Chapter 14

PCR Primer Design for the Rapidly Evolving SARS-CoV-2 Genome

Wubin Qu, Jiangyu Li, Haoyang Cai, and Dongsheng Zhao

Abstract

Real-time quantitative PCR is currently the most widely used method for the human pathogen severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) identification. Due to the rapid evolution of the SARS-CoV-2 genome, novel mutations on the primer binding sites will cause the failure of PCR. Therefore, in addition to a well-designed primer set, these primers need to be updated and evaluated regularly to ensure that the rapidly evolving genome primers can be amplified. In this protocol, (1) we firstly use assembled genome sequences in the SARS-CoV-2 database to identify and characterize indels and point mutations; (2) design primers skipping the sites of mutations; (3) check the coverage of the primers with the daily update SARS-CoV-2 database; (4) redesign them if novel mutations found in the primer binding sites. Although this protocol takes SARS-CoV-2 as an example, it is suitable for other species that have genomes accumulating mutations over time.

Key words SARS-CoV-2, Primer Design, Variants

1 Introduction

SARS-CoV-2 is an RNA virus with limited proofreading capability of correcting replication errors [1] and evolving continuously with new mutations [2, 3]. If these mutations are located in the binding regions of PCR primers, it will cause the failure [4] of PCR-based method [5]. Generally, we will use multiple sequence alignment methods [6] to find conserved regions and then design primers in the conserved regions. This method is effective in most cases. However, the SARS-CoV-2 is a rapidly evolving virus that mutates every day, which brings great challenges to multiple sequence alignment tasks. As of December 13, 2020, there are 30,645 complete new coronavirus genome sequences in the NCBI Virus database [7], which are updated daily.

With the advance of next-generation sequencing (NGS) technology, in addition to get the full genome sequence of SARS-CoV-2, there are a lot of free or open-source bioinformatics software

Chhandak Basu (ed.), *PCR Primer Design*, Methods in Molecular Biology, vol. 2392, https://doi.org/10.1007/978-1-0716-1799-1_14,

available to identify the variants of the genome [8, 9]. For example, MicroGMT [3] is a Python-based package, which takes either raw sequence reads or assembled genome sequence as input and compares against database sequences to identify and characterize small indels and point mutations in the microbial genomes. With the help of MicroGMT, we can use one of the sequences as a reference genome and compare other sequences to this reference genome. The inconsistent area is point mutations or indels. Then, the regions without point mutations or indels on the reference genome are suitable for primer design. In the following sections, (1) we firstly use assembled genome sequences in the SARS-CoV-2 database to identify and characterize indels and point mutations; (2) design primers skipping the sites of mutations; (3) check the coverage of the primers with the daily update SARS-CoV-2 database; (4) redesign them if novel mutations found in the primer binding sites. Although this protocol takes SARS-CoV-2 as an example, it is suitable for other species that have genomes accumulating mutations over time.

2 Materials

This protocol: (1) use MicroGMT [3] to identify the point mutations and indels; (2) use the famous Primer3 [10, 11] for primer design; (3) use the local command-line version of MFEprimer-3.0 [12] for primer evaluation. And the operating system is Linux (*see* **Note 1**) with minimum disk size 40Gb and minimum memory 8Gb. Create a directory named "SARS-CoV-2" as a working directory for this protocol and two subdirectories "bin" and "data." Python3 (https://www.python.org/) is also required.

2.1 Prepare SARS-CoV-2, Human and Influenza Virus Genome Database

1. Download and prepare the SARS-CoV-2 genome database in FASTA format (*see* **Note 2**). Here, we download the sequences from the National Center for Biotechnology Information (NCBI) Virus database [7]. Visit the website. "https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/", click "Search by virus," enter "Severe acute respiratory syndrome coronavirus 2," and select option "complete" for "Nucleotide Completeness" from the left panel. Then, click the "Download" button, select "Nucleotide" for step 1, select "Download All Records" for step 2, and "Use default" for step 3. Then, rename the downloaded file to a name like "SARS-CoV-2-20200927-17111.fasta," which means there are 17,111 sequences and downloaded on September 27, 2020. Put the file "SARS-CoV-2-20200927-17111.fasta" into directory "data."
2. Download the FASTA format of the reference genome of SARS-CoV-2 from https://www.ncbi.nlm.nih.gov/nuccore/

NC_045512.2. And also place the file into the "data" directory.

