Genetic variation of coronavirus SARS-CoV-2 in target regions of CRISPR-based diagnostic assays

Kashif Aziz Khan

Department of Biology, York University, Toronto, Ontario, Canada.

Corresponding Author:
Email address: kakhan@yorku.ca

Abstract:

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; initially named as 2019-nCoV) is the cause of the novel coronavirus disease 2019 (COVID-19) pandemic. Its diagnosis relies on the molecular detection of the viral RNA by polymerase chain reaction (PCR) while newer rapid CRISPR-based diagnostic tools are being developed for point-of-care and in-home diagnosis. As molecular diagnostic assays rely on the detection of unique sequences of viral nucleic acid, these are prone to mismatches due to mutations in the viral genome. This stage 1 protocol proposes the verification of the sequence variability within the target regions of recently developed CRISPR-based diagnostic assays in the SARS-CoV-2 genome. This would be accomplished using more than 60 000 publicly available sequences of viral isolates from around the world. The absence of any mutations in target regions of the assay used would provide a higher degree of confidence in the alternative tests being developed while the presence of mutations could help guide the assay development efforts. We believe that this study would provide potentially important information for clinicians, researchers, and policy-makers.

Keywords: SARS-CoV-2, coronavirus, sequence variation, mutations, diagnosis, CRISPR, COVID-19

Short Title: Mutations in CRISPR-based COVID-19 tests

Introduction:

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; initially named as 2019-nCoV) was firstly isolated from a cluster of pneumonia patients in Wuhan, China and is the cause of novel coronavirus disease termed COVID-19 (Wu et al. 2020; Zhou et al. 2020; Zhu et al. 2020). The rapid spread of the virus has resulted in a global pandemic reaching more than 200 countries and territories (Worldometers.info 2020). SARS-CoV-2 has been classified as a member of family *Coronaviridae* in the genus *Betacoronavirus* along with SARS-CoV and Middle East respiratory syndrome (MERS)-CoV (Gorbalenya et al. 2020). The sequencing of the virus from the patients early in the outbreak has shown that its single-stranded RNA genome is ~30 kb in size (Chan et al. 2020; Lu et al. 2020; Wu et al. 2020). SARS-CoV-2 genome has been predicted to encode at least 10 open reading frames (ORFs) for structural and accessory proteins, based on similarity with SARS-CoV. As per current annotation (NC_045512.2), these viral ORFs encode replicase ORF1ab, spike (S), envelope (E), membrane (M) and nucleocapsid (N), and at least 6 accessory proteins (3a, 6, 7a, 7b, 8, and 10) (NCBI 2020).

The pandemic has serious public health and economic implications. The day-to-day life of billions of people has been affected due to different forms of social distancing measures in place in different parts of the world to mitigate the spread of the virus. Thus, the widespread availability of rapid and reliable diagnostic testing is an important tool for policymakers to make public health decisions. The current diagnosis of COVID-19 relies on the molecular detection of the viral RNA from the patient samples using nucleic acid amplification tests (NAAT) like polymerase chain reaction (PCR) (WHO 2020). However, PCR requires specialized equipment and trained staff to perform the test and interpret the results and thus is a challenge for remote low-resource settings (Giri & Rana 2020). One of the alternatives being explored is the CRISPR-based nucleic-acid detection methods that may be particularly useful for point-of-care and in-house diagnosis. CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR-associated), a component of the bacterial immune system to infectious nucleic acid, has been widely used as a gene-editing tool. This technology exploits the ability of Cas proteins to accurately target any region in DNA in association with CRISPR RNA (crRNA) that matches the target DNA with or without the requirement of a protospacer adjacent motif (PAM) (Moon et al. 2019). Initially explored for Cas9 protein, the application of CRISPR in nucleic acid detection

57 emerged as a viable tool with the discovery of promiscuous collateral cleavage activity of

58 Cas12a (also known as C2c2) and Cas13a (also known as Cpf1) after target recognition (Chen et

59 al. 2018; Gootenberg et al. 2017). Subsequently, Cas14 was also shown to have this property

60 (Harrington et al. 2018). Several CRISPR-based methods have been developed for the detection

61 of RNA and DNA viruses (reviewed in (Jia et al. 2020; Strich et al. 2019)). With the emergence

62 of the novel coronavirus, scientists are rapidly employing these tools for the detection of SARS-

63 CoV-2 from patient samples as an alternative to PCR. The Cas12a based DETECTR assay has

64 been employed for diagnosis of COVID-19 from patient samples targeting viral genes N and E

65 (Broughton et al. 2020). Similarly, Cas12b based CASdetec has also been developed for the

66 detection of SARS-CoV-2 (Guo et al. 2020). Cas13a based SHERLOCK has also been approved

67 as the first CRISPR-based assay by the US Food and Drug Administration (FDA) under

68 Emergency Use Authorization (Sherlock 2020). Several other CRISPR-based methods are also

69 being developed (Esbin et al. 2020).

