Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Palindromic target site identification in SARS-CoV-2, MERS-CoV and SARS-CoV-1 by adopting CRISPR-Cas technique

Nimisha Ghosh [a,b,1], Indrajit Saha [c,*,1], Nikhil Sharma [d]

[a] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland
[b] Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India
[c] Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India
[d] Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

ARTICLE INFO

Edited by Xavier Carette

Keywords:
CRISPR-Cas
Primer
SARS-CoV-2
Palindrome
Protospacer Adjacent Motif
Target site identification

ABSTRACT

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) associated Cas protein (CRISPR-Cas) has turned out to be a very important tool for the rapid detection of viruses. This can be used for the identification of the target site in a virus by identifying a 3–6 nt length Protospacer Adjacent Motif (PAM) adjacent to the potential target site, thus motivating us to adopt CRISPR-Cas technique to identify SARS-CoV-2 as well as other members of Coronaviridae family. In this regard, we have developed a fast and effective method using *k*-mer technique in order to identify the PAM by scanning the whole genome of the respective virus. Subsequently, palindromic sequences adjacent to the PAM locations are identified as the potential target sites. Palindromes are considered in this work as they are known to identify viruses. Once all the palindrome-PAM combinations are identified, PAMs specific for the RNA-guided DNA Cas9/Cas12 endonuclease are identified to bind and cut the target sites. In this regard, PAMs such as 5'-TGG-3' and 5'-TTTA-3' in NSP3 and Exon for SARS-CoV-2, 5'-GGG-3' and 5'-TGG-3' in Exon and NSP2 for MERS-CoV and 5'-AGG-3' and 5'-TTTG-3' in Helicase and NSP3 respectively for SARS-CoV-1 are identified corresponding to SpCas9 and FnCas12a endonucleases. Finally, to recognise the target sites of Coronaviridae family as cleaved by SpCas9 and FnCas12a, complements of the palindromic target regions are designed as primers or guide RNA (gRNA). Therefore, such complementary gRNAs along with respective Cas proteins can be considered in assays for the identification of SARS-CoV-2, MERS-CoV and SARS-CoV-1.

1. Introduction

COVID-19, the disease caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) has affected a lot of people around the globe and has claimed more than 5.3 million lives as of 15*th* December 2021.[2] SARS-CoV-2 belongs to the family of Coronaviridae which also accommodates MERS-CoV and SARS-CoV-1 viruses (Zhou et al., 2020). The symptoms of COVID-19 include cough, fever, dyspnoea, diarrhoea, myalgia (Hosseini et al., 2020) and in some extreme cases may also lead to severe respiratory distress leading to eventual death. Moreover, comorbidity issue in COVID-19 is relatively high and targets different organs like kidney, liver, heart, brain, etc. (Dey et al., 2020; Qi et al., 2020).

Since its spread, symptom-based diagnosis of COVID-19 is being performed which includes chest X-ray and CT scan, quantitative reverse transcription polymerase chain reaction (qt-PCR) and antibody test. Most recently, another rapid detection method based on CRISPR-Cas has been proposed by the researchers (Broughton et al., 2020b). Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) associated Cas protein (CRISPR-Cas) is an adaptive immune system in prokaryotic organism that can provide resistance to foreign elements. This system has been exploited in recent times as a powerful gene editing tool and for the diagnosis and inactivation of viruses (Jia et al., 2020). The efficiency of CRISPR-Cas technique is dependent on the design of guide RNA

Abbreviations: CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats; PAM, Protospacer Adjacent Motif; gRNA, guide RNA.

* Corresponding author.

E-mail address: indrajit@nitttrkol.ac.in (I. Saha).

[1] Equally contributed.

[2] https://www.worldometers.info/coronavirus/.

https://doi.org/10.1016/j.gene.2021.146136
Received 1 October 2021; Received in revised form 16 December 2021; Accepted 20 December 2021
Available online 6 January 2022

(gRNA). gRNA guides the Cas protein to the intended DNA site and then creates a DNA double-strand break resulting in its repair which leads to different DNA sequence modifications (Rahman et al., 2021). The study by Bhat et al. (2020) provides an overview of using CRISPR-Cas system for editing plant genomes. The study also includes information on the approaches, procedural programs and applications in editing plant genomes for improving resistance against emerging pathogens, crop yield, herbicide tolerance and abiotic stresses. Some of the platforms that are being used by CRISPR-Cas systems include DNA endonuclease-targeted CRISPR trans reporter or DETECTR (Chen et al., 2018), Cas13-assisted restriction of viral expression and readout or CARVER (Freije et al., 2019), 1-h low-cost multipurpose highly efficient system or HOLMES (Li et al., 2019) and specific high sensitivity enzymatic reporter unlocking or SHERLOCK (Gootenberg et al., 2017). In their study, Lyu et al. (2020) have highlighted the potential of CRISPR platforms as a tool for diagnosing tuberculosis in children. They have recommended further studies to evaluate the performance of CRISPR in non-invasive specimens collected from children. In (Kayesh et al., 2020), Kayesh et al. have used CRISPR-Cas9 systems to target Hepatitis B. In this regard, they have used systems which target HBsAg, HBV DNA, and HBV cccDNA and investigate the potential of virus-based vectors as a suitable delivery system. They have further designed 16 sets of HBV-specific gRNAs which target different conserved regions of the HBV genome of HBV genotype C. The feasibility and efficiency of using CRISR-based methodologies have been explored in (King and Munger, 2019) to engineer human cytomegalovirus (HCV).

From all the aforementioned works, it can be said that CRISPR-Cas is a well established system for the rapid detection of viruses and therefore can be employed for the detection of SARS-CoV-2, MERS-CoV and SARS-CoV-1 as well. In this regard, Zhang et al. (2020) have recently reported the basic framework of specific high sensitivity enzymatic reporter unlocking or SHERLOCK which utilises CRISPR-Cas13 technique for SARS-CoV-2 detection. Cas13 can identify previously determined target sequence and binds to them leading to cleavage of surrounding single-strand RNA (ssRNA) molecules. SHERLOCK uses a quenched fluorescent ssRNA reporter. The presence of such reporters activates Cas13 leading to quantifiable signals. Amplification of targeted DNA or RNA by Recombinase Polymerase Amplification (RPA) or reverse transcriptase-RPA (RTRPA) before a reaction improves the sensitivity of the assay. Thereafter, amplified DNA gets transformed to RNA by the combination of RPA and T7 transcription. Finally, by simultaneous incorporation of the ssRNA reporter (Biotin-RNA-FITC), the virus is detected. Furthermore, Broughton et al. (2020a) have proposed DNA Endonuclease-Targeted CRISPR Trans Reporter or DETECTR to report the development and validation of CRISPR-Cas12 based assay for the detection of SARS-CoV-2. Simultaneous reverse transcription and isothermal amplification using loop-mediated amplification (RTLAMP) is performed for the extracted RNA of the virus using this assay. Thereafter, predefined coronavirus sequences are defined using Cas12 protein followed by which cleavage of a reporter molecule verifies the virus detection. Although the above techniques are quite advantageous, genome-wide analysis of the virus for different lengths of k and palindromes have not yet been explicitly reported.