3. Download the human genome database (hg19) from UCSC [13] http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz and unzip it into the "data" directory.
4. Download the influenza virus database [14] from https://ftp.ncbi.nih.gov/genomes/INFLUENZA/.

2.2 Install MicroGMT, Primer3, MFEprimer-3.0, and Other Software

1. Download the source code of MicroGMT from https://github.com/qunfengdong/MicroGMT/archive/master.zip and put the file into the "bin" directory and unzip the file.
2. MicroGMT requires minimap2 [9] software for sequence alignment. Change the working directory into the "bin," and run the command "curl -L https://github.com/lh3/minimap2/releases/download/v2.17/minimap2-2.17_x64-linux.tar.bz2 | tar -jxvf - ./minimap2-2.17_x64-linux/minimap2", then minimap2 will be downloaded and install into "bin/minimap2-2.17_x64-linux/" directory.
3. MicroGMT also requires bcftools [15] for calling variants. Download the source code from http://www.htslib.org/download/ and install it by following the instructions on the same page.
4. Install the latest version MFEprimer-3.0 from https://www.mfeprimer.com/mfeprimer-3.1/#2-command-line-version.
5. Install the latest version Primer3 from https://sourceforge.net/projects/primer3/.
6. Install "seqkit" [16] command for sequence manipulation from https://github.com/shenwei356/seqkit/releases.
7. Install "rush" command for parallel jobs from https://github.com/shenwei356/rush.
8. Install "bedtools" [17, 18] command for manipulating BED files from https://bedtools.readthedocs.io/en/latest/content/installation.html.
9. Install "blat" [19] command for sequence alignment from http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/blat/.
10. At last, add the path of minimap2 and bcftools into $PATH. If errors are arises during the installation, Bioconda [20] (*see* **Note 3**) can be used to install this bioinformatics software easily.

3 Methods

Change into the working directory "SARS-CoV-2" and create a "variants" in this directory to store the mutations file. And the directory tree should like Fig. 1.

3.1 Identify the Mutations Compared to the Reference Genome

1. In the working directory, run the command: "python3 bin/MicroGMT-master/sequence_to_vcf.py -r data/NC_045512.fa -i assembly -fs data/SARS-CoV-2-20200927-17111.fasta -o variants". This command will generate 17,111 files with the suffix ".vcf". Each VCF (*see* **Note 4**) file [21] contains the variants of this record compared to the reference genome. Figure 2 is one of the VCF files. Take the line in the red box as an example, this line means that the record "MW035376.1" has base "T" in position 241, while the reference genome NC_045512 has base "C" in the same position, indicating this position is a point mutation. Primers should not select from this region, at least the 3′ end of the primers should skip this region.