70 The molecular diagnosis of SARS-CoV-2 may be jeopardized by potential preanalytical and

71 analytical vulnerabilities leading to false-positive or false-negative results (Lippi et al. 2020). As

72 molecular diagnostic assays rely on the detection of unique sequences of viral nucleic acid, these

73 are prone to mismatches due to genetic variability in the viral genome. It is known that mutations

74 at primer/probe binding regions of the viral genome can result in potential mismatches and false-

75 negative PCR diagnosis (Lefever et al. 2013; Stadhouders et al. 2010; Whiley & Sloots 2005).

76 We have demonstrated the genetic variability in the primer/probe binding regions of the SARS-

77 CoV-2 genome using more than 17 000 sequences highlighting the importance of periodic

78 sequence verification for optimal virus detection (Khan & Cheung 2020). Although analyzed

79 much fewer sequences, other concurrent studies reached the same conclusion (Farkas et al. 2020;

80 Osorio & Correia-Neves 2020). One of the focus of CRISPR-diagnostics remained to improve

81 assay specificity as tolerance of mismatches by Cas9 can result in a false-positive diagnosis. This

82 risk has been minimized with the newer Cas proteins, Cas12 and Cas13, that have a lower

83 tolerance for mismatches compared to Cas9 especially in the "seed" region (Safari et al. 2019).

84 *Francisella novicida* Cas9 (FnCas9) has been reported to have higher specificity and lower

85 tolerance for mismatches compared to *Streptococcus pyogenes* Cas9 (SpCas9) especially at the

86 PAM-distal end (Acharya et al. 2019). However, this raises the possibility that these tests may

87 miss certain viral variants due to genetic variability in the regions targeted by these assays.

88   Mismatch intolerant seed region of ~6 nucleotides is located in the PAM-proximal region for

89   Cas12 (Chen et al. 2018; Kim et al. 2016) while the seed region is located in the central region of

90   crRNA for Cas13a (Abudayyeh et al. 2017; Cox et al. 2017).

91   The objective of this study is the verification of the sequence variability within the target regions

92   of recently developed CRISPR-based COVID-19 diagnostic assays using publicly available

93   SARS-CoV-2 sequences from around the world. The absence of any mutations in target regions

94   of the assay used would provide a higher degree of confidence in the alternative tests being

95   developed while the presence of mutations could help guide the assay development efforts. We

96   believe that this study would provide potentially important information for clinicians, researchers

97   and policy-makers.

98

Methods:

100   CRISPR-based diagnostic assays and SARS-CoV-2 sequences

101   At least 10 crRNA of recently published* CRISPR-based methods would be selected based on

102   the literature review. The sequence variability would be determined in the publicly available

103   viral sequences within the target regions of CRISPR-based diagnostic assays using the protocol

104   described previously (Khan 2020; Khan & Cheung 2020). The summary of the sequence tracing

105   pipeline is shown in Figure 1. The design planner is included in Table 1.

106   Complete genome sequences of the virus would be obtained from the Global Initiative on

107   Sharing All Influenza Data (GISAID) EpiCoV database (Shu & McCauley 2017) that are

108   available upon free registration (https://www.gisaid.org/). A total of 60 000** near full-length

109   sequences would be downloaded by applying the full length filter (>29,000bp). The RNA

110   genome of SARS-CoV-2 is shown in DNA format as per scientific convention. The viral

111   sequences are deposited in GISAID EpiCoV database by the laboratories around the world and a

112   list of accession numbers would be included as supplemental data.

113   * Assay published in the coming weeks would also be included.

114   ** The number of sequences in EpiCoV database is growing on a daily basis and the exact

115   number of included sequences would be updated in the 2nd stage submission.