Taking cues from these recent works, we have adopted the concept of CRISPR-Cas system to identify the target sites for the identification of SARS-CoV-2 and other viruses of Coronaviridae family, that is MERS-CoV and SARS-CoV-1. In this regard, identification of protospacer adjacent motif or PAM is carried out in this work. PAM is a short DNA sequence having usually a length of about 3–6 nt that is present adjacent to CRISPR in the genomic sequence. The genomic locations that are the potential target sites for the identification of viruses are limited by the presence and locations of the PAM. Thus, in order to find the target sites for the identification of SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses, initially the PAM and their corresponding genomic locations are identified. Once the PAM are identified, instead of finding short palindromic repeats as required by CRISPR-Cas, we have modified the idea to consider palindromic sequences which are adjacent to PAM to be the target sites for virus identification. Thereafter, to bind and cut the target sites, specific PAMs are identified for the RNA-guided DNA Cas9/Cas12 endonuclease. In this regard, PAMs such as 5’-TGG-3’ and 5’-TTTA-3’ in NSP3 and Exon for SARS-CoV-2, 5’-GGG-3’ and 5’-TGG-3’ in Exon and NSP2 for MERS-CoV and 5’-AGG-3’ and 5’-TTTG-3’ in Helicase and NSP3 respectively for SARS-CoV-1 are identified corresponding to SpCas9 and FnCas12a. It is worth mentioning that studies performed by Cain et al. (2001), Dirac et al. (2002), Chew et al. (2004) have suggested that palindromes can be considered to be involved in target identification, viral packaging and defence mechanisms. A palindromic sequence is a symmetrical sequence so that when read from the reverse direction, it is the exact complement of itself. For example, TGCA is a palindrome of length 4. It is to be noted that a palindrome is always even in length. Thereafter, to recognise these target sites in a virus genome as cleaved by SpCas9 and FnCas12a, primers are designed as complementary to the target site sequences. Thus, these complementary palindromic primers can be considered in assays for the rapid identification of SARS-CoV-2, MERS-CoV and SARS-CoV-1. These primers are akin to guide RNA (gRNA) in CRISPR-Cas technology.

2. Materials and methods

To find PAM and the corresponding palindromic sequences, initially the three reference genomic sequences of SARS-CoV-2 (NC_045512.2),[3] MERS-CoV (NC_019843.3)[4] and SARS-CoV-1 (NC_004718.3)[5] are collected from NCBI (National Center for Biotechnology Information). Once the palindrome and PAM combinations for each virus are identified on their respective reference sequence, their presence or coverage is verified for 108246, 291 and 340 virus sequences of SARS-CoV-2, MERS-CoV and SARS-CoV-1 respectively. SARS-CoV-2 virus sequences are collected from Global Initiative on Sharing All Influenza Data (GISAID)[6] in fasta format while 291 and 340 virus sequences of MERS-CoV[7] and SARS-CoV-1[8] are downloaded from NCBI. It is to be noted that 108246 SARS-CoV-2 sequences are considered up to September 2021 after performing filtering for considering only complete SARS-CoV-2 genomes. For docking purposes, P3DOCK Server[9] is used and PyMOL[10] is considered for visualization. For such docking, the PDB ID of PAMs like FnCas12a is 5NG6 and that of SpCas9 is 4O08 while the corresponding palindromes of SARS-CoV are designed with the help of Chimera v.1.15 RNA/DNA build tool.

Algorithm 1 presents the Palindrome-PAM finding method in details. To identify PAMs ($\mathcal{P}$), the reference sequence ρ for each of the viruses is initially divided into patterns of sequences of length k using the popular k-mer technique. Thereafter, these k-mers are searched for in the reference sequence to learn about their corresponding genomic locations in the *FindPAMpositions* step. These sequences of length k are considered as PAMs in this work. Subsequently, all these k-mer PAMs are used to identify the palindromic sequences in the reference sequence. Based on the intended palindromic sequence length η, some calculations are performed to find the starting and ending genomic locations of the target site of a virus. Since, the palindromic sequence and PAM are adjacent, the ending (*end_loc_pal*) and the starting locations of the palindrome

[3] https://www.ncbi.nlm.nih.gov/nuccore/1798174254.

[4] https://www.ncbi.nlm.nih.gov/nuccore/667489388.

[5] https://www.ncbi.nlm.nih.gov/nuccore/30271926.

[6] https://gisaid.org/CoV2020.

[7] https://www.ncbi.nlm.nih.gov/genomes/VirusVariation/Database/nph-select.cgi?cmd=database&taxid=1335626.

[8] https://www.ncbi.nlm.nih.gov/labs/virus/vssi/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome-related%20coronavirus,%20taxid:694009.

[9] http://www.rnabinding.com/P3DOCK/P3DOCK.html.

[10] https://pymol.org/2/.

($strt\_loc\_pal$) are calculated as $end\_loc\_pal = strt\_loc\_PAM - 1$ and $strt\_loc\_pal = end\_loc\_pal - (\eta - 1)$ respectively. If these starting and ending locations in the reference sequence of a virus genome return a palindromic sequence in the *PalindromeCheck* step, then they are chosen as the target sites ($\mathcal{M}$) for the identification of the virus. Subsequently, all the virus genomic sequences are checked for the presence of the set of identified palindrome and PAM combinations (S_{PalPAM}) to check for the presence of these combinations (Population Coverage) in the *Coverage* step. This method for identification of the target site for the subsequent identification of virus is applied for SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses. The CRISPR-Cas9 technology for gene editing is shown in Fig. 1(a) while the proposed method for virus identification are shown in Fig. 1(b).