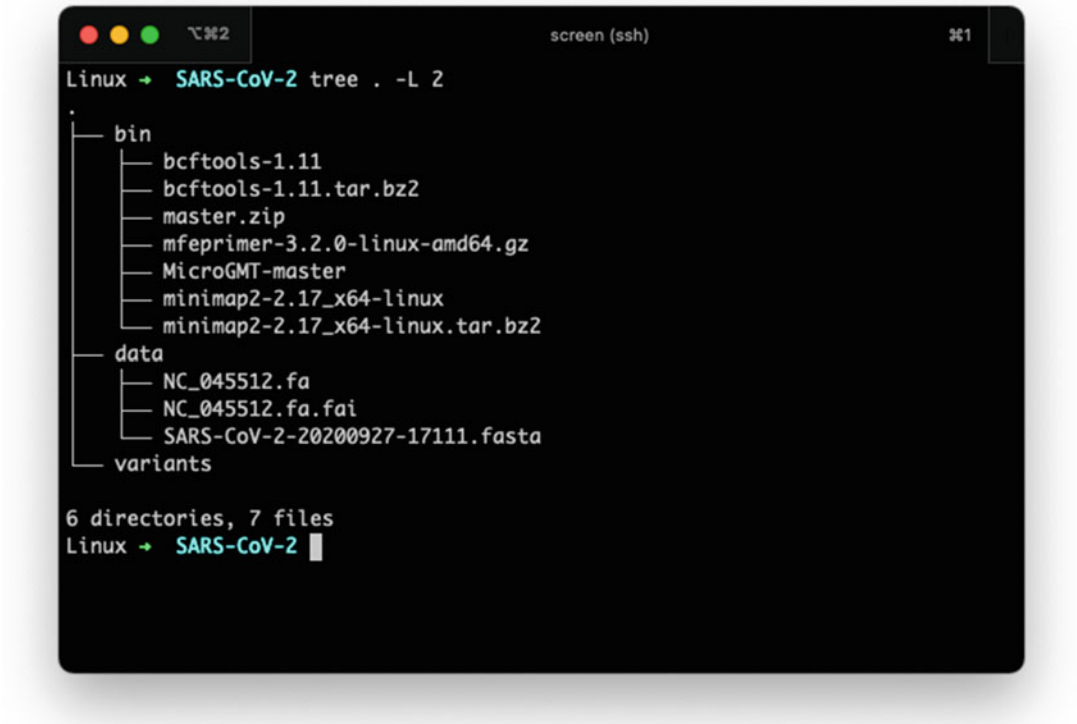

Fig. 1 Working directory structure

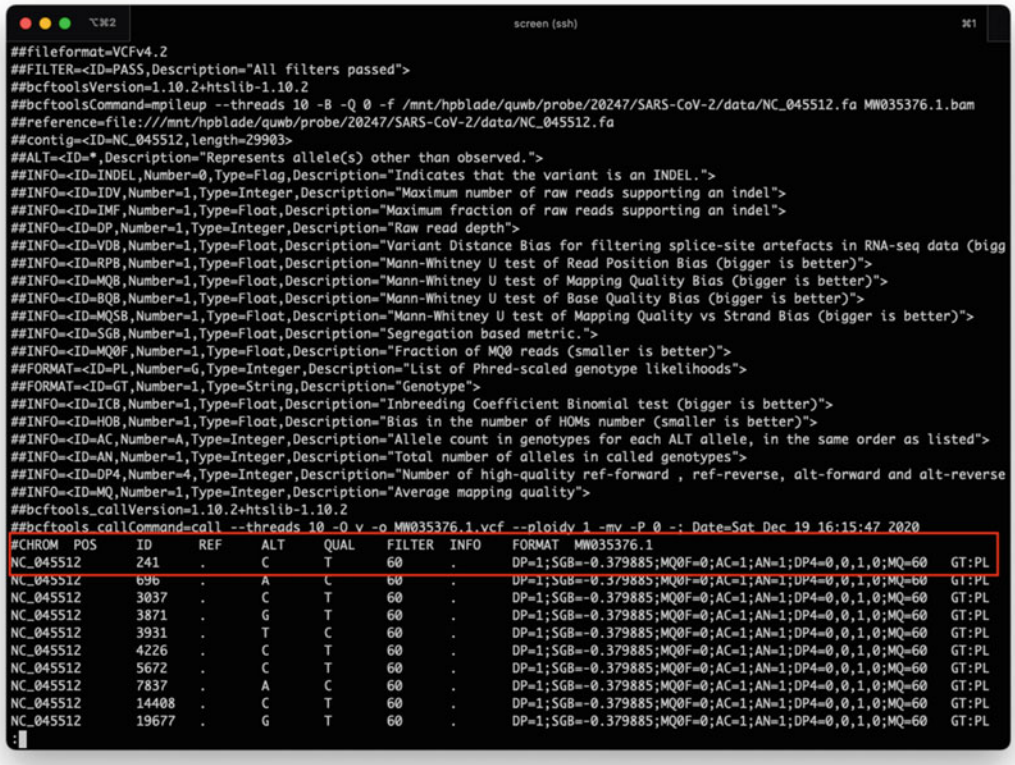

Fig. 2 VCF file example

2. Merge all VCF files into one with command: "find variants -name "*.vcf" | rush 'grep -v "#" {}s >> all.variants.vcf'", please ignore the error messages (means that no variants found for a certain record).
3. Convert VCF format to BED format with command: "awk '! /\#/' all.variants.vcf | awk '{if(length($4) > length($5)) print $1"\t"($2-1)"\t"($2+length($4)-1); else print $1"\t"($2-1)"\t"($2+length($5)-1)}' > all.variants.bed".
4. Sort and merge the variant file in BED format with command: "cat all.variants.bed | bedtools sort | bedtools merge > all.variants.merged.bed".

