ZOOLOGICAL RESEARCH

# An online coronavirus analysis platform from the National Genomics Data Center

**DEAR EDITOR,**

Since the first reported severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection in December 2019, coronavirus disease 2019 (COVID-19) has become a global pandemic, spreading to more than 200 countries and regions worldwide. With continued research progress and virus detection, SARS-CoV-2 genomes and sequencing data have been reported and accumulated at an unprecedented rate. To meet the need for fast analysis of these genome sequences, the National Genomics Data Center (NGDC) of the China National Center for Bioinformation (CNCB) has established an online coronavirus analysis platform, which includes *de novo* assembly, BLAST alignment, genome annotation, variant identification, and variant annotation modules. The online analysis platform can be freely accessed at the 2019 Novel Coronavirus Resource (2019nCoVR) (https://bigd.big.ac.cn/ncov/online/tools).

As of 1 October 2020, the Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org/) (Shu & McCauley, 2017) contained 131 424 SARS-CoV-2 sequences, the 2019 Novel Coronavirus Resource (2019nCoVR) (Song et al., 2020; Zhao et al., 2020) contained 135 979 genome sequences, and the National Center for Biotechnology Information (NCBI) (Leinonen et al., 2011) contained 61 551 high-throughput sequencing runs. In addition, the Genome Sequence Archive (GSA) (Wang et al., 2017) has also released more than 200 accessions of SARS-CoV-2 sequencing runs. These data provide important information for SARS-CoV-2-based studies on viral classification, viral tracing, viral mutations, genome evolution, and antiviral drug development. Thus, there is an urgent need for a comprehensive online analysis platform to deal with the massive amount of data available.

To promote studies and applications based on SARS-CoV-2 sequencing data, specific sequence analysis tools have been established in several online platforms worldwide. For example, NCBI has provided the BLAST alignment tool (Altschul et al., 1990) in SARS-CoV-2 Resources (https://www.ncbi.nlm.nih.gov/sars-cov-2/). The University of California, Santa Cruz (UCSC) SARS-CoV-2 Genome Browser has integrated the visualization browser with BLAT alignment and variant annotation tools (https://genome.ucsc.edu/covid19.html) (Fernandes et al., 2020). The National Microbiology Data Center (NMDC) has provided various analysis tools, such as BLAST alignment and phylogenetic analysis, in the Global Coronavirus Data Sharing and Analysis System (http://nmdc.cn/coronavirus/). The Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences (CAS), has established the Virus Identification Cloud (VIC, https://www.biosino.org/vic/), offering online analysis services for viral sequence identification and genome assembly. The Genome Detective webserver has also provided a virus identification workflow for high-throughput sequencing data (https://www.genomedetective.com/) (Cleemput et al., 2020). Although the above SARS-CoV-2 analysis tools provide online services, their functions are relatively limited and do not cover all aspects of SARS-CoV-2 research (Table 1).

Thus, to provide a unified and convenient approach for processing SARS-CoV-2 sequencing data, the National Genomics Data Center (NGDC) of the China National Center for Bioinformation (CNCB) established an online coronavirus analysis platform based on viral genomes collected in 2019nCoVR (https://bigd.big.ac.cn/ncov/online/tools), offering free analysis services for researchers. The platform includes five functional modules (Figure 1), which cover various SARS-CoV-2 genomic data analyses.

1. *De novo* assembly module

This module can be used for *de novo* assembly of next-generation sequencing (NGS) data. First, raw reads are trimmed for quality using Trimmomatic (Bolger et al., 2014)



Received: 17 September 2020; Accepted: 12 October 2020; Online: 12 October 2020

Foundation items: This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB38030200，XDB38050300, XDA19090116, XDA19050302) and National Key R&D Program of China (2020YFC0848900, 2020YFC0847000)

DOI: 10.24272/j.issn.2095-8137.2020.065

**Table 1 Analysis function comparison of SARS-CoV-2 online resources**

| Functions or features | | 2019nCoVR | NCBI SARS-CoV-2 resources | UCSC SARS-CoV-2 browser | Genome detective | NMDC | VIC |
|---|---|---|---|---|---|---|---|
| Genome sequences | Sequence comparison | √ | √ | √ | | √ | |
| | Gene annotation | √ | | | | √ | |
| | Variant identification | √ | | | √ | | |
| | Phylogenetic tree | √ | | | √ | √ | |
| NGS raw reads | *De novo* assembly | √ | | | √ | | √ |
| | Variant identification | √ | | | | | |
| | Variant annotation | √ | | √ | | | |
| Open access | Do not need login | √ | √ | √ | √ | √ | |

*: 2019nCoVR: 2019 Novel Coronavirus Resource; NCBI: National Center for Biotechnology Information; UCSC: University of California, Santa Cruz; NMDC: National Microbiology Data Center; VIC: Virus Identification Cloud.

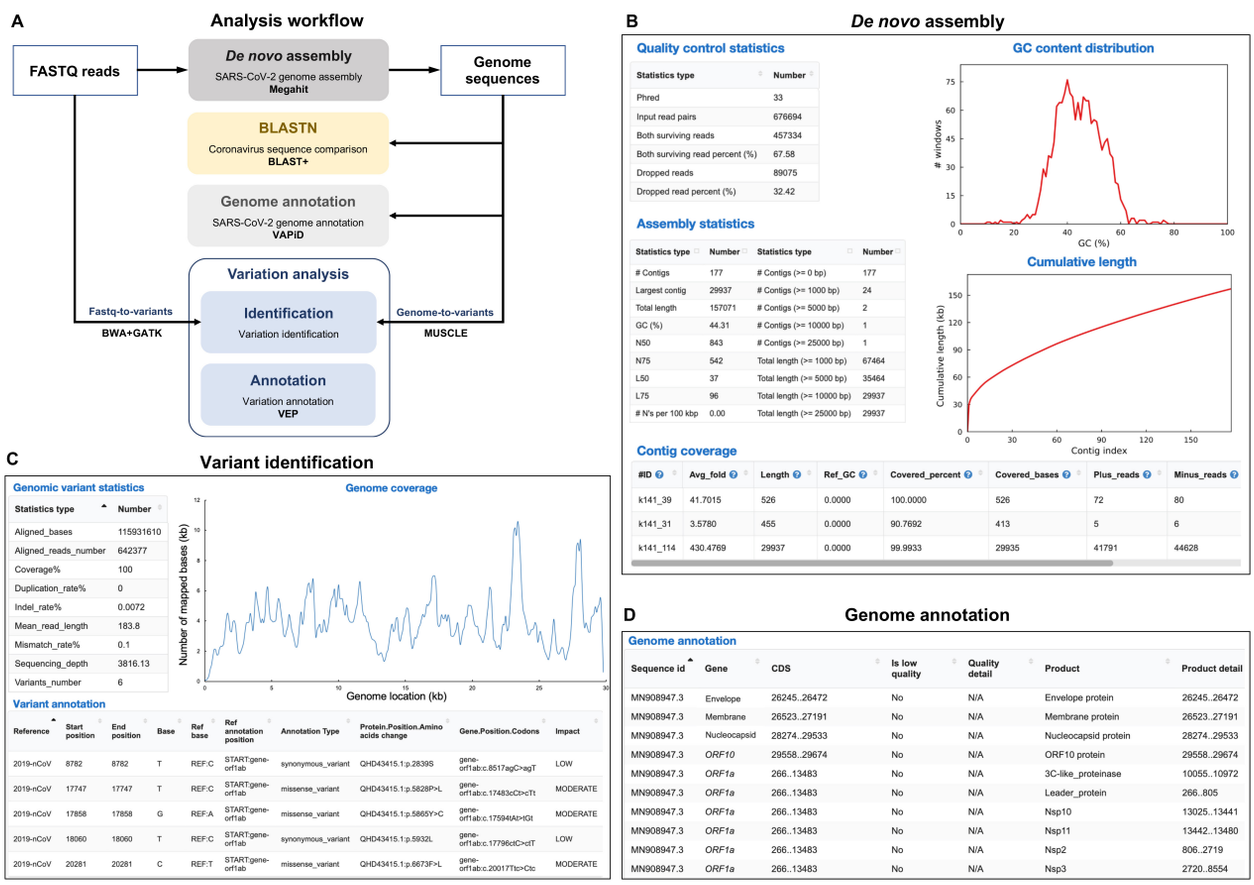


**Figure 1 Processing workflow and webpage demonstration of analysis results**

A: Analysis modules are in the middle of the figure. Main software used in the workflow is shown beside each module. B–D: Analysis demonstration of *de novo* assembly, variant identification, and genome annotation modules. N/A: Not available.

with the settings SLIDINGWINDOW: 4:15, LEADING: 3, TRAILING: 3 and MINLEN: 36. Megahit (Li et al., 2015) is then used for sequence assembly with default parameters. The assembled sequences are compared with the SARS-CoV-2 reference genome (NC_045512.2) using BLASTN (Altschul et al., 1990) to identify target sequence(s), and assembly quality is evaluated using QUAST (Gurevich et al., 2013). The assembly results depend on the qualities of samples and sequencing data and may consist of a complete genome or several contigs. In the future, we plan to assemble those contigs into a single sequence by alignment with the reference genome, and to support genome assembly for third-generation sequencing data.

2. BLAST module

To compare sequences among virus strains, the analysis platform includes a BLAST alignment module, with three

**Table 2 Reference running time**

| | Data1 | Data2 | Data3 | Data4 | Data5 |
|---|---|---|---|---|---|
| NCBI accession No. | SRR11247077 | SRR11092064 | SRR11092057 | SRR11092058 | SRR10971381 |
| Calculation time* | 0 m 37 s | 0 m 55 s | 1 m 10 s | 1 m 36 s | 3 m 42 s |
| Data size (bp) | 118 M | 1.0 G | 1.5 G | 2.2 G | 8.0 G |

*: Run on 24 CPU cores.

algorithms (BLASTN, Mega BLAST and discontinuous Mega BLAST) (Altschul et al., 1990). Users can select the SARS-CoV-2 reference genome, 2019nCoVR genome database, or coronavirus genome database (including alpha/beta/delta/gamma genus) for online BLAST.

3. Genome annotation module

To perform sequence comparison and evolutionary analysis on specific viral genes, gene annotations are required. However, most viral genomes in the above SARS-CoV-2 databases are not annotated. Therefore, we built a genome annotation module based on VAPiD (Shean et al., 2019), which can identify coding sequences (CDS) or protein sequences and generate a GenBank annotation file.

4. Variant identification modules

The variant identification function consists of the Genome-to-Variants and Fastq-to-Variants modules. Both modules use the genome NC_045512.2 as a default reference, but users can customize the reference by uploading a genome file. Genome-to-Variants can detect mutation sites from complete or partial genomes, using Muscle (Edgar, 2004) for sequence alignment. Fastq-to-Variants can identify genome variants from NGS raw data and connect seamlessly to the GSA system to load massive raw sequencing data to the server automatically. Sequencing reads are aligned to the SARS-CoV-2 reference genome (NC_045512.2) using BWA (Li & Durbin, 2009), after which Picard is used to remove duplicate reads and calculate aligned read number, error rate, sequencing depth, and genome coverage (http://broadinstitute.github.io/picard/). Single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) are identified using GATK (McKenna et al., 2010).

5. Variation annotation module

To clarify the mutation influence on gene function, the variation annotation module integrates the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) to show codon and amino acid changes, and then calculates the degree of function influence.

It is worth mentioning that the parameters for the data analysis modules have been highly optimized to improve efficiency and reduce computing time. For example, when testing the running time with the Fastq-to-Variants module using one 24-core server, it cost ~1 min to process 1 Gb of NGS data and less than 4 min for handling 8 Gb of NGS data (Table 2). For this online platform, we established five servers to provide public service, which indicates that the platform has the capacity to analyze 7 200 NGS data in one day if the data size is less than 1 Gb. In general, a notification email will be automatically sent to users when computing jobs are finished.

For future applications, we will continue to improve this specialized online platform by integrating more tools, software, and pipelines for SARS-CoV-2 data analysis and provide one-click and public data analysis services for coronavirus researchers.

## COMPETING INTERESTS

The authors declare that they have no competing interests.

## AUTHORS' CONTRIBUTIONS

W.M.Z., Y.M.B., Y.B.X., J.F.X., and Z.Z. designed the research. Z.L.D., Z.G., L.N.M., S.J., S.H.S., M.L.C., and C.P.L. implemented the analysis modules. J.W.Z., B.X.T., D.Z., and Y.B.S. built the web server. Z.L.D. and Z.G. wrote the manuscript. Y.B.X., W.M.Z., and Z.L.D. revised the manuscript. All authors read and approved the final version of the manuscript.

Zheng Gong[1,3,#], Jun-Wei Zhu[1,2,#], Cui-Ping Li[1,2,#], Shuai Jiang[1,2,#], Li-Na Ma[1,2], Bi-Xia Tang[1,2], Dong Zou[1,2], Mei-Li Chen[1,2], Yu-Bin Sun[1,2], Shu-Hui Song[1,2], Zhang Zhang[1,2,3], Jing-Fa Xiao[1,2,3], Yong-Biao Xue[1,2,3], Yi-Ming Bao[1,2,3], Zheng-Lin Du[1,2,*], Wen-Ming Zhao[1,2,3,*]

[1] *China National Center for Bioinformation & National Genomics Data Center*, *Beijing Institute of Genomics*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*

[2] *CAS Key Laboratory of Genome Sciences and Information*, *Beijing Institute of Genomics*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*

[3] *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*

[#]Authors contributed equally to this work

*Corresponding authors, E-mail: duzhl@big.ac.cn; zhaowm@big.ac.cn

## REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3): 403−410.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15): 2114−2120.

Cleemput S, Dumon W, Fonseca V, Abdool Karim W, Giovanetti M, Alcantara LC, et al. 2020. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics*, **36**(11): 3552−3555.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with

reduced time and space complexity. *BMC Bioinformatics*, **5**(1): 113.
Fernandes JD, Hinrichs AS, Clawson H, Gonzalez JN, Lee BT, Nassar LR, et al. 2020. The UCSC SARS-CoV-2 genome browser. *Nature Genetics*, **52**: 991−998.
Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**(8): 1072−1075.
Leinonen R, Sugawara H, Shumway M. 2011. The sequence read archive. *Nucleic Acids Research*, **39**(S1): D19−D21.
Li DH, Liu CM, Luo RB, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**(10): 1674−1676.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14): 1754−1760.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9): 1297−1303.
McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. 2016. The ensembl variant effect predictor. *Genome Biology*, **17**(1): 122.
Shean RC, Makhsous N, Stoddard GD, Lin MJ, Greninger AL. 2019. VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank. *BMC Bioinformatics*, **20**(1): 48.
Shu YL, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance*, **22**(13): 30494.
Song SH, Ma L, Zou D, Tian DM, Li CP, Zhu JW, et al. 2020. The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. *bioRxiv*, doi: 10.1101/2020.08.30.273235.
Wang YQ, Song FH, Zhu JW, Zhang SS, Yang YD, Chen TT, et al. 2017. GSA: genome sequence archive. *Genomics, Proteomics & Bioinformatics*, **15**(1): 14−18.
Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK, et al. 2020. The 2019 novel coronavirus resource. *Hereditas (Beijing)*, **42**(2): 212−221. (in Chinese)