Algorithm 1. Pseudo-code for palindromic target site identification

```
Input  : $\rho$ (reference sequence of virus genome), $k$ (k-mer), $\eta$ (length of palindrome), $\chi$ (all virus genomic sequences)
Output: $S_{PalPAM}$ (set of Palindrome and PAM combination) ,$\theta$ (Poulation Coverage)
begin
    Initialize a NULL Set, $S_{PalPAM}$
    $\mathcal{P} \leftarrow kmercount(\rho, k)$
    $\lambda \leftarrow FindPAMpositions(\mathcal{P})$
    $\alpha \leftarrow numberofrows(\lambda)$
    $\beta \leftarrow numberofcolumns(\lambda)$
    for $i \leftarrow 1$ to $\alpha$ do
        for $j \leftarrow 1$ to $\beta$ do
            $end\_loc\_pal = strt\_loc\_PAM - 1$
            $strt\_loc\_pal = end\_loc\_pal - (\eta - 1)$
            $\mathcal{M} \leftarrow PalindromeCheck(strt\_loc\_pal, end\_loc\_pal)$
            $PalPAM \leftarrow \mathcal{M} \| \mathcal{P}$
            $S_{PalPAM} \leftarrow S_{PalPAM} \cup PalPAM$
        end
    end
    $\theta \leftarrow Coverage(\chi, S_{PalPAM})$
    return $S_{PalPAM}, \theta$
end
```

Once all the palindrome-PAM combinations are identified, PAMs like 5'-NGG-3'and 5'-TTTN-3' (N is any nucleotide base) specific for the RNA-guided DNA SpCas9 and FnCas12a endonucleases are identified out of all the combinations to bind and cut the target sites. Thereafter, SpCas9 and FnCas12a cleave the target regions for the final virus identification. Such virus identification is performed by the gRNAs which are the complementary primers of the palindromic sequences. For example, if SARS-CoV-2 is present in a sample collected from a person, the corresponding gRNA will attach itself to the fragment of the virus as cleaved by Cas9/Cas12 endonuclease. Such fragments can be then amplified thereby identifying the corresponding virus. Similar experiments can be performed for MERS-CoV and SARS-CoV-1 as well.

The time complexity of the proposed method for finding the palindrome-PAM combinations can be calculated as: let the length of the reference sequence be $\mathcal{N}$. Thus, the time complexity of finding PAM using *kmercount* is $O(\mathcal{N})$. To find the corresponding genomic locations of these PAM in *FindPAMpositions* step, the time taken is $O(log\mathcal{N})$. Computing the palindromic sequences takes $O(\alpha\beta)$ time, where α and β are the number of rows and columns respectively of the matrix returned by *FindPAMpositions* step. Finally, the computation of population coverage takes $O(\mathcal{XYD})$ time, where $\mathcal{X}$ and $\mathcal{Y}$ are the dimensions of the set which holds the palindrome-PAM combinations and $\mathcal{D}$ is the number of virus genomic sequences. Thus, the overall complexity of the proposed target site identification method is $O(\mathcal{N} + log\mathcal{N} + \alpha\beta + \mathcal{XYD})$.

3. Results

The results for the total number of PAM in the reference genomic sequence of each of SARS-CoV-2, MERS-CoV and SARS-CoV-1 are shown in Fig. 1(c), where the value of k is varied from 3 to 6. From the figure, we can see that with increasing k, the number of PAM is also increasing. For example, for SARS-CoV-2, with $k = 3$, 4, 5 and 6, number of PAM are 64, 256, 1023 and 3756 respectively. With these PAM, the corresponding palindromic sequences are identified in the reference genomic sequence following Algorithm 1. The results for the number of identified palindrome-PAM combinations in the reference sequences of SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses are depicted in Fig. 1 respectively for varying lengths of k and where the palindrome of even length is taken from 10 to 20. For example, 38 such combinations exist in the reference sequence of SARS-CoV-2 for PAM lengths of 3 to 6 and palindrome length of 10. For MERS-CoV, the identified palindrome-PAM combination sequences exist for lengths of 10 to 18 but not for 20 while for SARS-CoV-1 such combinations exist for all the palindromic lengths as considered in this work. It is to be noted that beyond the length of 20, no palindrome-PAM combinations exist for any of the virus sequences. Table 1 reports the percentage of coverage of the identified palindrome-PAM combinations in 108246, 291 and 340 SARS-CoV-2, MERS-CoV

