diagnostic methods, vaccines, and drugs [1,2]. Moreover, a key role has been recognized for sequence data production and sharing in outbreak response and management [3-5]. In the current COVID-19 epidemic, more than 6 million full genome sequences of SARS-CoV-2 have been deposited in publicly accessible databases in the arc of 1 year (ie, GISAID) [6,7]. SARS-CoV-2 genome surveillance on a global scale is permitting real-time analysis of the outbreak, with a direct impact on the public health response. This contribution includes the tracing of SARS-CoV-2 spread over

time and space, evidence of emerging variants that may affect pathogenicity, transmission capacity, diagnostic methods, therapeutics, or vaccines [8-11]. Recently divergent SARS-CoV-2 variants are emerging in rapid succession, harboring multiple deletions and amino acid mutations. Some mutations occur in the receptor-binding domain of the spike protein and are associated with an increase of angiotensin-converting enzyme 2 (ACE2) affinity as well as a potential reduction of polyclonal human plasma antibody efficacy [12,13]. The growing contribution of sequence information to public health is driving global investment in sequencing facilities and scientific programs [14,15]. The falling cost of generating genomic NGS data provides new chances for sequencing capacity expansion; however, many laboratories have low sequencing capacity and even a lack of expertise for data elaboration.

While sequencing runs can be performed without consolidated experience in the infectious disease field, virus genomic sequence assembly is often a demanding task. Translating SARS-CoV-2 raw read data into reliable and informative results is complex and requires solid bioinformatics knowledge, particularly for low-coverage samples. Some steps can lead to incorrect variant calling and produce erroneous assembled sequences.

Supervision of the sequence assembly to avoid inconsistent or misleading assignment of a virus to a taxonomic lineage or clade [9,10] as well as evaluation of low-coverage samples to prevent loss of epidemiological information are mandatory.

Many tools have been developed to support whole genome sequence reconstruction, starting with reads produced by different NGS platforms. However, most tools have been designed for genome assembly of other viruses and often are able to elaborate only a specific type of data. Some of these tools, for example, have implemented the assembly method for one specific platform (ie, Loretta for PacBio data) [16] or for a specific sequencing approach (ie, UNAGI for Nanopore and Illumina data) [17]. Some sequencing platform manufacturers have proposed pipelines for SARS-CoV-2 genome reconstruction that have been designed to obtain the most accurate sequence from one specific technological output. For example, Illumina developed the DRAGEN tool for SARS-CoV-2 genome analysis, a commercial tool that is temporarily free and available online, while Ion Torrent suggests the iterative refinement meta-assembler (IRMA) for SARS-CoV-2 data analyses, an open-source program developed by the Centers for Disease Control and Prevention (CDC) [18].

We propose the Easy-to-use SARS-CoV-2 Assembler (ESCA) pipeline: a novel reference-based genome assembly pipeline specifically designed for SARS-CoV-2 data analysis. This pipeline was created to support laboratories with limited experience in bioinformatics for SARS-CoV-2 analysis. ESCA can be easily installed and runs in most Linux environments.

Methods

Overview

The ESCA pipeline is a reference-based assembly algorithm written for Linux environments and requires only raw reads as input files, without any other information. Two versions of the software are available: one for Illumina paired-end reads in the “fastq.gz” file format and the other for Ion Torrent reads in the “ubam” file format.

The software is designed to process several samples in a single run. All reads (paired or unpaired) must be copied into the same working directory, and then, the program is launched through the command line by typing “StartEasyTorrent” for IonTorrent input or “StartEasyIllumina” for Illumina input. The pipeline than performs all the other passages automatically, as described in the following paragraphs.

The program processes all input reads, dividing them into different samples using file names as identifiers. Illumina paired-end reads are expected to be divided into 2 files that contain “R1” or “R2” to distinguish forward reads from reverse reads.

Sample preprocessing is performed by filtering out all reads with a mean Phred quality score lower than 20 and that are less than 30 nucleotides long.

Filtered reads are mapped on the SARS-CoV-2 reference genome Wuhan-Hu-1 (GenBank Accession Number NC_045512.2) with bwa-mem software [15]; all reads that do not map on the reference genome are then discarded.

Genome coverage is then analyzed: The read-mapping file is converted into “sorted-bam” and “mpileup” files using samtools software [19], and these data are translated into a detailed coverage table that reports the count of nucleotides observed at each position.