116

117   Multiple sequence alignment and alignment processing

118   Multiple sequence alignment (MSA) would be performed using MAFFT (Multiple Alignment

119   with Fast Fourier Transform) program version 7 dedicated to closely-related viral genomes

120   available online (https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html)

121   (Katoh et al. 2002; Katoh et al. 2019). The complete genome of Wuhan-Hu-1, which is 29,903

122   bp long, would be downloaded from NCBI nucleotide and would be included as a reference

123   (NCBI Reference Sequence: NC_045512.2). The aligned sequences would be downloaded in

124   PIR format and the alignment would be visually inspected using the AliView program 1.26

125   (Larsson 2014). MAFFT online service dedicated to MSA of closely-related viral genomes

126   supports up to 20 000 sequences of ~30 kb in length. Thus the tracing pipeline would be applied

127   to sequences in batches of less than 20 000 sequences and results would be subsequently

128   combined. To evaluate the sequence variability in target regions referred here as the region of

129   interest (ROI), the sequence targeted by each assay will be saved as a separate file in FASTA

130   format.

131

132   Sequence variation in target regions of CRISPR-based assays

133   The MSA sequence for each ROI would be stratified using the SequenceTracer module

134   (http://entropy.szu.cz:8080/EntropyCalcWeb/sequences) of the Alignment Explorer (Nagy et al.

135   2019). This tool segregated sequences into discrete groups of identical sequence variants along

136   with their frequency. SequenceTracer automatically segregates the sequences with ambiguous

137   sequences, stretches of NNNs, and missing sequences in ROI and these sequences will be

138   excluded from the analysis. To remove extremely low prevalent variants and sequencing errors

139   in the data, only the sequence variants occurring more than once would be further considered

140   (Bioinformatics-Institute 2020). The viral isolates would be reported as the frequency of hits

141   with 100% match and hits with mismatches along with a summary of mutated nucleotides for

142   each crRNA. In addition, the frequency of sequence variants with mismatches in the seed and

143   PAM region of each crRNA would be determined. The base composition of each nucleotide

144   position of highly variable regions, if any, would further be analyzed using the positional

145   nucleotide numerical summary (PNNS) calculator

146   (http://entropy.szu.cz:8080/EntropyCalcWeb/pnns) of the Alignment Explorer (Nagy et al.

147   2019).

148

References

150   Abudayyeh OO, Gootenberg JS, Essletzbichler P, Han S, Joung J, Belanto JJ, Verdine V, Cox
151   DBT, Kellner MJ, Regev A, Lander ES, Voytas DF, Ting AY, and Zhang F. 2017. RNA
152   targeting with CRISPR–Cas13. *Nature* 550:280-284. 10.1038/nature24049
153   Acharya S, Mishra A, Paul D, Ansari AH, Azhar M, Kumar M, Rauthan R, Sharma N, Aich M,
154   Sinha D, Sharma S, Jain S, Ray A, Jain S, Ramalingam S, Maiti S, and Chakraborty D.
155   2019. Francisella novicida Cas9 interrogates genomic DNA with very high specificity
156   and can be used for mammalian genome editing. *Proceedings of the National Academy of*
157   *Sciences* 116:20959-20968. 10.1073/pnas.1818461116
158   Bioinformatics-Institute. 2020. CoVsurver: Frequently Asked Questions. *Available at*
159   *https://corona.bii.a-star.edu.sg/faq.html* (accessed July 4 2020).
160   Broughton JP, Deng X, Yu G, Fasching CL, Servellita V, Singh J, Miao X, Streithorst JA,
161   Granados A, Sotomayor-Gonzalez A, Zorn K, Gopez A, Hsu E, Gu W, Miller S, Pan C-
162   Y, Guevara H, Wadford DA, Chen JS, and Chiu CY. 2020. CRISPR–Cas12-based
163   detection of SARS-CoV-2. *Nature Biotechnology*. 10.1038/s41587-020-0513-4
164   Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, and Yuen KY. 2020. Genomic
165   characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient
166   with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect* 9:221-236.
167   10.1080/22221751.2020.1719902
168   Chen JS, Ma E, Harrington LB, Da Costa M, Tian X, Palefsky JM, and Doudna JA. 2018.
169   CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity.
170   *Science* 360:436-439. 10.1126/science.aar6245
171   Cox DBT, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J, and Zhang F. 2017.
172   RNA editing with CRISPR-Cas13. *Science* 358:1019-1027. 10.1126/science.aaq0180
173   Esbin MN, Whitney ON, Chong S, Maurer A, Darzacq X, and Tjian R. 2020. Overcoming the
174   bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for
175   COVID-19 detection. *Rna* 26:771-783. 10.1261/rna.076232.120
176   Farkas C, Fuentes-Villalobos F, Garrido JL, Haigh J, and Barría MI. 2020. Insights on early
177   mutational events in SARS-CoV-2 virus reveal founder effects across geographical
178   regions. *PeerJ* 8. 10.7717/peerj.9255
179   Giri AK, and Rana DR. 2020. Charting the challenges behind the testing of COVID-19 in
180   developing countries: Nepal as a case study. *Biosafety and Health* 2:53-56.
181   10.1016/j.bsheal.2020.05.002
182   Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, Verdine V, Donghia
183   N, Daringer NM, Freije CA, Myhrvold C, Bhattacharyya RP, Livny J, Regev A, Koonin
184   EV, Hung DT, Sabeti PC, Collins JJ, and Zhang F. 2017. Nucleic acid detection with
185   CRISPR-Cas13a/C2c2. *Science* 356:438-442. 10.1126/science.aam9321
186   Gorbalenya AE, Baker SC, Baric RS, Groot RJd, Drosten C, Gulyaeva AA, Haagmans BL,
187   Lauber C, Leontovich AM, Neuman BW, Penzar D, Perlman S, Poon LLM, Samborskiy