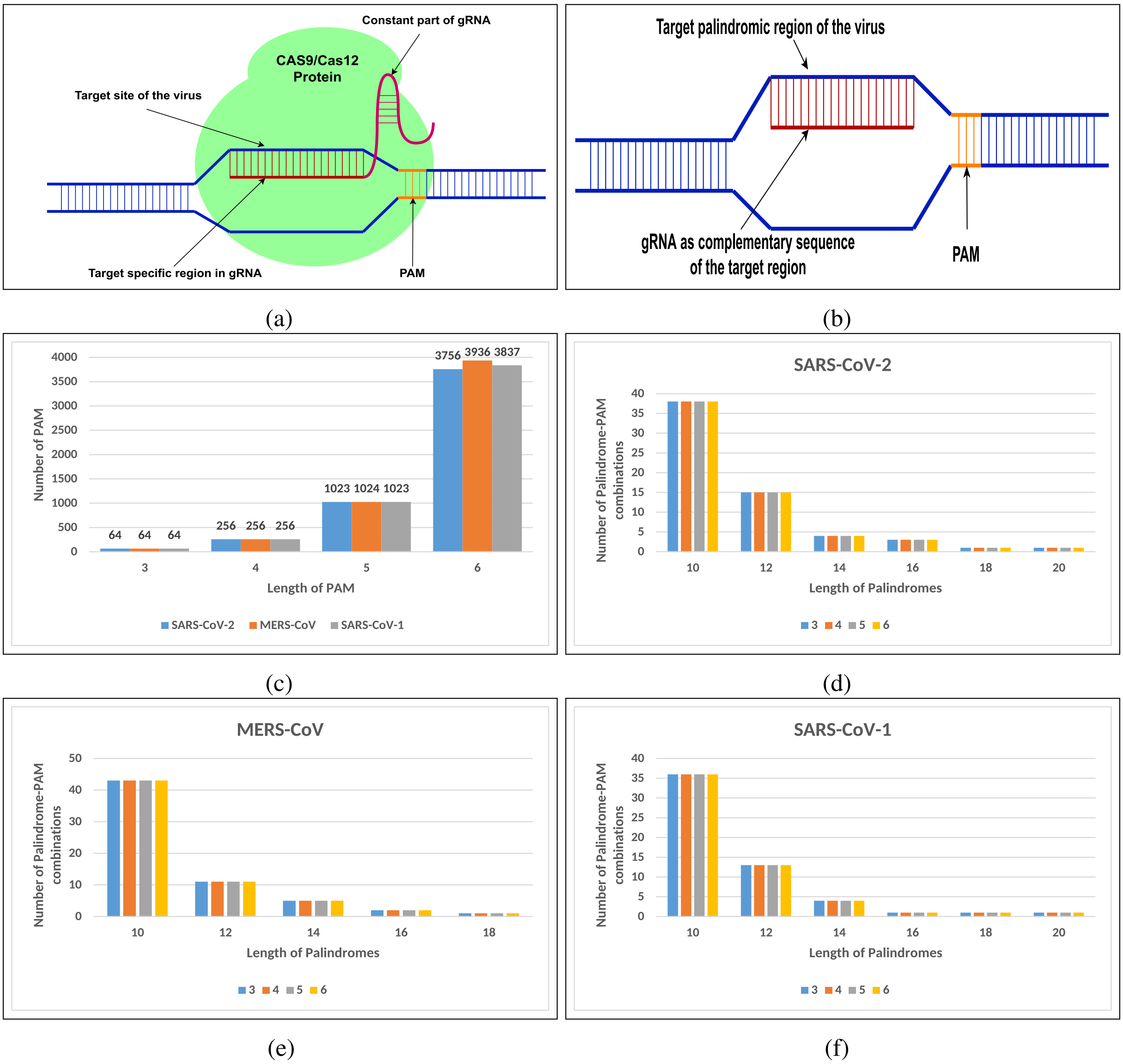


Fig. 1. (a) CRISPR-Cas9 gene editing system (b) Proposed target site identification technique (c) Number of PAM in reference sequence for varying length of *k* and Number of palindrome-PAM combinations for (d) SARS-CoV-2 (e) MERS-CoV and (f) SARS-CoV-1.

and SARS-CoV-1 virus sequences respectively for palindromic lengths of 14 to 20. For all the palindromic sequences, the results are provided in the supplementary as an excel file. For example, for SARS-CoV-2 with $k = 6$ and target palindromic sequence length of 20, palindrome-PAM combination of ACACTGGTAATTACCAGTGT and GGTCAC is present in 108048 out of 108246 virus sequences, i.e. this combination is present in 99.81% of the sequences. This coverage percentage shows the validity of the identified palindrome-PAM combinations for the respective virus genomic sequences. The results are similarly reported for MERS-CoV and SARS-CoV-1. Table 1 also reports the corresponding GC content of the palindromes. According to (Reynolds et al., 2004; Haeussler et al., 2016), it is difficult to target GC-rich genes and thus it can be said that sequences with moderate GC-content are good candidates for being target sites of a virus. As can be observed from Table 1, the GC content of the identified target sites for each of SARS-CoV-2, MERS-CoV and SARS-CoV-1 are quite moderate. For example, for $k = 6$ and target palindromic sequence length of 20, the target palindromic site of SARS-CoV-2 has a GC content of 40%.

Once all the palindrome-PAM combinations are identified for all lengths of palindrome and PAM, specific PAMs which are recognised by either SpCas9 or FnCas12a endonuclease for cleaving the target sites are further identified. These palindrome-PAM combinations are specifically important for the identification of the viruses. Out of the total 248 unique palindrome-PAM combinations for SARS-CoV-2, 2 such combinations are identified corresponding to SpCas9 and FnCas12a, for MERS-CoV, out of 248 combinations, 3 such combinations are identified while for SARS-CoV-1 2 such combinations out of 224 are identified. Table 2 reports such PAMs along with the corresponding palindromes and their

Table 1
Population Coverage and GC Content of the target sites for SARS-CoV-2, MERS-CoV and SARS-CoV-1.

Virus	Length of Palindrome	Length of PAM							
		3		4		5		6	
		Population Coverage (%)	GC Content (%)	Population Coverage (%)	GC Content (%)	Population Coverage (%)	GC Content (%)	Population Coverage (%)	GC Content (%)
SARS-CoV-2	14	99.89	42.85	99.88	42.85	99.79	42.85	90.81	0
		99.80	42.85	70.37	14.28	99.88	42.85	99.88	42.85
		70.37	14.28	99.80	42.85	91.63	0	99.79	42.85
		91.93	0	91.68	0	70.34	14.28	70.33	14.28
	16	70.37	25	70.34	25	99.84	37.5	70.32	25
		99.85	37.5	99.79	37.5	70.33	25	99.79	37.5
		99.79	37.5	99.85	37.5	99.79	37.5	99.84	37.5
	18	99.82	44.44	99.82	44.44	99.82	44.44	99.82	44.44
	20	99.81	40	99.81	40	99.81	40	99.81	40
MERS-CoV	14	99.65	57.14	99.65	57.14	99.65	57.14	99.65	57.14
		100	14.28	100	14.28	100	14.28	98.62	14.28
		28.52	28.57	28.52	28.57	98.62	14.28	100	14.28
		99.65	28.57	98.62	14.28	28.52	28.57	28.52	28.57
		98.62	14.28	99.65	14.28	99.65	28.57	99.65	28.57
	16	98.62	25	98.62	25	98.62	25	98.62	25
		28.52	25	28.52	25	28.52	25	28.52	25
	18	28.52	33.33	28.52	33.33	28.52	33.33	28.52	33.33
		92.64	14.28	92.64	14.28	82.64	42.85	92.05	14.28
SARS-CoV-1	14	91.17	28.57	90.58	28.57	92.64	14.28	82.64	42.85
		82.64	42.85	82.64	42.85	87.05	28.57	71.47	0
		92.05	0	91.76	0	90.58	0	87.05	28.57
	16	87.05	25	87.05	25	87.05	25	87.05	25
	18	87.05	22.22	87.05	22.22	87.05	22.22	87.05	22.22
	20	87.05	30	87.05	30	87.05	30	87.05	30