The consensus sequence is then reconstructed on the basis of 3 parameters: (1) frequency of nucleotides observed at each position, (2) nucleotide coverage, (3) reference genome sequence.

Briefly, sample parameters for consensus sequence reconstruction are designed to call the nucleotide observed with >50% frequency and with a coverage of >50 reads, but the minimum coverage is reduced at >10 reads if the most frequent nucleotide observed is identical to the nucleotide observed in the reference genome.

For all positions where these parameters are not satisfied, the ESCA pipeline is designed to call “N” to indicate a low coverage position or an intrasample nucleotide variant.

After whole genome reconstruction of all samples, the consensus sequences are aligned with the Wuhan-Hu-1 reference genome using MAFFT software [20], and a mutation table is generated, reporting nucleotide mutations of all the genomes assembled.

Illumina Data

To test the efficiency of the ESCA pipeline, 228 SARS-CoV-2–positive samples were sequenced with Illumina

platforms using the Ion AmpliSeq SARS-CoV-2 Research Panel following the manufacturer's instructions (ThermoFisher, Waltham, MA). For Illumina samples, whole SARS-CoV-2 genome sequences were assembled using both ESCA and DRAGEN RNA Pathogen Detection v.3.5.15 (BaseSpace) with default parameters.

Ion Torrent Data

A resequencing assay on Ion Torrent platforms was carried out for the same 228 SARS-CoV-2–positive samples using the Ion AmpliSeq SARS-CoV-2 Research Panel following the manufacturer's instructions (ThermoFisher).

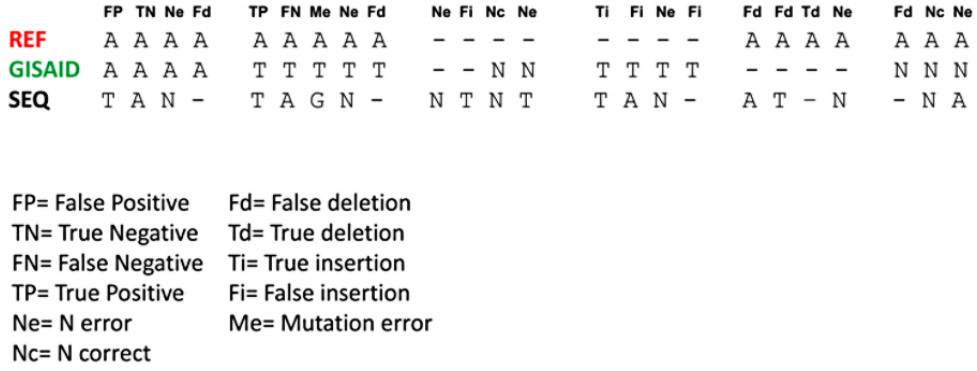
For Ion Torrent samples, whole SARS-CoV-2 genome sequences were assembled with ESCA and IRMA software [18] using the setting parameters indicated by ThermoFisher, in order to test the consistency of ESCA and IRMA outputs.

Performance Test

The respective results were compared, aligning the sequences obtained using the 2 methods with the reference sequence Wuhan-Hu-1 (NCBI Acc. Numb. NC_045512.2), and the corrected sequence was submitted to GISAID, using MAFFT [20]. Then, each discordant position was evaluated following the classification reported in Figure 1. In particular, we evaluated true positives (TP; mutations correctly classified as real); false negatives (FN; mutations correctly classified as unreal); false positives (FP; mutations incorrectly classified as real); true negatives (TN; mutations correctly classified as unreal); corrected TN (positions unknown correctly classified as N); and TN error (positions unknown, incorrectly classified as N).

To test the performances with respect to mean coverage, linear regression correlation analysis was carried out for mean coverage and specific measures of accuracy.

Figure 1. Classification scheme for genome assemblers, in which assembled genome sequences (SEQ) were compared with the corresponding submitted sequences (on GISAID) and with reference genome sequence "Wuhan-Hu-1" (REF). Nucleotide threesomes were classified using the following 11 categories: false deletion (Fd), false insertion (Fi), false negative (FN), false positive (FP), mutation error (Me), N correct (Nc), N error (Ne), true deletion (Td), true insertion (Ti), true negative (TN), true positive (TP).