188   DV, Sidorov IA, Sola I, Ziebuhr J, and Coronaviridae Study Group of the International
189   Committee on Taxonomy of Viruses. 2020. The species Severe acute respiratory
190   syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat*
191   *Microbiol* 5:536-544. 10.1038/s41564-020-0695-z
192   Guo L, Sun X, Wang X, Liang C, Jiang H, Gao Q, Dai M, Qu B, Fang S, Mao Y, Chen Y, Feng
193   G, Gu Q, Wang RR, Zhou Q, and Li W. 2020. SARS-CoV-2 detection with CRISPR
194   diagnostics. *Cell Discovery* 6. 10.1038/s41421-020-0174-y
195   Harrington LB, Burstein D, Chen JS, Paez-Espino D, Ma E, Witte IP, Cofsky JC, Kyrpides NC,
196   Banfield JF, and Doudna JA. 2018. Programmed DNA destruction by miniature CRISPR-
197   Cas14 enzymes. *Science* 362:839-842. 10.1126/science.aav4294
198   Jia F, Li X, Zhang C, and Tang X. 2020. The expanded development and application of CRISPR
199   system for sensitive nucleotide detection. *Protein & Cell*. 10.1007/s13238-020-00708-8
200   Katoh K, Misawa K, Kuma K, and Miyata T. 2002. MAFFT: a novel method for rapid multiple
201   sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059-3066.
202   10.1093/nar/gkf436
203   Katoh K, Rozewicki J, and Yamada KD. 2019. MAFFT online service: multiple sequence
204   alignment, interactive sequence choice and visualization. *Brief Bioinform* 20:1160-1166.
205   10.1093/bib/bbx108
206   Khan KA. 2020. In silico inclusivity evaluation for PCR diagnostic assays of SARS-CoV-2. *Bio-*
207   *protocol*.
208   Khan KA, and Cheung P. 2020. Presence of mismatches between diagnostic PCR assays and
209   coronavirus SARS-CoV-2 genome. *Royal Society Open Science* 7:200636.
210   doi:10.1098/rsos.200636
211   Kim D, Kim J, Hur JK, Been KW, Yoon S-h, and Kim J-S. 2016. Genome-wide analysis reveals
212   specificities of Cpf1 endonucleases in human cells. *Nature Biotechnology* 34:863-868.
213   10.1038/nbt.3609
214   Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets.
215   *Bioinformatics* 30:3276-3278. 10.1093/bioinformatics/btu531
216   Lefever S, Pattyn F, Hellemans J, and Vandesompele J. 2013. Single-nucleotide polymorphisms
217   and other mismatches reduce performance of quantitative PCR assays. *Clin Chem*
218   59:1470-1480. 10.1373/clinchem.2013.203653
219   Lippi G, Simundic AM, and Plebani M. 2020. Potential preanalytical and analytical
220   vulnerabilities in the laboratory diagnosis of coronavirus disease 2019 (COVID-19). *Clin*
221   *Chem Lab Med*. 10.1515/cclm-2020-0285
222   Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X,
223   Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin
224   Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W,
225   Shi W, and Tan W. 2020. Genomic characterisation and epidemiology of 2019 novel
226   coronavirus: implications for virus origins and receptor binding. *Lancet* 395:565-574.
227   10.1016/S0140-6736(20)30251-8
228   Moon SB, Kim DY, Ko J-H, and Kim Y-S. 2019. Recent advances in the CRISPR genome
229   editing tool set. *Experimental & Molecular Medicine* 51:1-11. 10.1038/s12276-019-
230   0339-7
231   Nagy A, Jirinec T, Jirincova H, Cernikova L, and Havlickova M. 2019. In silico re-assessment of
232   a diagnostic RT-qPCR assay for universal detection of Influenza A viruses. *Sci Rep*
233   9:1630. 10.1038/s41598-018-37869-w