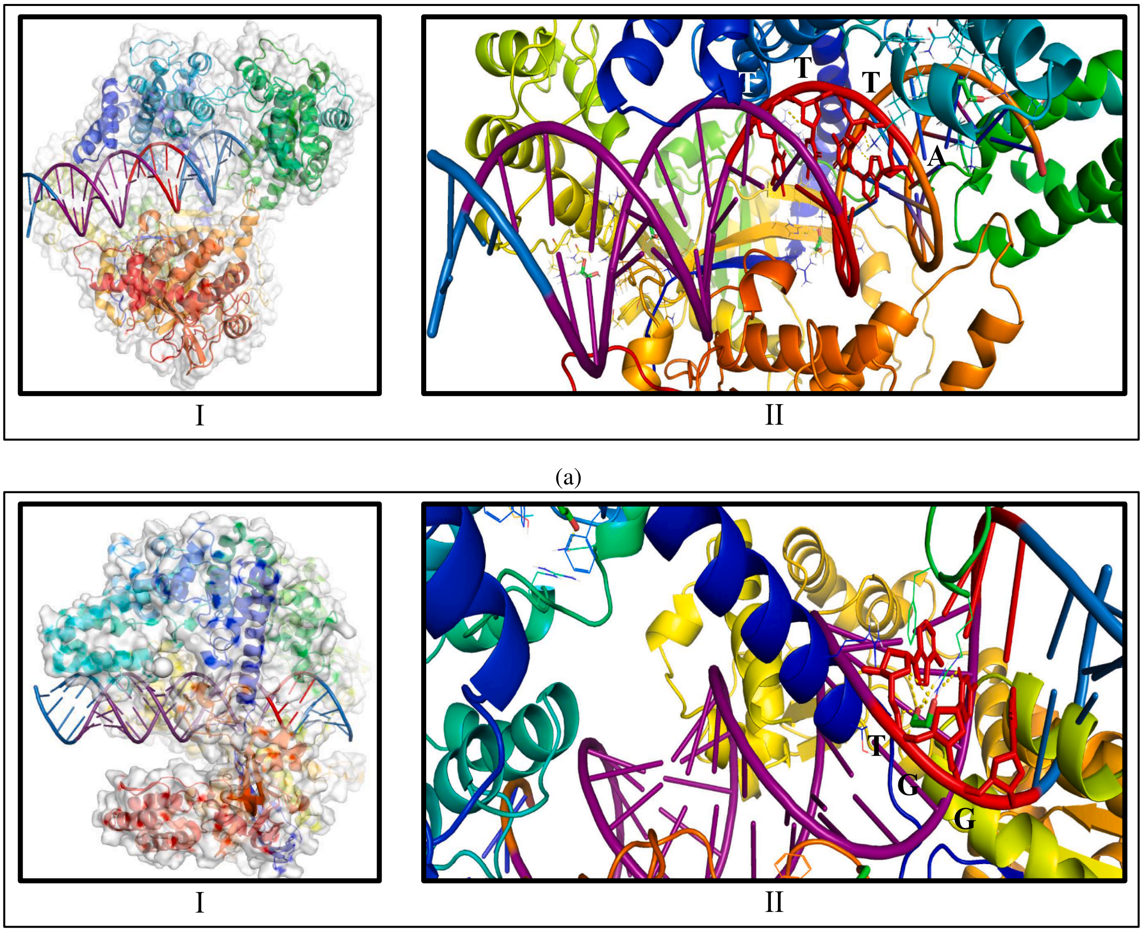
Table 2
Combination of Palindrome and PAM in SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses along with the complementary primer.

Virus	Palindrome as Target	Length of Palindrome	Start Coordinate of Target	PAM	Start Coordinate of PAM	Complementary Primer (gRNA)	Coding Gene	Population Coverage (%)	GC Content (%)
SARS-CoV-2	TTGGTACCAA	10	18341	TTTA	18351	AACCATGGTT	Exon	99.71	40
	CACTGGTAATTACCAGTG	18	5746	TGG	5764	GTGACCATTAATGGTCAC	NSP3	99.82	44.44
MERS-CoV	TTATGCATAA	10	18456	GGG	18466	AATACGTATT	Exon	100	20
	ATCTATATAGAT	12	1018	TGG	1030	ATCTATATAGAT	NSP2	98.62	16.66
	CTTATGCATAAG	12	18455	GGG	18467	GAATACGTATTC	Exon	99.31	33.33
SARS-CoV-1	ACACATGTGT	10	16450	AGG	16460	TGTGTACACA	Helicase	94.41	40
	TAACAATTGTTA	12	5208	TTTG	5220	ATTGTTAACAAT	NSP3	92.64	16.67

associated details along with the gRNA, coding gene, percentage of coverage and GC content. For example, for identifying SARS-CoV-2, TTTA is a PAM in Exon needed to guide FnCas12a endonuclease for binding to and cleaving TTGGTACCAA with the help of gRNA AACCATGGTT as a complementary primer. Such palindrome-PAM is present in 99.71% of the total number of sequences while the GC content of the palindrome is 40. On the other hand, TGG is a PAM in NSP3 for SpCas9 for cleaving CACTGGTAATTACCAGTG with GTGACCATTAATGGTCAC and this combination is present in 99.82% of the sequences with the GC content of the palindrome being 44.44%. Similarly, PAMs corresponding to SpCas9 and FnCas12a are also identified for MERS-CoV and SARS-CoV-1 and reported in Table 2. The corresponding target sequences as DNA in SpCas9 and FnCas12a along with the Cas Binding with PAM in order to cleave the target region are shown in Figs. 2–4. In each figure, (I) shows the palindrome-PAM combinations with Cas Protein and (II) shows the binding of PAM with the Cas protein for cleaving the target site. For example, Fig. 2(a)(I) shows the palindromic sequence TTGGTACCAA in purple and PAM TTTA in red and (II) shows the binding of TTTA with FnCas12a. This binding is a proof that indeed FnCas12a can recognise TTTA and cleave the target region TTGGTACCAA.

Please note that all the palindrome-PAM combinations are unique to each virus, thereby confirming the fact that they can indeed be used for virus identification. Furthermore, we have also checked for the combinations from Table 2 in the reference sequence of Ebola, Dengue, Influenza and Zika viruses. Also, nucleotide BLAST[11] is used to check the specificity of the same and it has been observed that such palindrome-PAM combinations are not present in any of the other viruses. Apart from aforementioned results, all the palindrome-PAM combinations are provided in the supplementary as an excel file. It is to be further noted that though this work specifically focuses on PAMs as recognised by Cas9 or Cas12 endonuclease, we have reported other palindrome-PAM combinations as well in the hope that if any new endonuclease is engineered, our work can serve as a way for further virus identification.

