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

João Carneiro [1], Catarina Gomes [2], Cátia Couto[2], 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.

Running Head: Oligo Database for SARS-CoV-2

Address for correspondence:

Filipe Pereira

IDENTIFICA
Science and Technology Park of the University of Porto (UPTEC)
Rua Alfredo Allen, N.º455/461
Porto 4200-135, Portugal
Tel: +351937546703; E-mail: 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, LAMP; polymorphisms

1. Introduction

The SARS-CoV-2 genome consists of a single, positive-stranded RNA with approximately 30 000 nucleotides. Thousands of genomic sequences are now available in public databases as the epidemic progresses. The great adaptability and infection capacity of RNA viruses depends in part from their high mutation rates.[1] As expected, available SARS-CoV-2 genomic sequences already show a large number of polymorphisms. 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.[2] For example, PCR primers and RT-qPCR probes are been widely used to detect SARS-CoV-2.[3,4] It is likely that oligonucleotides complementary to the virus RNA will be tested as possible antiviral agents.[5,6] However, polymorphisms can be a challenge for the efficiency of available assays since they may lead to false-negative results in detection tests or inefficient therapeutics.

2. Materials and Methods

2.1 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.

2.2. Oligonucleotides

The oligonucleotides were retrieved from peer reviewed publications [e.g.,[7-14]] and protocols provided by the World Health Organization (WHO). 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), as previously described.[15,16]

2.3 Genomic sequences

The SARS-CoV-2 'isolate Wuhan-Hu-1' (NC_045512.2) was used as reference. Genomes were obtained from the NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/) and the GISAID Initiative (https://www.gisaid.org/). 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* - SARS-CoV-2 complete genomes available at the NCBI.

b) *CoV2ID_alig02* - SARS-CoV-2 complete genomes from the GISAID with high coverage and no ambiguities or gaps.

c) *CoV2ID_alig03* - Alignment of the consensus sequence of each human coronavirus: SARS-CoV-2, HCoV-OC43, HCoV-HKU1, HCoV-NL63, HCoV-229E, MERS-CoV and SARS-CoV.

The genomes from the NCBI 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 sequences 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/) tools.

3. Results and Discussion

The CoV2ID database currently includes 145 oligonucleotides from 21 protocols: 64 PCR primers, 57 LAMP primers, 20 probes and 4 target generation oligonucleotides. 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 (*CoV2ID_alig02*) has 3251 complete SARS-CoV-2 genomes. The alignment has a PIS of 88.60% and a PPI of 99.96%. The 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. 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.

The database indicates which oligonucleotides bind to the most conserved regions of the SARS-CoV-2 using different measures of sequence conservation (Table 1). Several oligonucleotides have a perfect homology to all available genomes (CoV2ID score of 100%). For example, nine oligonucleotides are 100% complementary to all genomes. On the contrary, some oligonucleotides have several mismatches to SARS-CoV-2 genomes. There are six oligonucleotides with a CoV2ID score of below 60%. For example, primers *NIID_WH-1_F24381* and *NIID_WH-1_Seq_F519* have a CoV2ID score of below 50%. Previous works have already detected polymorphisms in primers and probes that may cause problems when performing the testing [18,19]. In terms of pairs of primers, we identified 238 pairs with a CoV2ID score above 90%. For example, the pair of primers *Pasteur_nCoV_IP4-14059Fw* and *Pasteur_nCoV_IP4-14146Rv* had a CoV2ID score of 99.58%.

SARS-CoV-2 oligonucleotides with a high divergence to other strains should be preferred to avoid false positives resulting from the putative binding in non-target species. We have identified the most divergent oligonucleotides in other coronaviruses, i.e., the best ones to avoid false positives (Table 2). Fifty-six oligonucleotides have a CoV2ID score below 20% regarding other coronaviruses (*CoV2ID_alig03*), meaning they are highly divergent. On the contrary, seven oligonucleotides have a CoV2ID score above 70%. The most conserved oligonucleotide (CoV2ID102; *Chan_RdRp_gene_R*) has a CoV2ID score of 91.43%, but was designed to target all SARS-related coronaviruses [14], which explains its high conservation. In general, available SARS-CoV-2 oligonucleotides diverge from other human coronaviruses by several positions, and therefore are unlikely to cause false-positives.

4. 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. 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 with 50 nucleotides of overlap. The list of the most conserved genomic regions can be found in a table. In this case, the genomic regions 15901-16000 and 15951-16050 had the highest PIS value (100%) considering alignment *CoV2ID_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

Acknowledgements
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.

Conflict of Interests
The authors declare that there are no conflict of interests.

References

1. 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):1059-1065.

2. Weissleder R, Lee H, Ko J, Pittet MJ. COVID-19 diagnostics in context. *Science Translational Medicine.* 2020;12(546):eabc1931.

3. Liu R, Han H, Liu F, 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.

4. 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.

5. Kole R, Krainer AR, Altman S. RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nature reviews Drug discovery.* 2012;11(2):125-140.

6. 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.

7. Lu R, Wu X, Wan Z, et al. Development of a Novel Reverse Transcription Loop-Mediated Isothermal Amplification Method for Rapid Detection of SARS-CoV-2. *Virologica Sinica.* 2020:1.

8. Huang WE, Lim B, Hsu CC, et al. RT-LAMP for rapid diagnosis of coronavirus SARS-CoV-2. *Microbial Biotechnology.* 2020.

9. Yan C, Cui J, Huang L, et al. Rapid and visual detection of 2019 novel coronavirus (SARS-CoV-2) by a reverse transcription loop-mediated isothermal amplification assay. *Clinical Microbiology and Infection.* 2020.

10. Yip CC-Y, Ho C-C, Chan JF-W, et al. Development of a Novel, Genome Subtraction-Derived, SARS-CoV-2-Specific COVID-19-nsp2 Real-Time RT-PCR Assay and Its Evaluation Using Clinical Specimens. *International Journal of Molecular Sciences.* 2020;21(7):2574.

11. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance.* 2020;25(3):2000045.

12. Broughton JP, Deng X, Yu G, et al. CRISPR–Cas12-based detection of SARS-CoV-2. *Nature Biotechnology.* 2020:1-5.

13. Nalla AK, Casto AM, Huang M-LW, et al. Comparative Performance of SARS-CoV-2 Detection Assays Using Seven Different Primer-Probe Sets and One Assay Kit. *Journal of clinical microbiology.* 2020;58(6).

14. Chan JF-W, Yuan S, Kok K-H, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet.* 2020;395(10223):514-523.

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):1160-1166.

18. Vogels CBF, Brito AF, Wyllie AL, et al. Analytical sensitivity and efficiency comparisons of SARS-COV-2 qRT-PCR assays. *medRxiv.* 2020:2020.2003.2030.20048108.

19. Wang C, Liu Z, Chen Z, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. *Journal of Medical Virology.* 2020.

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

Database reference	*Type*	*Original name*	*Sequence (5'-3')*	*Genome position*	*Genomic region*	*Mean PIS**	*Mean PPI**	*CoV2ID score*
CoV2ID001	PCR primer forward	China_CDC_Meta1_F	CCCTGTGGGTTTTACACTTAA	13342-13362	ORF1ab	100	100	100
CoV2ID006	Probe	China_CDC_Meta2_P	TTGCTGCTGCTTGACAGATT	28934-28953	N	100	100	100
CoV2ID008	PCR primer reverse	Charite_RdRP_SARSr-R1	CARATGTTAAASACACTATTAGCATA	15505-15530	ORF1ab	100	100	100
CoV2ID047	PCR primer forward	Pasteur_nCoV_IP2-12669Fw	ATGAGCTTAGTCCTGTTG	12690-12707	ORF1ab	100	100	100
CoV2ID050	PCR primer forward	Pasteur_nCoV_IP4-14059Fw	GGTAACTGGTATGATTTCG	14080-14098	ORF1ab	100	100	100
CoV2ID054	PCR primer reverse	RBD-qR1	CTCAAGTGTCTGTGGATCACG	23291 - 23311	S	100	100	100
CoV2ID073	LAMP forward inner primer	FIP 2019-nCoV E-gene	CTAGCCATCCTTACTGCGCTACTCACGTTAACAATATTGCA	26374 - 26394	E	100	100	100
CoV2ID075	LAMP loop backward primer	LB 2019-nCoV E-gene	TGAGTACATAAGTTCGTAC	26235 - 26253	E	100	100	100
CoV2ID086	PCR primer forward	S-F	CTTCCCTCAGTCAGCACCTC	24715 - 24734	S	100	100	100
CoV2ID078	Probe	N_Sarbeco_P	ACTTCCTCAAGGAACAACATTGCCA	28753 - 28777	N	98	100	99.33
CoV2ID090	LAMP forward inner primer	orf1ab-4FIP	GGCATCACAGAATTGTACTGTTTTTGCGTATACGCCAACTTAGG	13958 - 13976	ORF1ab	98	100	99.33
CoV2ID085	PCR primer reverse	orf1ab-R	CCCTGGTCAAGGTTAATATAGGCA	14165 - 14188	ORF1ab	97.91	100	99.3
CoV2ID122	LAMP loop forward primer	LF 2019-nCoV N15-gene	GCAATGTTGTTCCTTGAGGAAGTT	28752 - 28775	N	97.91	100	99.3

*Percentage of Identical Sites (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')*	*Genome position*	*Genomic region*	*Mean PIS**	*Mean PPI**	*CoV2ID score*
CoV2ID001	PCR primer forward	China_CDC_Meta 1_F	CCCTGTGGGTTTTACA CTTAA	13342-13362	ORF1ab	0	0	28.57
CoV2ID124	LAMP outer forward primer	F3 2019-nCoV S17-gene	TCTTTCACACGTGGTG TT	21653 - 21670	S	0	0	28.57
CoV2ID072	LAMP backward inner primer	BIP 2019-nCoV E-gene	ACCTGTCTCTTCCGAA ACGAATTTGTAAGCAC AAGCTGATG	26214 - 26233	E	0	0	29.22
CoV2ID127	LAMP backward inner primer	BIP 2019-nCoV S17-gene	CTCTGGGACCAATGG TACTAAGAGGACTTCT CAGTGGAAGCA	21838 - 21855	S	0	0	29.37
CoV2ID075	LAMP loop backward primer	LB 2019-nCoV E-gene	TGAGTACATAAGTTCG TAC	26235 - 26253	E	0	0	29.46
CoV2ID128	LAMP loop forward primer	LF 2019-nCoV S17-gene	GAAAGGTAAGAACAA GTCCTGAGT	21713 - 21736	S	0	0	29.85
CoV2ID132	LAMP forward inner primer	FIP 2019-nCoV O117-gene	GGTTTTCAAGCCAGAT TCATTATGGATGTCAC AATTCAGAAGTAGGA	1381 - 1402	ORF1ab	0	0	30.24
CoV2ID018	PCR primer reverse	HKU-NR	CGAAGGTGTGACTTC CATG	29236-29254	N	0	0	30.33
CoV2ID130	LAMP outer forward primer	F3 2019-nCoV O117-gene	CCCCAAAATGCTGTTG TT	1346 - 1363	ORF1ab	0	0	28.57

*Percentage of Identical Sites (PIS); Percentage of Pairwise Identity (PPI).