Results

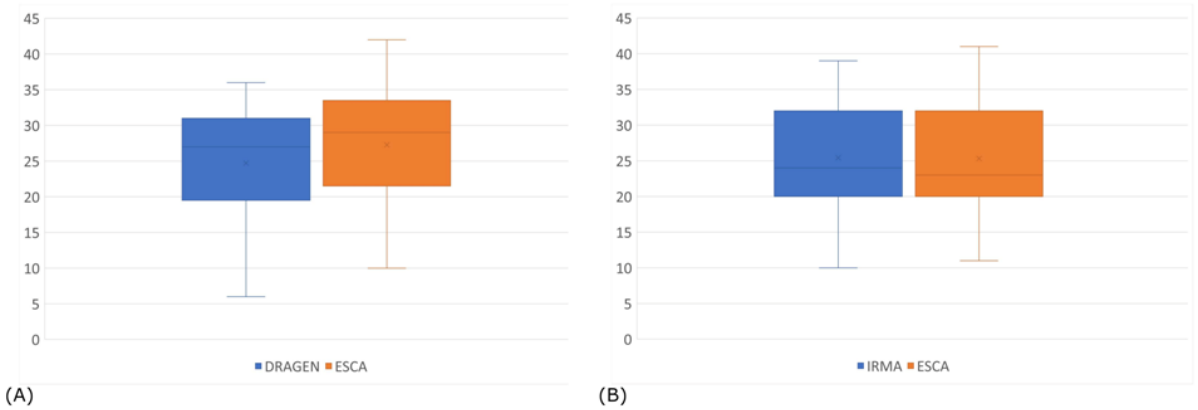
In the computational evaluation, ESCA software was compared with the most often used assemblers for SARS-CoV-2 genome analysis on 228 SARS-CoV-2–positive samples.

Sequencing was performed on Illumina MiSeq for 65 libraries, obtaining a median of 1.50 x 106 paired-end reads per sample (range: 0.02 x 106 to 4.56 x 106), and on Ion Gene Studio S5 Sequencer for 163 libraries, obtaining a median of 0.61 x 106 single-end reads per sample (range: 0.02 x 106 to 3.02 x 106). Using the ESCA reconstruction, the coverage point by point was calculated, and we observed that, in the Illumina sample, the point coverage was not uniform, although the mean coverage was quite high in all samples (average 3508X; range: 70-10,733). This could introduce error in genome reconstruction using some software. In this context, ESCA could reduce the error in regions with low coverage. In parallel, mean coverage obtained with Ion Torrent was 4966X (range: 94-19,917), but a higher uniformity was observed. The comparison of the coverage distribution is shown in Figure 2.

To evaluate the ESCA and DRAGEN/IRMA results, assembled genomes, the reference Wuhan-Hu-1, and the corrected genome of GISAID (Accession IDs available in Multimedia Appendix 1) were aligned with MAFFT [20].

At each position along the SARS-CoV-2 genome, the 24 available nucleotide combinations were classified in 11 mutation categories (Figure 1). For all sequences, the number of occurrences of mutation categories for each assembly software was then evaluated.

Figure 2. Comparison of true positive mutations between our Easy-to-use SARS-CoV-2 Assembler (ESCA) and the (A) Illumina DRAGEN tool and (B) iterative refinement meta-assembler (IRMA) recommended by Ion Torrent.



Illumina Data

The comparison of ESCA with DRAGEN showed that, as expected, the mean number of mutations in genomes was very low (in the mean 28 position) and ESCA could correctly identify a mean 27 of 28 mutations (Figure 2A). Moreover, no FN positions were identified by ESCA. This is due to the pipeline design that reduces the error of introducing N where the coverage is not sufficient. The DRAGEN genome, instead, showed a mean of 25 of 28 TP and 3 FN positions. The absence of a mutation in specific positions could be essential to assigning the lineage, and the presence of FNs could modify the identification of the variants.

On the other hand, both ESCA and DRAGEN did not introduce FN, identifying 29,308 and 28,027 TN positions, respectively.

These results show an accuracy of 100% for ESCA and 99.99% for DRAGEN. Moreover, the sensitivity of ESCA compared with DRAGEN was 96.43% for ESCA and 89.29% for DRAGEN, and the specificity with both methods was 100%.

Ion Torrent Data

Parallel to the previous comparison, ESCA compared with IRMA showed that both methods identified a mean 25 of 26 TP positions (Figure 2B) but did not induce FN. However, IRMA introduced a certain number of errors. In fact, the FP was 20 for IRMA, compared with 0 for ESCA. Once again, the introduction of mutations could induce error in the lineage assignation.