[11] https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.



(b)

Fig. 2. (I) Palindrome-PAM combinations with Cas Proteins and (II) Cas binding with PAM for cleaving the target site in SARS-CoV-2 for (a) FnCas12a (b) SpCas9. In the figures, purple represents the palindromic sequence and red represents the PAM sequence.

4. Conclusion

This study adopts the idea of CRISPR-Cas technology in order to identify palindrome-PAM combinations as target sites for virus identification. To achieve this, initially PAMs are identified using *k*-mer technique. Thereafter, palindromic sequences which are adjacent to the PAM locations are identified as the potential target sites. Next, PAMs specific for the RNA-guided DNA Cas9/Cas12 endonuclease to bind and cut the target sites are detected. In this regard, corresponding to SpCas9 and FnCas12a endonuclease, PAMs such as 5’-TGG-3’ and 5’-TTTA-3’ in NSP3 and Exon for SARS-CoV-2, 5’-GGG-3’ and 5’-TGG-3’ in Exon and NSP2 for MERS-CoV and 5’-AGG-3’ and 5’-TTTG-3’ in Helicase and NSP3 respectively for SARS-CoV-1 are identified. Finally, gRNAs as primers complementary to the identified target sites are designed to recognise these target sites in a virus genome. The palindrome-PAM combinations are initially identified in the corresponding reference sequences of SARS-CoV-2, MERS-CoV and SARS-CoV-1. By varying the length of *k* in *k*-mer technique and by considering different length of palindromes, the highest number of palindrome-PAM combinations for each of SARS-CoV-2, MERS-CoV and SARS-CoV-1 are found to be 38, 43 and 36 respectively for $k = 3$ to 6 and palindrome length = 10. The palindrome-PAM combinations suitable for CAS binding in SARS-CoV-2 are TTGGTACCAATTTA in Exon and CACTGGTAATTACCAGTGTGG in NSP3. For MERS-CoV such combinations are TTATGCATAAGGG in Exon, ATCTATATAGATTGG in NSP2 and CTTATGCATAAGGGG in Exon while for SARS-CoV-1 such combinations are ACACATGTGTAGG in Helicase and TAACAATTGTTAGGG in NSP3. These palindrome-PAM combinations when searched for in 108246, 291 and 340 SARS-CoV-2, MERS-CoV and SARS-CoV-1 virus sequences respectively, they are present in 99.71%, 99.82%, 100%, 98.62%, 99.31%, 94.41% and 92.64% of the virus sequences. Furthermore, the GC content of these identified target sites are evaluated to judge their candidature. The method proposed in this work can be deemed to be a very efficient and quick way to detect SARS-CoV-2, MERS-CoV and SARS-CoV-1 where the gRNAs as primers complementary to the target sites along with the respective Cas proteins can be considered in assays for the identification method.

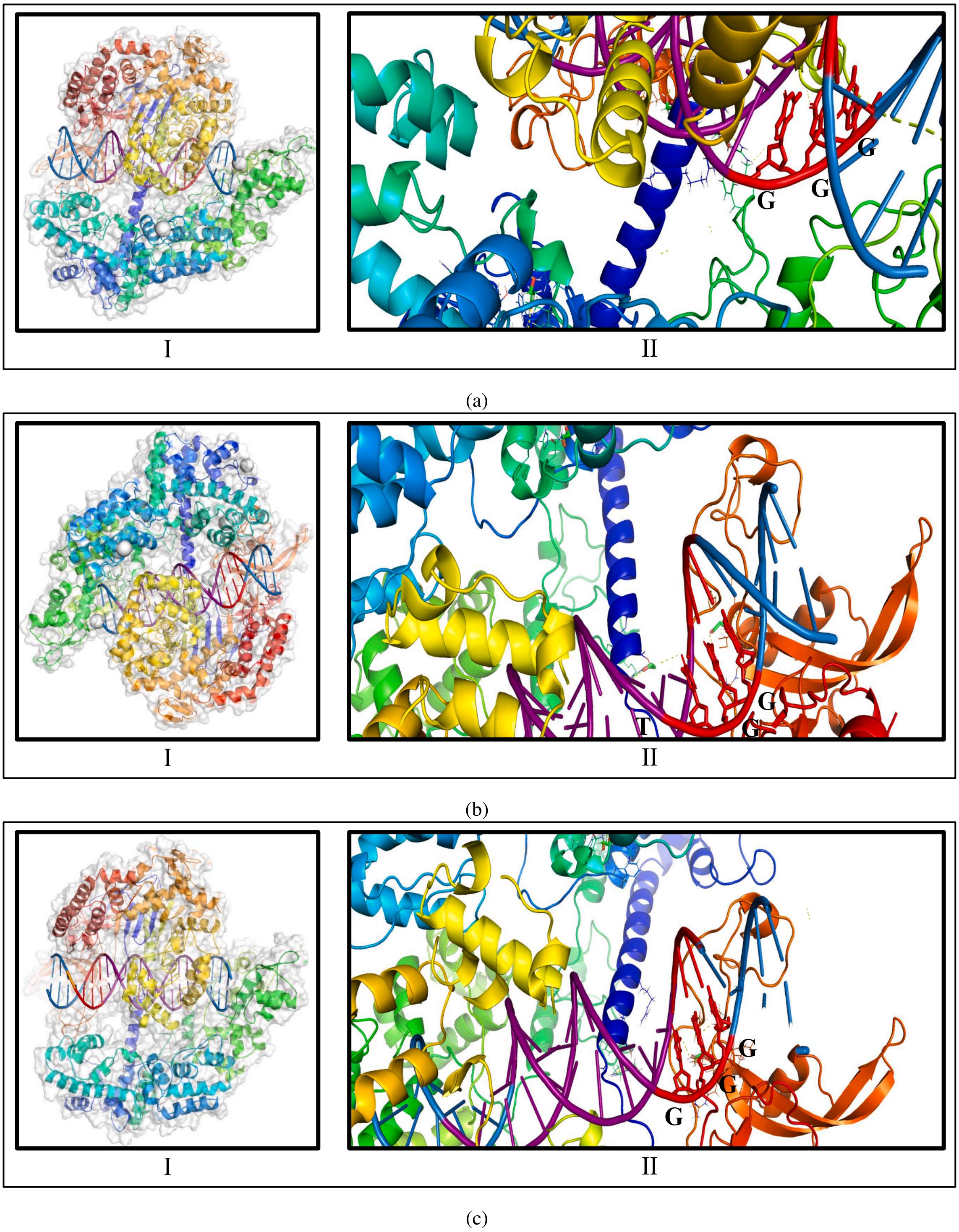
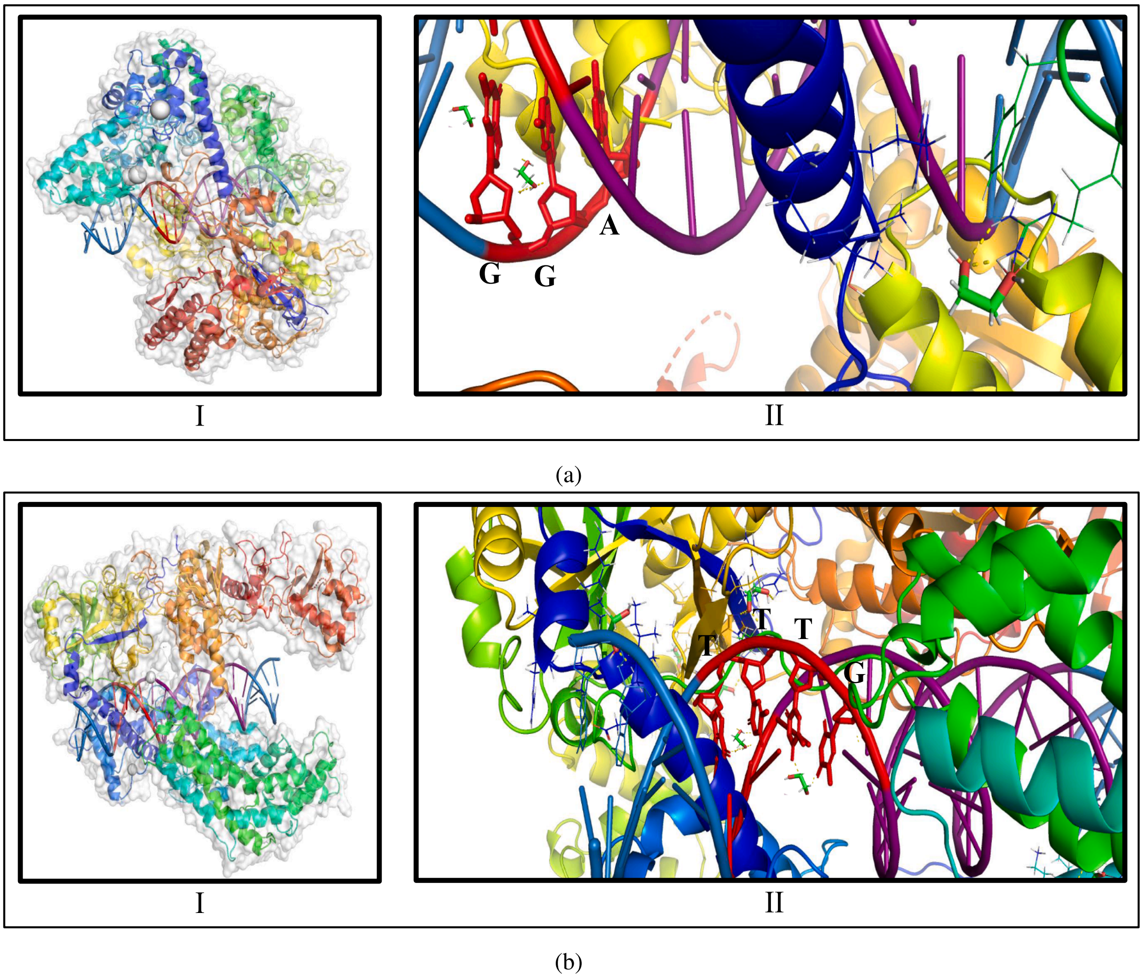


Fig. 3. (I) Palindrome-PAM combinations with Cas Proteins and (II) Cas binding with PAM for cleaving the target site in MERS-CoV for (a) (b) and (c) SpCas9. In the figures, purple represents the palindromic sequence and red represents the PAM sequence.



(a)

(b)

Fig. 4. (I) Palindrome-PAM combinations with Cas Proteins and (II) Cas binding with PAM for cleaving the target site in SARS-CoV-1 for (a) SpCas9 and (b) FnCas12a. In the figures, purple represents the palindromic sequence and red represents the PAM sequence.

Furthermore, apart from the Cas recognised PAMs we have provided a list of other palindrome-PAM combinations as well for all the three viruses in the hope that if in the future some other endonucleases are engineered, such combinations may be further used for successful virus identification. Moreover, the in vitro verification of the identified palindrome-PAM combinations may also be further investigated.

Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

Availability of data and materials

All the SARS-CoV-2, MERS-CoV and SARS-CoV-1 virus genomes with their corresponding reference sequences and the final results of this work are available at "http://www.nitttrkol.ac.in/indrajit/projects/COVID-CRISPR-Cas/".

Funding

This work has been partially supported by CRG short term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India.

Author contributions

Nimisha Ghosh: Conceptualization; Methodology; Visualization; Writing – original draft, **Indrajit Saha:** Conceptualization; Data curation; Supervision; Funding acquisition; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Writing – review & editing, **Nikhil Sharma:** Methodology; Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Program awarded to Nimisha Ghosh. Also, the authors would like to thank all those who have contributed sequences to GISAID and NCBI databases.

References

Bhat, M.A., Bhat, M.A., Kumar, V., Wani, I.A., Bashir, H., Shah, A.A., Rahman, S., Jan, A. T., 2020. The era of editing plant genomes using crispr/cas: A critical appraisal. J. Biotechnol. 324, 34–60. https://doi.org/10.1016/j.jbiotec.2020.09.013.

Broughton, J.P., Deng, X., Yu, G., Fasching, C.L., Servellita, V., Singh, J., Miao, X., Streithorst, J., Granados, A., Sotomayor-Gonzalez, A., Zorn, K., Gopez, A., Hsu, E., Gu, W., Miller, S., Pan, C., Guevara, H., Wadford, D.A., Chen, J.S., Chiu, C.Y., 2020a. Crisprcas12-based detection of sars-cov-2. Nat. Biotechnol. 38, 870–874. https://doi.org/10.1101/2020.03.06.20032334.

Broughton, J.P., Deng, X., Yu, G., Fasching, C.L., Singh, J., Streithorst, J., Granados, A., Sotomayor-Gonzalez, A., Zorn, K., Gopez, A., Hsu, E., Gu, W., Miller, S., Pan, C., Guevara, H., Wadford, D.A., Chen, J.S., Chiu, C.Y., 2020b. Rapid detection of 2019 novel coronavirus sars-cov-2 using a crispr-based detectr lateral flow assay. medRxiv: the preprint server for health sciences, doi: https://doi.org/10.1101/2020.03.06.20032334.

Cain, D., Erlwein, O., Grigg, A., Russell, R., Mcclure, M., 2001. Palindromic sequence plays a critical role in human foamy virus dimerization. J. Virol. 75, 3731–3739. https://doi.org/10.1128/JVI.75.8.3731-3739.2001.

Chen, S.J., Ma, E., Harrington, B.L., Costa, M.D., Tian, X., Palefsky, J.M., Doudna, A.J., 2018. Crispr-cas12a target binding unleashes indiscriminate single-stranded dnase activity. Science 360, 436–439. https://doi.org/10.1126/science.aar6245.

Chew, D., Choi, K.P., Heidner, H., Leung, M.-Y., 2004. Palindromes in sars and other coronaviruses. INFORMS J. Comput. 16, 331–340. https://doi.org/10.1287/ijoc.1040.0087.

Dey, A., Sen, S., Maulik, U., 2020. Unveiling COVID-19-associated organ-specific cell types and cell-specific pathway cascade. Briefings Bioinformat. https://doi.org/10.1093/bib/bbaa214.

Dirac, A.M.G., Huthoff, H., Kjems, J., Berkhout, B., 2002. Requirements for rna heterodimerization of the human immunodeficiency virus type 1 (hiv-1) and hiv-2 genomes. J. Gen. Virol. 83, 3731–3739. https://doi.org/10.1099/0022-1317-83-10-2533.

Freije, C.A., Myhrvold, C., Boehm, C.K., Lin, A.E., Welch, N.L., Carter, A., Metsky, H.C., Luo, C.Y., Abudayyeh, O.O., Gootenberg, J.S., Yozwiak, N.L., Zhang, F., Sabeti, P.C., 2019. Programmable inhibition and detection of rna viruses using cas13. Mol. Cell 76, 826–837. https://doi.org/10.1016/j.molcel.2019.09.013 e11.

Gootenberg, S.J., Abudayyeh, O.O., Lee, J.W., Essletzbichler, P., Dy, A.J., Joung, J., Verdine, V., Donghia, N., Daringer, N.M., Freije, C.A., Myhrvold, C., Bhattacharyya, R.P., Livny, J., Regev, A., Koonin, E.V., Hung, D.T., Sabeti, P.C., Collins, J.J., Zhang, F., 2017. Nucleic acid detection with crispr-cas13a/c2c2. Science 356, 438–442. https://doi.org/10.1126/science.aam9321.

Haeussler, M., Schnig, K., Eckert, H., Eschstruth, A., Miann, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., Joly, J.-S., Concordet, J.-P., 2016. Evaluation of off-target and on-target scoring algorithms and integration into the guide rna selection tool crispor. Nat. Biotechnol. 17 https://doi.org/10.1186/s13059-016-1012-2.

Hosseini, E.S., Kashani, N.R., Nikzad, H., Azadbakht, J., Bafrani, H.H., Kashani, H.H., 2020. The novel coronavirus disease-2019 (covid-19): Mechanism of action, detection and recent therapeutic strategies. Virology 551, 1–9. https://doi.org/10.1016/j.virol.2020.08.011.

Jia, F., Li, X., Zhang, C., Tang, X., 2020. The expanded development and application of crispr system for sensitive nucleotide detection. Protein Cell 11, 624–629. https://doi.org/10.1007/s13238-020-00708-8.

Kayesh, M.E.H., Amako, Y., Hashem, M.A., Murakami, S., Ogawa, S., Yamamoto, N., Hifumi, T., Miyoshi, N., Sugiyama, M., Tanaka, Y., Mizokami, M., Kohara, M., Tsukiyama-Kohara, K., 2020. Development of an in vivo delivery system for crispr/cas9-mediated targeting of hepatitis b virus cccdna. Virus Res. 290, 198191. https://doi.org/10.1016/j.virusres.2020.198191.

King, M.W., Munger, J., 2019. Editing the human cytomegalovirus genome with the crispr/cas9 system. Virology 529, 186–194. https://doi.org/10.1016/j.virol.2019.01.021.

Li, L., Li, S., Wu, N., Wu, J., Wang, G., Zhao, G., Wang, J., 2019. Holmesv2: A crispr-cas12b-assisted platform for nucleic acid detection and dna methylation quantitation. ACS Synthetic Biol. 8, 2228–2237. https://doi.org/10.1021/acssynbio.9b00209.

Lyu, C., Shi, H., Cui, Y., Li, M., Yan, Z., Yan, L., Jiang, Y., 2020. Crispr-based biosensing is prospective for rapid and sensitive diagnosis of pediatric tuberculosis. Int. J. Infectious Dis. 101, 183–187. https://doi.org/10.1016/j.ijid.2020.09.1428.

Qi, F., Qian, S., Zhang, S., Zhang, Z., 2020. Single cell rna sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. Biochem. Biophys. Res. Commun. 526, 135–140. https://doi.org/10.1016/j.bbrc.2020.03.044.

Rahman, M.R., Hossain, M.A., Mozibullah, M., et al., 2021. CRISPR is a useful biological tool for detecting nucleic acid of SARS-CoV-2 in human clinical samples. Biomed. Pharmacotherapy 140, 111772. https://doi.org/10.1016/j.biopha.2021.111772.

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., Khvorova, A., 2004. Rational sirna design for rna interference. Nat. Biotechnol. 22, 326–330. https://doi.org/10.1038/nbt936.

Zhang, F., Abudayyeh, O.O., Gootenberg, J.S., 2020. A protocol for detection of covid-19 using crispr diagnostics. In: A protocol for detection of COVID-19 using CRISPR diagnostics, p. 8.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C., Chen, H., Chen, J., Luo, Y., Guo, H., Jiang, R., Liu, M., Chen, Y., Shen, X., Wang, X., Zheng, X., Zhao, K., Chen, Q., Deng, F., Liu, L., Yan, B., Zhan, F., Wang, Y., Xiao, G., Shi, Z., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273. https://doi.org/10.1038/s41586-020-2012-7.