3.2 Design Primers

1. Create a "design" directory in the working directory "SARS-CoV-2."
2. Mask the reference genome sequences with variants file: bedtools maskfasta -soft -fi data/NC_045512.fa -bed all.variants.merged.bed -fo data/NC_045512.softmask.fa.
3. Merge the multiple lines into one line: awk '/^>/ {printf("\n%s\n",$0);next;} {printf("%s",$0);} END {printf("\n");}' < data/NC_045512.softmask.fa > data/NC_045512.softmask.one-line.fa.txt.

4. Change into the "design" directory, copy the Primer3 example input file into current directory: cp ../bin/primer3-2.4.0/example input.txt.
5. Edit the "input.txt" file: Add "PRIMER_ LOWERCASE _MASKING=1" (*see* **Note 5**) and "PRIMER_THERMODY NAMIC _P ARAMETERS _PATH=../bin/primer3–-2.4.0/src/primer3_config/", change "PRIMER_NUM_RETURN=100" and "PRIMER_PRODUCT_SIZE_RANGE=150-200" (*see* **Note 6**). Also replace the sample sequence with the sequence in file "../data/NC_045512.softmask.one-line.fa.txt" for tag "SEQUENCE_TEMPLATE=". Other parameters leaves with default values.
6. Run Primer3 to design primers: "../bin/primer3-2.4.0/src/primer3_core < input.txt > output.txt."
7. Manually review the "output.txt" primer file and select the primer pairs with 3′ end (*see* **Note 7**) with no lowercase bases. In my test for this protocol, the 3′ end of primer 9 (red box in Fig. 3) has no lowercase letters. The sequences of primer 9 are: forward GAAGTggGTttTGTCGTGCC, reverse TCaGCAGC CAAAACACAAGC.