The accuracy of IRMA was calculated to be 99.93%, while it was 100% for ESCA.

Moreover, although the sensitivity was identical with the 2 methods (96.15%), the specificity was 99.93% for IRMA and 100% for ESCA.

Performance Test

To evaluate the performance of each of the methods, linear regression correlation analysis was carried out with respect to mean coverage (Multimedia Appendix 2).

For IonTorrent single-end sequencing data, a significant positive correlation was found comparing coverage and TN for both IRMA and ESCA ($r>0.15$, $P<.05$), while for Illumina pair-end sequencing data, such a correlation was found only for DRAGEN ($r>0.40$, $P<.05$). This difference could be caused by a different error rate for the 2 sequencing techniques. These data suggest that all assembly methods are comparable in the case of high coverage samples, while ESCA seems to perform better for low coverage data.

Discussion

Principal Findings

The importance of rapidly obtaining and sharing high-quality whole genomes of SARS-CoV-2 is increasing with the emerging variant strains [14]. For this reason, the use of NGS custom amplicon panels can be a rapid and performant method for identifying viral variants. However, a lack of bioinformatic skills could be a problem in handling NGS raw data. Our pipeline ESCA provides help to laboratories with low bioinformatic capacity using a single command. Both of the more common methods for the analysis of Ion Torrent and Illumina data (IRMA and DRAGEN, respectively) have shown a certain amount of error that could induce false identification in variant assignation. On the contrary, the SARS-CoV-2 genome obtained by ESCA shows a reduced number of false insertions and false mutations and a higher number of real mutations.

Limitations

This pipeline should be tested on a larger number of sequences and with other sequencing technologies.

Conclusions

ESCA automatically produces a variant table output file, fundamental for rapidly recognizing variants of interest.

These results show how ESCA could be a useful method for obtaining a rapid, complete, and correct analysis even with minimal skill in bioinformatics.

Acknowledgments

We gratefully acknowledge the contributors of the genome sequences of the newly emerging coronavirus (ie, the originating laboratories) for sharing sequences and other metadata through the GISAID Initiative, on which this research is based. We also acknowledge Ornella Butera, Francesco Santini, and Giulia Bonfiglio for their contribution to sample preparation. National Institute for Infectious Diseases Lazzaro Spallanzani IRCCS received financial support from the Italian Ministry of Health grants “Ricerca Corrente” (Progetto 1-2763705) and “5 PER MILLE 2020” (iSNV study for early detection and risk analysis of SARS-CoV-2 mutations with impact on clinical management of COVID-19) research funds.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comparison of Easy-to-use SARS-CoV-2 Assembler (ESCA) and (first tab: Ion Torrent) iterative refinement meta-assembler (IRMA) and (second tab: Illumina) DRAGEN results for every assembled SARS-CoV-2 genome. DELerror: deletion unknown incorrectly identified; errorMut: number of incorrected mutations described; FN: false negative; FP: false positive; INScorr: insertion unknown correctly identified; INSerror: insertion unknown incorrectly identified; NCorr: positions unknown correctly classified as N; Nerror: position unknown incorrectly classified as N; TN: true negative; TP: true positive.
[XLSX File (Microsoft Excel File), 36 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Pairwise linear regression correlation analysis (shown with the *P* value) for every parameter: DELerror: deletion unknown incorrectly identified; FN: false negative; FP: false positive; INScorr: insertion unknown correctly identified; INSerror: insertion unknown incorrectly identified; NCorr: positions unknown correctly classified as N; Nerror: position unknown incorrectly classified as N; TN: true negative; TP: true positive.
[XLSX File (Microsoft Excel File), 41 KB-Multimedia Appendix 2]