234 NCBI. 2020. Sequence Viewer 3.36.0: Severe acute respiratory syndrome coronavirus 2 isolate
235   Wuhan-Hu-1, complete genome. *Available at*
236   *https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=NC_045512.2* (accessed May 15
237   2020).
238 Osorio NS, and Correia-Neves M. 2020. Implication of SARS-CoV-2 evolution in the sensitivity
239   of RT-qPCR diagnostic assays. *Lancet Infect Dis*. 10.1016/S1473-3099(20)30435-7
240 Safari F, Zare K, Negahdaripour M, Barekati-Mowahed M, and Ghasemi Y. 2019. CRISPR Cpf1
241   proteins: structure, function and implications for genome editing. *Cell & Bioscience* 9.
242   10.1186/s13578-019-0298-7
243 Sherlock. 2020. Sherlock CRISPR SARS-CoV-2 kit. *Available at*
244   *https://www.fda.gov/media/137746/download* (accessed July 3 2020).
245 Shu Y, and McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from
246   vision to reality. *Euro Surveill* 22. 10.2807/1560-7917.ES.2017.22.13.30494
247 Stadhouders R, Pas SD, Anber J, Voermans J, Mes TH, and Schutten M. 2010. The effect of
248   primer-template mismatches on the detection and quantification of nucleic acids using the
249   5' nuclease assay. *J Mol Diagn* 12:109-117. 10.2353/jmoldx.2010.090035
250 Strich JR, Chertow DS, and Kraft CS. 2019. CRISPR-Cas Biology and Its Application to
251   Infectious Diseases. *Journal of Clinical Microbiology* 57. 10.1128/jcm.01307-18
252 Whiley DM, and Sloots TP. 2005. Sequence variation in primer targets affects the accuracy of
253   viral quantitative PCR. *J Clin Virol* 34:104-107. 10.1016/j.jcv.2005.02.010
254 WHO. 2020. Laboratory testing for coronavirus disease (COVID-19) in suspected human cases:
255   Interim Guidance. *Available at https://www.who.int/publications/i/item/laboratory-*
256   *testing-for-2019-novel-coronavirus-in-suspected-human-cases-20200117* (accessed July
257   3 2020).
258 Worldometers.info. 2020. COVID-19 Coronavirus Pandemic. *Available at*
259   *https://www.worldometers.info/coronavirus/* (accessed April, 16 2020).
260 Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan
261   ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, and Zhang YZ.
262   2020. A new coronavirus associated with human respiratory disease in China. *Nature*
263   579:265-269. 10.1038/s41586-020-2008-3
264 Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen
265   HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS,
266   Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, and Shi ZL.
267   2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin.
268   *Nature* 579:270-273. 10.1038/s41586-020-2012-7
269 Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan
270   F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W, China Novel Coronavirus I, and
271   Research T. 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N*
272   *Engl J Med* 382:727-733. 10.1056/NEJMoa2001017

273

274

275   Table 1: Study design planner.

276

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis Plan	Interpretation given different outcomes
Are there any mutations in the SARS-CoV-2 genome within the target regions of CRISPR-based COVID-19 diagnostic assays?	Given the potential of virus to mutate during natural evolution, genetic variability in the target regions of CRISPR-based COVID-19 diagnostic assays can result in crRNA-template mismatches.	>60 000* viral isolates would be download from GISAID EpiCov database. *The number of sequences is growing on a daily basis and the exact number of included sequences would be updated in the 2nd stage Inclusion criteria Only full length (>29,000 bp) Exclusion Criteria The sequences with stretches of NNNs, ambiguous sequences, and missing sequences in the region of interest (ROI) will be considered low quality and would be excluded.	- Sequences would be aligned using MAFFT - The alignment would be inspected visually using AliView - Sequence variability would be traced using SequenceTracer. - The highly variable region, if any, would be further analyzed for nucleotide composition at each position using positional nucleotide numerical summary (PNNS) - The complete genome of Wuhan-Hu-1 would act as a positive control (NCBI Reference Sequence: NC_045512.2).	- In the event of a negative result, it would be concluded that there is no evidence of a difference between crRNA and viral isolates. - This would serve as a reference for clinicians and researchers using CRISPR-based assays for the detection of SARS-CoV-2.

277

278   Figure 1: Sequence tracing pipeline to be used in the study. Adapted from (Khan & Cheung
279   2020) shared under a CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

x