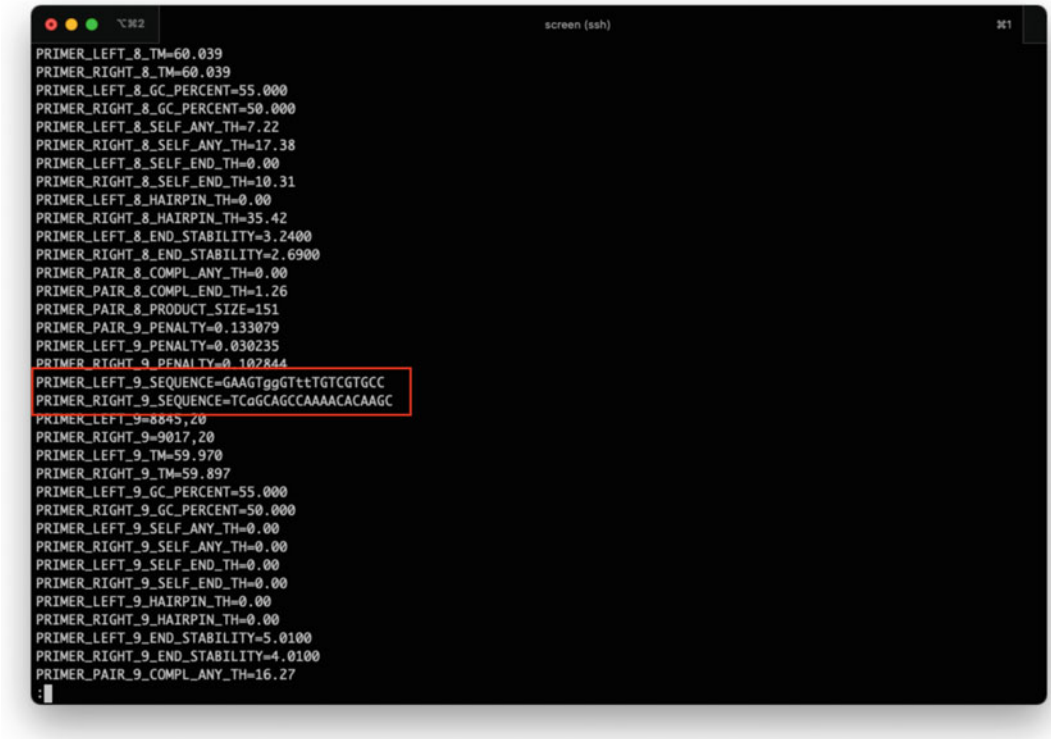

Fig. 3 Primer3 output

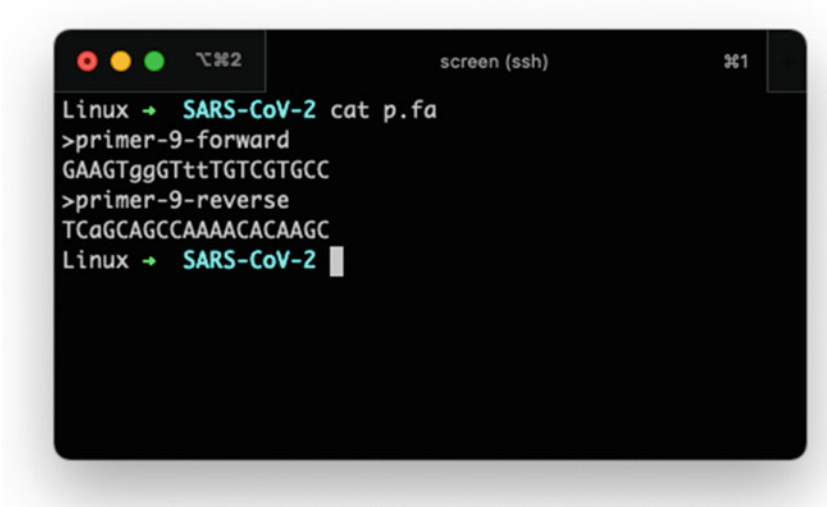

Fig. 4 Primers in FASTA format

3.3 Evaluate Primers

We use MFEprimer to analyze whether this pair of primers can amplify all the genome sequences in SARS-CoV-2 and whether it can amplify the human genome and other influenza viruses; these databases are also called background databases (*see* **Note 8**).

1. Prepare primers in the FASTA format shown in Fig. 4.
2. Index SARS-CoV-2 database: mfeprimer index -i data/SARS-CoV-2-20200927-17111.fasta.
3. Index human genome database: mfeprimer index -i data/hg19.fa.
4. Index influenza virus: mfeprimer index -i data/influenza.fna.
5. Run MFEprimer to check the coverage of this pair of primers on the SARS-CoV-2 database: mfeprimer -d data/SARS-CoV-2-20200927-17111.fasta -i p.fa -S 300 -t 55 --virus -o p.mfe.txt. Figure 5 shows that this pair of primers can coverage 99.87% sequences in database SARS-CoV-2-20200927-17111.fasta, while misses 23 sequences. The missed sequence ID information is automatically stored in file "p.mfe.txt.virus.Failed.txt".
6. To check whether this primer pair failed on these missed records. Firstly, copy and paste the first amplicon sequence in FASTA format from file "p.mfe.txt" into a file named "amp.fa" (Fig. 6). Secondly, run command "cat data/SARS-CoV-2-20200927-17111.fasta | seqkit grep -n -f p.mfe.txt.virus.Failed.txt > p.mfe.txt.virus.Failed.fa" to get the missed sequences and saved as file "p.mfe.txt.virus.Failed.fa." Thirdly, run command "blat p.mfe.txt.virus.Failed.fa amp.fa amp.fa.