References

1. World Health Organization (WHO). Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health. World Health Organization. 2021 Jan 08. URL: https://www.who.int/publications/i/item/9789240018440 [accessed 2022-02-22]
2. Greaney AJ, Loes AN, Crawford KH, Starr TN, Malone KD, Chu HY, et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe 2021 Mar 10;29(3):463-476.e6 [FREE Full text] [doi: 10.1016/j.chom.2021.02.003] [Medline: 33592168]
3. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 2009 Jun 25;459(7250):1122-1125. [doi: 10.1038/nature08182] [Medline: 19516283]
4. Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ, ECDC National Microbiology Focal Points and Experts Group. Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of European National Capacities, 2015-2016. Front Public Health 2017;5:347 [FREE Full text] [doi: 10.3389/fpubh.2017.00347] [Medline: 29326921]
5. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, FWD-NEXT Expert Panel. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill 2017 Jun 08;22(23):1 [FREE Full text] [doi: 10.2807/1560-7917.ES.2017.22.23.30544] [Medline: 28662764]
6. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 2017 Jan 10;1(1):33-46 [FREE Full text] [doi: 10.1002/gch2.1018] [Medline: 31565258]
7. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2018 Jan 04;46(D1):D8-D13 [FREE Full text] [doi: 10.1093/nar/gkx1095] [Medline: 29140470]
8. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [FREE Full text] [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
9. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018 Dec 01;34(23):4121-4123 [FREE Full text] [doi: 10.1093/bioinformatics/bty407] [Medline: 29790939]
10. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020 Nov;5(11):1403-1407 [FREE Full text] [doi: 10.1038/s41564-020-0770-5] [Medline: 32669681]
11. Outbreak.info. URL: https://outbreak.info/ [accessed 2022-02-22]

12. Cameroni E, Bowen JE, Rosen LE, Saliba C, Zepeda SK, Culap K, et al. Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. Nature 2022 Feb 10;602(7898):664-670 [FREE Full text] [doi: 10.1038/s41586-021-04386-2] [Medline: 35016195]
13. Focosi D, Maggi F. Neutralising antibody escape of SARS-CoV-2 spike protein: Risk assessment for antibody-based Covid-19 therapeutics and vaccines. Rev Med Virol 2021 Nov;31(6):e2231 [FREE Full text] [doi: 10.1002/rmv.2231] [Medline: 33724631]
14. SARS-CoV-2 genomic sequencing for public health goals: Interim guidance, 8 January 2021. World Health Organization. 2021 Jan 08. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-genomic_sequencing-2021.1 [accessed 2022-02-22]
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009 Jul 15;25(14):1754-1760 [FREE Full text] [doi: 10.1093/bioinformatics/btp324] [Medline: 19451168]
16. Al Qaffas A, Nichols J, Davison AJ, Ourahmane A, Hertel L, McVoy MA, et al. LoReTTA, a user-friendly tool for assembling viral genomes from PacBio sequence data. Virus Evol 2021 Jan;7(1):veab042 [FREE Full text] [doi: 10.1093/ve/veab042] [Medline: 33996146]
17. Al Kadi M, Jung N, Ito S, Kameoka S, Hishida T, Motooka D, et al. UNAGI: an automated pipeline for nanopore full-length cDNA sequencing uncovers novel transcripts and isoforms in yeast. Funct Integr Genomics 2020 Jul 18;20(4):523-536 [FREE Full text] [doi: 10.1007/s10142-020-00732-1] [Medline: 31955296]
18. Shepard SS, Meno S, Bahl J, Wilson MM, Barnes J, Neuhaus E. Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. BMC Genomics 2016 Sep 05;17:708 [FREE Full text] [doi: 10.1186/s12864-016-3030-6] [Medline: 27595578]
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009 Aug 15;25(16):2078-2079 [FREE Full text] [doi: 10.1093/bioinformatics/btp352] [Medline: 19505943]
20. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002 Jul 15;30(14):3059-3066 [FREE Full text] [doi: 10.1093/nar/gkf436] [Medline: 12136088]

Abbreviations

ACE2: angiotensin-converting enzyme 2
CDC: Centers for Disease Control and Prevention
ESCA: Easy-to-use SARS-CoV-2 Assembler
FN: false negative
FP: false positive
IRMA: iterative refinement meta-assembler
NGS: next-generation sequencing
TN: true negative
TP: true positive

Edited by A Mavragani; submitted 24.06.21; peer-reviewed by S Tausch, Y Miao; comments to author 30.07.21; revised version received 02.11.21; accepted 05.02.22; published 14.03.22

Please cite as:
Rueca M, Giombini E, Messina F, Bartolini B, Di Caro A, Capobianchi MR, Gruber CEM
The Easy-to-Use SARS-CoV-2 Assembler for Genome Sequencing: Development Study
JMIR Bioinform Biotech 2022;3(1):e31536
URL: https://bioinform.jmir.org/2022/1/e31536
doi: 10.2196/31536
PMID: