CoV2ID: Detection and Therapeutics Oligo Database for SARS-CoV-2

João Carneiro[1] and Filipe Pereira[2]

[1]Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Portugal

[2] IDENTIFICA, Science and Technology Park of the University of Porto - UPTEC, Porto, Portugal.

* E-mails: joaomiguelsov@gmail.com; fpereirapt@gmail.com

Abstract

The ability to detect the SARS-CoV-2 in a widespread epidemic is crucial for screening of carriers and for the success of quarantine efforts. Methods based on real-time reverse transcription polymerase chain reaction (RT-qPCR) and sequencing are being used for virus detection and characterization. However, RNA viruses are known for their high genetic diversity which poses a challenge for the design of efficient nucleic acid-based assays. The first SARS-CoV-2 genomic sequences already showed novel mutations, which may affect the efficiency of available screening tests leading to false-negative diagnosis or inefficient therapeutics. Here we describe the CoV2ID (http://covid.portugene.com/), a free database built to facilitate the evaluation of molecular methods for detection of SARS-CoV-2 and treatment of COVID-19. The database evaluates the available oligonucleotide sequences (PCR primers, RT-qPCR probes, etc.) considering the genetic diversity of the virus. Updated sequences alignments are used to constantly verify the theoretical efficiency of available testing methods. Detailed information on available detection protocols are also available to help laboratories implementing SARS-CoV-2 testing.

Keywords: COVID-19; oligonucleotides; coronavirus; false negatives; RT-qPCR

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first detected in December 2019 [1-3]. Phylogenetic data implicate a zoonotic origin in Wuhan, the capital of Central China's Hubei Province, from where the novel virus rapidly spread worldwide becoming a pandemic [4]. The SARS-CoV-2 belongs to the β-coronavirus genus of the *Coronaviridae* family, being related with other virus causing human infections such as SARS-CoV and MERS-CoV. The novel SARS-CoV-2 shares 80% of identity with SARS-CoV (the causing agent of the 2002-2003 SARS outbreak in Asia) and nearly 96% similar to the bat coronavirus isolate RaTG13, suggesting these animals are the likely natural reservoir of the virus [4-6].

The SARS-CoV-2 genome consists of a single, positive-stranded RNA with approximately 30 000 nucleotides. Several genomic sequences have been made available in public databases by researchers worldwide as the epidemic progresses. The great adaptability and infection capacity of RNA viruses depends in part from their high mutation rates [7]. As expected, available SARS-CoV-2 genomic sequences show a large number of new mutations. In April 2020, more than 2500 mutations have been reported in the 2019 Novel Coronavirus Resource (2019nCoVR) of the China National Center for Bioinformation (https://bigd.big.ac.cn/ncov/variation/annotation).

Many techniques use molecules that interact with the virus RNA genome or the reverse transcribed DNA, either for clinical testing, diagnosis or determination of viral loads. PCR primers and RT-qPCR probes are been used to detect the SARS-CoV-2 (e.g., [8-11]) using molecular biology techniques. It is likely that oligonucleotides complementary to the virus RNA will be tested as possible

antiviral agents [12, 13]. However, the SARS-CoV-2 genetic diversity can be a challenge for the efficiency of available assays since it may lead to false-negative results in detection tests or inefficient therapeutics. Polymorphisms at binding sites of PCR primers, RT-qPCR probes, small interfering RNAs may be a problem for available techniques. Here, we describe CoV2ID, a database whose objective is to help the scientific community to improve the testing and therapeutic capacity and efficiency.

Methods

Database features

The CoV2ID database (http://covid.portugene.com/) uses java graphics and dynamic tables and works with major web browsers (e.g. Internet Explorer, Mozilla Firefox, Chrome). The database provides descriptive webpages for each oligonucleotide and a search engine to access dynamic tables with numeric data and multiple sequence alignments. A SQLite local database is used for data storage and runs on an Apache web server. The dynamic HTML pages were implemented using CGI-Perl and JavaScript and the dataset tables using the JQuery plugin DataTables v1.9.4 (http://datatables.net/). Python and Perl in-house algorithms were written and used to perform identity and pairwise calculations.

Oligonucleotides

The oligonucleotides were retrieved from seven molecular assays to diagnose the SARS-CoV-2 provided by the World Health Organization (WHO) [14]. Future

updates of the database will include peer-reviewed protocols when available. Each oligonucleotide has a specific database code (for example, CoV2ID001). The CoV2ID database ranks oligonucleotides using three measures of sequence conservation:

a) *Percentage of identical sites* (PIS), calculated by dividing the number of equal positions in the alignment for an oligonucleotide by its length;

b) *Percentage of identical sites in the last five nucleotides at the 3' end of the oligonucleotide* (3'PIS) - the most critical regions for an efficient binding to the template DNA during PCR and

c) *Percentage of pairwise identity* (PPI), calculated by counting the average number of pairwise matches across the positions of the alignment, divided by the total number of pairwise comparisons.

The 'CoV2ID ranking score' considers the mean value of the three different measures (PIS, 3'PIS and PPI). Further details can be found in our previous publications for the Ebola [15] and HIV [16] databases.

Genomic sequences

The 'Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1' with accession number NC_045512.2 was selected as reference. Genomes were obtained from the GenBank (https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/) and the GISAID Initiative (https://www.gisaid.org/). The complete genomes were obtained for all know human coronaviruses: SARS-CoV-2, HCoV-OC43, HCoV-HKU1, HCoV-NL63, HCoV-229E, MERS-CoV and SARS-CoV. The list of acknowledgments to the original source of the data available at GISAID can be found in 'Acknowledgments' section of our database.

The first release of the database includes three multiple sequence alignments:

a) CoV2ID_alig01 - All human SARS-CoV-2 complete genomes from the NCBI Virus resource (www.ncbi.nlm.nih.gov/labs/virus/).

b) CoV2ID_alig02 - All human SARS-CoV-2 complete genomes with high coverage from the GISAID initiative (www.gisaid.org/).

c) CoV2ID_alig03 - Alignment of the consensus sequence of each human coronavirus. The consensus was obtained by aligning all complete available genomes for each virus obtained from the NCBI Virus resource (www.ncbi.nlm.nih.gov/labs/virus/).

The genomes from the NCBI Virus resource were aligned using an optimized version of MUSCLE running at the NCBI Variation Resource. The genomes from GISAID were aligned using the default parameters of the MAFFT version 7 [17]. The annotated reference of the SARS-CoV-2 genomes and the alignments can be visualized, edited and exported using the NCBI (https://www.ncbi.nlm.nih.gov/tools/sviewer/) and the Wasabi (http://wasabi2.biocenter.helsinki.fi/) sequence viewers.

Data analyses

The first release of the CoV2ID database (March 2020) includes 52 SARS-CoV-2 oligonucleotides (38 primers and 14 probes) retrieved from seven molecular assays to diagnose the SARS-CoV-2 provided by the World Health Organization (WHO) [14]. The oligonucleotides are located in the ORF1ab, S, ORF3a, E, M and N genes. The database provides an interface for browsing, filtering and downloading data from the different oligonucleotides annotated according to the SARS-CoV-2 reference genome. For each oligonucleotide, it is possible to find

information on the sequence, type of technique where it was originally used, location in the reference genome, etc.

The largest multiple sequence alignment (alig02) has currently 956 complete SARS-CoV-2 genomes. The alignment has a PIS of 64.90% and a PPI of 99.80%. The smaller NCBI alignment (currently with 106 genomes) has similar values (PIS of 62.40% and a PPI of 99.40%). These results demonstrate the existence of several mutated positions across the genome leading to relatively low percentage of identical sites (62 to 64%). However, the level of genetic diversity is relatively low, as shown by the high percentage of pairwise identity (>99%), suggesting that most mutations only occur in a few genomes, in line with other studies [18-20].

The database indicates which oligonucleotides bind to the most conserved regions of the SARS-CoV-2 using different measures of sequence conservation (Table 1). Our analyses revealed that oligonucleotides from different protocols have a perfect homology to all available genomes (CoV2ID score of 100%). For example, we identified two probes (*HKU-NP* and *Pasteur_nCoV_IP4-14084Probe*) that are 100% complementary to all genomes. The values will probably change as more sequences are added to the alignments, but these results are already a good indication of their sequence conservation.

On the contrary, some oligonucleotides have several mismatches to SARS-CoV-2 genomes. There are 16 oligonucleotides with a CoV2ID score of below 80%. For example, primers *NIID_WH-1_F24381* and *NIID_WH-1_Seq_F24383* have a CoV2ID score of below 50%. The primer *NIID_WH-1_Seq_F519* has a PIS of only 15%, meaning that only 15% of its positions are conserved across all sequences in the alignments. Previous works have already detected polymorphisms in primers and probes that may cause problem when performing

the testing [20, 21].

In terms of pairs of primers, we identified two pairs with a CoV2ID score of 100%: CoV2ID020 - CoV2ID036 and CoV2ID036 - CoV2ID041. Nevertheless, many other pairs of primers have high CoV2ID scores. For example, 25 pairs of primers have a CoV2ID score above 97.52%.

False positives could be a problem when using PCR primers and probes due to binding in non-target species. SARS-CoV-2 oligonucleotides with a high divergence to other strains should be preferred. Therefore, we have identified the most divergent oligonucleotides in other coronaviruses, i.e., the best ones to avoid false positives (Table 2). Twenty-eight primers and probes have a CoV2ID score in CoV2ID_alig03 below 20%, meaning they are highly divergent from other human coronaviruses. On the contrary, only three oligonucleotides have a CoV2ID score above 50%. For example, two probes from the Corman et al. protocol [11], *RdRP_SARSr-P1 and RdRP_SARSr-P2*, have CoV2ID scores above 67%. In general, available SARS-CoV-2 oligonucleotides diverge from other human coronaviruses by several positions. Nevertheless, caution is recommended when performing the experiments as some homology is observed in primer- and probe-binding sites.

We also analyzed the genetic diversity across the SARS-CoV-2 genome by measuring the diversity scores in 100 nucleotide sliding windows with 50 nucleotides of overlap. The PIS and PPI values revealed several 100 nucleotide regions completely conserved across the genome (see table on the database tab ‘Genome variation’), which may be used for the design of new oligonucleotides. Twenty-one 100 nucleotide windows (3.4%) from a total of 614 windows had a value of PIS of 100%. A total of 88 windows has a PPI of 100%.

Example of use

If the aim is to choose an oligonucleotide located in a conserved genomic region, the user can navigate through the “Search” tab on the top menu bar and open the “The best oligonucleotides” tab. The table with oligonucleotides is automatically ordered by the “CoV2ID Score” column filter. The user can also access the oligonucleotide summary information by clicking in the ID hyperlink. The database can also be used to filter all columns using the search tool. For instance, to access the best oligonucleotide located in ORF1a genomic region, the user can type “ORF1a” in the search box. The database table filter and only display the records related with the ORF1a region. In this example, the oligonucleotide CoV2ID028 has the highest CoV2ID score in the selected region. If the purpose is to design a new oligonucleotide, the database section “Genome variation” should be selected in the tab on the top menu bar. The user can then visualize the PIS and PPI values in 100 nucleotide sliding windows. The list of the most conserved genomic regions can be found in a table. In this case, the genomic region located between 7701-7800 has the highest PIS value (100%) considering alignment *alig02*. This section of the alignment can be visualized by clicking on the position value in the table. The user can also visualize any window of the alignment by using the ‘Show window in alignment’ box.

Funding

This research was supported by national funds through FCT - Foundation for Science and Technology within the scope of UIDB/04423/2020 and UIDP/04423/2020. J.C. also acknowledges the FCT funding for his research contract at CIIMAR, established under the transitional rule of Decree Law 57/2016, amended by Law 57/2017.

Table 1. Oligonucleotides with the highest conservation score considering the multiple sequence alignments of complete SARS-CoV-2 genomes.

Database reference	***Target***	***Original name***	***Sequence (5'-3')***	***Position in reference genome***	***Genomic region***	***Mean PIS****	***Mean 3PIS****	***Mean PPI****	***CoV2ID score***
CoV2ID007	PCR primer forward	Charite_RdRP_SARSr-F2	GTGARATGGTCATGTGTGGCGG	15431-15452	RdRp	100	100	100	100
CoV2ID011	PCR primer forward	Charite_E_Sarbeco_F1	ACAGGTACGTTAATAGTTAATAGCGT	26269-26294	E	100	100	100	100
CoV2ID019	Probe	HKU-NP	GCAAATTGTGCAATTTGCGG	29177-29196	N	100	100	100	100
CoV2ID020	PCR primer forward	WH-NIC N-F	CGTTTGGTGGACCCTCAGAT	28320-28339	N	100	100	100	100
CoV2ID028	PCR primer reverse	NIID_WH-1_Seq_R840	GACATAGCGAGTGTATGCC	805-823	ORF1a	100	100	100	100
CoV2ID036	PCR primer reverse	NIID_2019-nCOV_N_R2	TGGCAGCTGTGTAGGTCAAC	29263-29282	N	100	100	100	100
CoV2ID041	PCR primer forward	CDC_2019-nCoV_N2-F	TTACAAACATTGGCCGCAAA	29164-29183	N	100	100	100	100
CoV2ID047	PCR primer forward	Pasteur_nCoV_IP2-12669Fw	ATGAGCTTAGTCCTGTTG	12690-12707	RdRp	100	100	100	100
CoV2ID050	PCR primer forward	Pasteur_nCoV_IP4-14059Fw	GGTAACTGGTATGATTTCG	14080-14098	RdRp	100	100	100	100
CoV2ID052	Probe	Pasteur_nCoV_IP4-14084Probe	TCATACAAACCACGCCAGG	14105-14123	RdRp	100	100	100	100
CoV2ID008	PCR primer reverse	Charite_RdRP_SARSr-R1	CARATGTTAAASACACTATTAGCATA	15505-15530	RdRp	98.08	100	100	99.36
CoV2ID010	Probe	Charite_RdRP_SARSr-P1	CCAGGTGGWACRTCATCMGGTGATGC	15469-15494	RdRp	98.08	100	100	99.36
CoV2ID009	Probe	Charite_RdRP_SARSr-P2	CAGGTGGAACCTCATCAGGAGATGC	15470-15494	RdRp	98	100	100	99.33

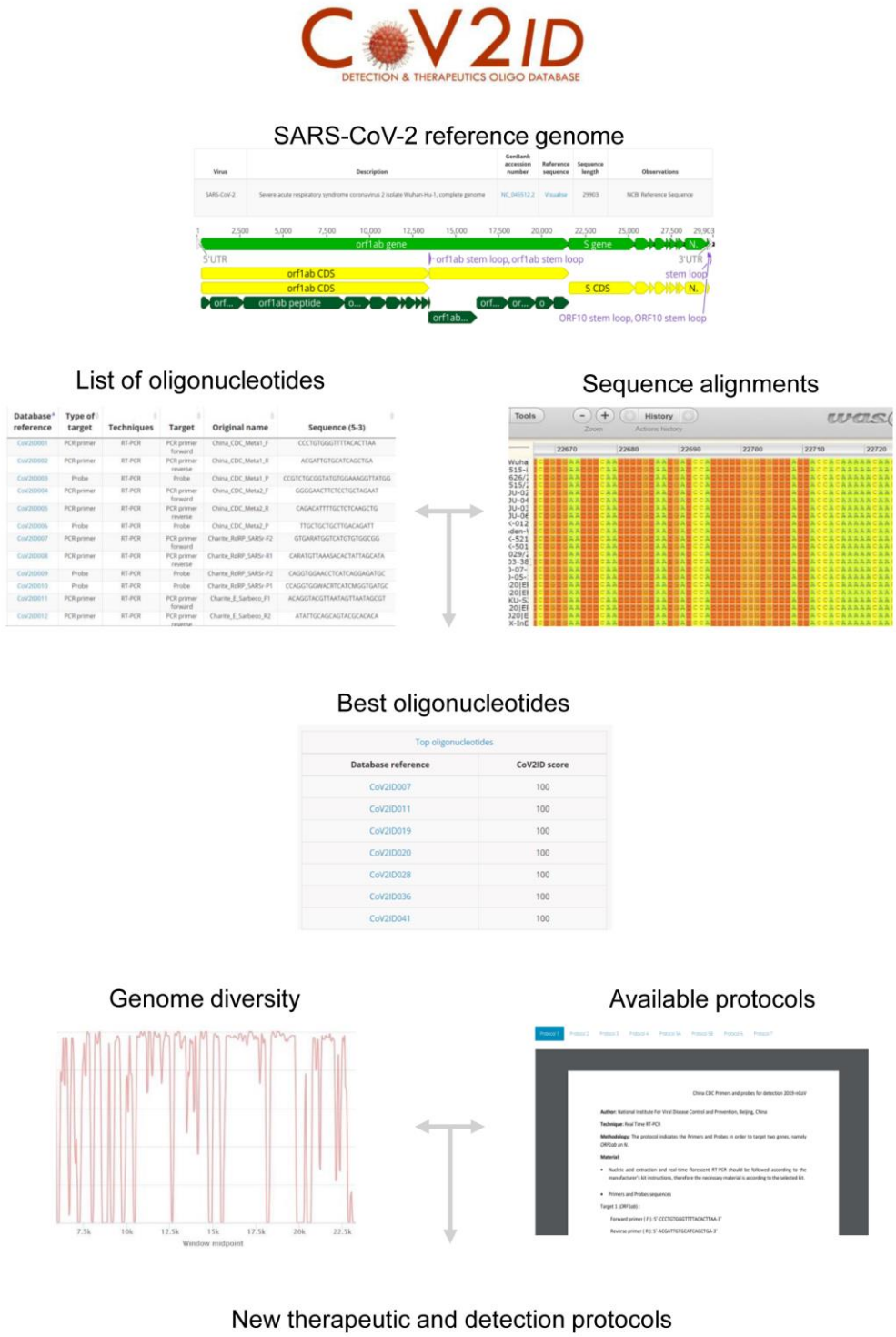
*Percentage of identical sites (PIS); percentage of identical sites in the last five nucleotides at the 3' end of the oligonucleotide (3'PIS); percentage of pairwise identity (PPI).

Table 2. Oligonucleotides with the lowest conservation score considering the alignment of consensus sequences of all human coronavirus.

Database reference	***Type***	***Original name***	***Sequence (5'-3')***	***Position in reference genome***	***Genomic region***	***Genomic position***	***PIS****	***PPI****	***CoV2ID score***
CoV2ID018	PCR primer reverse	HKU-NR	CGAAGGTGTGACTTCCATG	29236-29254	N	37104-37122	0	30.33	10.11
CoV2ID023	PCR primer forward	NIID_WH-1_F501	TTCGGATGCTCGAACTGCACC	484-504	ORF1a	924-942	0	35.59	11.86
CoV2ID006	Probe	China_CDC_Meta2_P	TTGCTGCTGCTTGACAGATT	28934-28953	N	36714-36731	0	36.77	12.26
CoV2ID026	PCR primer reverse	NIID_WH-1_R854	CAGAAGTTGTTATCGACATAGC	816-837	ORF1a	1421-1435	0	40.95	13.65
CoV2ID046	Probe	CDC_2019-nCoV_N3-P	AYCACATTGGCACCCGCAATCCTG	28704-28727	N	36464-36485	0	41.13	13.71
CoV2ID045	PCR primer reverse	CDC_2019-nCoV_N3-R	TGTAGCACGATTGCAGCATTG	28732-28752	N	36490-36510	0	44.67	14.89
CoV2ID004	PCR primer forward	China_CDC_Meta2_F	GGGGAACTTCTCCTGCTAGAAT	28881-28902	N	36660-36681	0	45.24	15.08
CoV2ID044	PCR primer forward	CDC_2019-nCoV_N3-F	GGGAGCCTTGAATACACCAAAA	28681-28702	N	36439-36460	0	45.45	15.15
CoV2ID037	Probe	NIID_2019-nCOV_N_P2	ATGTCGCGCATTGGCATGGA	29222-29241	N	37090-37109	5	41.19	15.4

*Percentage of identical sites (PIS); percentage of pairwise identity (PPI).

Figure 1. Screenshot of the data and tools included in the CoV2ID database. The website includes a NCBI sequence viewer of the SARS-CoV-2 reference genome with oligonucleotide annotations and feature tracks. The ‘Oligos’ section provides details on the available oligonucleotides. Multiple sequence alignments can be visualized using a multifunctional sequence viewer. The oligonucleotides are ranked according to their conservation in the multiple sequence alignments. The website also provided a scatter plot describing a sliding window analysis of diversity measures across the SARS-CoV-2 genome and the details of the detection protocols.



References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine. 2020.
2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine. 2020.
3. Lu H, Stratton CW, Tang YW. Outbreak of Pneumonia of Unknown Etiology in Wuhan China: the Mystery and the Miracle. Journal of Medical Virology.
4. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020:1-4.
5. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020;395(10224):565-74.
6. Li C, Yang Y, Ren L. Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. Infection, Genetics and Evolution. 2020:104285.
7. Holmes EC, Rambaut A. Viral evolution and the emergence of SARS coronavirus. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2004;359(1447):1059-65.
8. Liu R, Han H, Liu F, Lv Z, Wu K, Liu Y, et al. Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. Clinica Chimica Acta. 2020.
9. Ren X, Liu Y, Chen H, Liu W, Guo Z, Chen C, et al. Application and Optimization of RT-PCR in Diagnosis of SARS-CoV-2 Infection. Chaoqun and Zhou, Jianhui and Xiao, Qiang and Jiang, Guan-Min and Shan, Hong, Application and Optimization of RT-PCR in Diagnosis of SARS-CoV-2 Infection (2/25/2020). 2020.
10. Pfefferle S, Reucher S, Nörz D, Lütgehetmann M. Evaluation of a quantitative RT-PCR assay for the detection of the emerging coronavirus SARS-CoV-2 using a high throughput system. Eurosurveillance. 2020;25(9):2000152.
11. Corman V, Bleicker T, Brünink S, Drosten C, Zambon M. Diagnostic detection of 2019-nCoV by real-time RT-PCR. Berlin, Germany. 2020.
12. Spurgers KB, Sharkey CM, Warfield KL, Bavari S. Oligonucleotide antiviral therapeutics: antisense and RNA interference for highly pathogenic RNA viruses. Antiviral research. 2008;78(1):26-36.
13. Kole R, Krainer AR, Altman S. RNA therapeutics: beyond RNA interference and antisense oligonucleotides. Nature reviews Drug discovery. 2012;11(2):125-40.
14. Organization WH. Coronavirus disease (COVID-19) technical guidance: Laboratory testing for 2019-nCoV in human 2020. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratory-guidance
15. Carneiro J, Pereira F. EbolaID: An Online Database of Informative Genomic Regions for Ebola Identification and Treatment. PLoS neglected

tropical diseases. 2016;10(7).
16. Carneiro J, Resende A, Pereira F. The HIV oligonucleotide database (HIVoligoDB). Database. 2017;2017.
17. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in bioinformatics. 2019;20(4):1160-6.
18. Lv L, Li G, Chen J, Liang X, Li Y. Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13. BioRxiv. 2020.
19. Karamitros T, Papadopoulou G, Bousali M, Mexias A, Tsiodras S, Mentis A. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. bioRxiv. 2020.
20. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. Journal of Medical Virology. 2020.
21. Vogels CBF, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC, et al. Analytical sensitivity and efficiency comparisons of SARS-COV-2 qRT-PCR assays. medRxiv. 2020:2020.03.30.20048108. doi: 10.1101/2020.03.30.20048108.