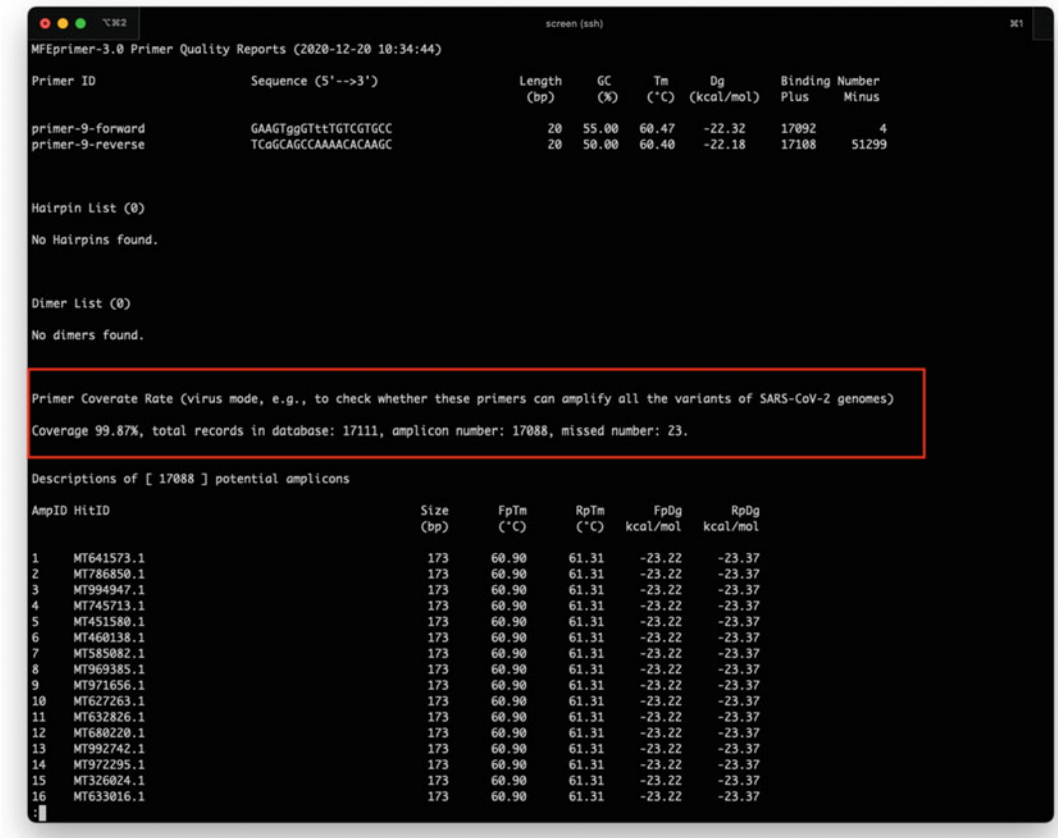

Fig. 5 Primer coverage analysis against SARS-CoV-2 database downloaded on September 27, 2020

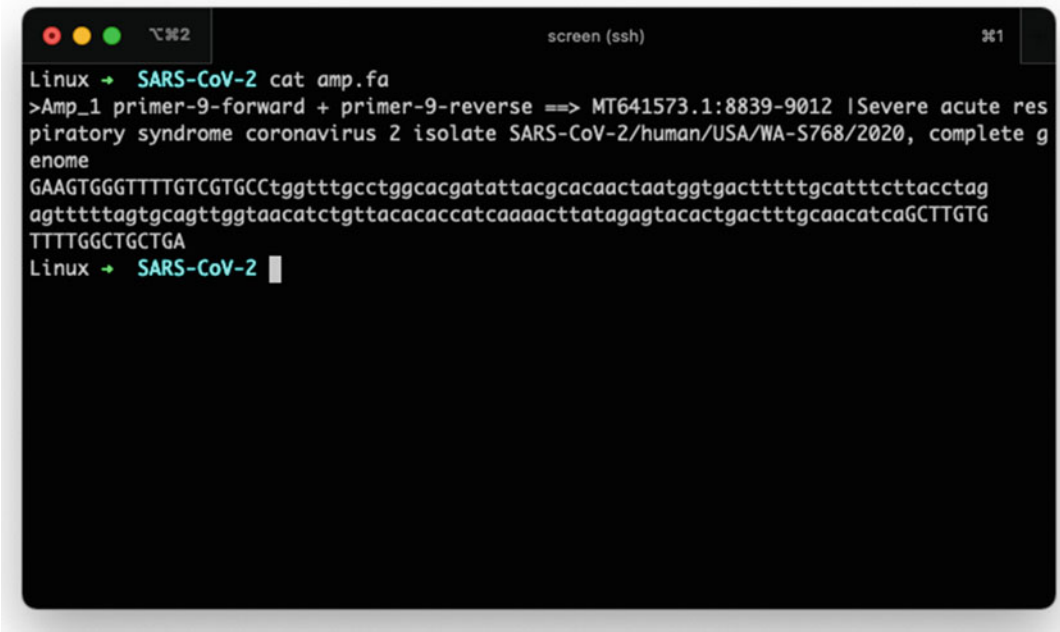

Fig. 6 First amplicon sequence in FASTA format

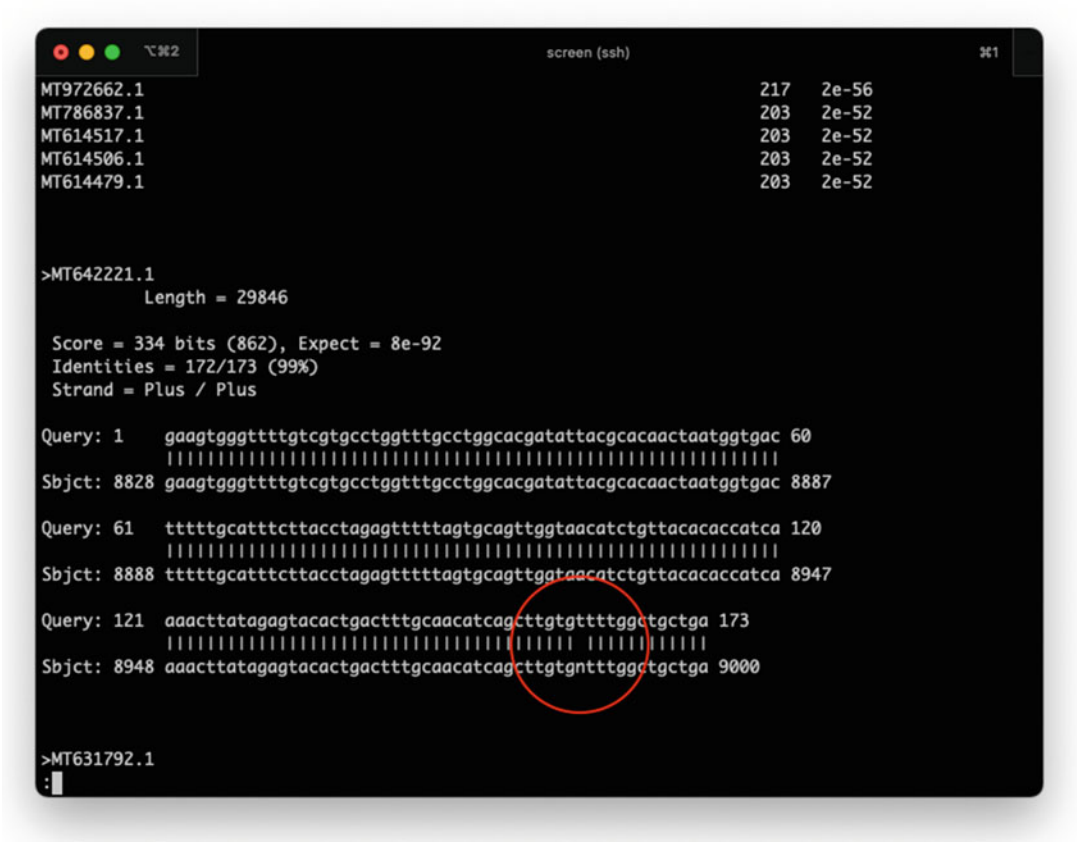

Fig. 7 Sequence alignment result shows that the missed sequences have N bases in primer binding regions shown in the red circle

blat.txt -out=blast" to do the sequence alignment. Figure 7 shows that the missed sequences have N bases in primer binding regions (shown in red circle).

7. Run MFEprimer to check the specificity of primers against human and influenza virus database with the command: mfeprimer -d data/hg19.fa -d data/influenza.fna -i p.fa -S 300 -t 55 -o p-against-human-influenza.mfe.txt. And the data (Fig. 8) shows that there are no nonspecific amplicons found on these two databases.

3.4 Re-evaluate Primers when Database Growing

In order to simulate the evolution of the SARS-CoV-2 genome, we used the database downloaded on November 11, 2020, which has 26,456 complete SARS-CoV-2 genome sequences. If a new mutation occurs at the primer binding site, the primer will not be able to amplify these variant sequences, so the coverage rate will decrease.

Run the command "mfeprimer -d data/SARS-CoV-2-20201111-26456.fa -S 300 -t 55 -i p.fa -o p.SARS-CoV-2-20201111.mfe.txt --virus". Figure 9 shows that the sequence

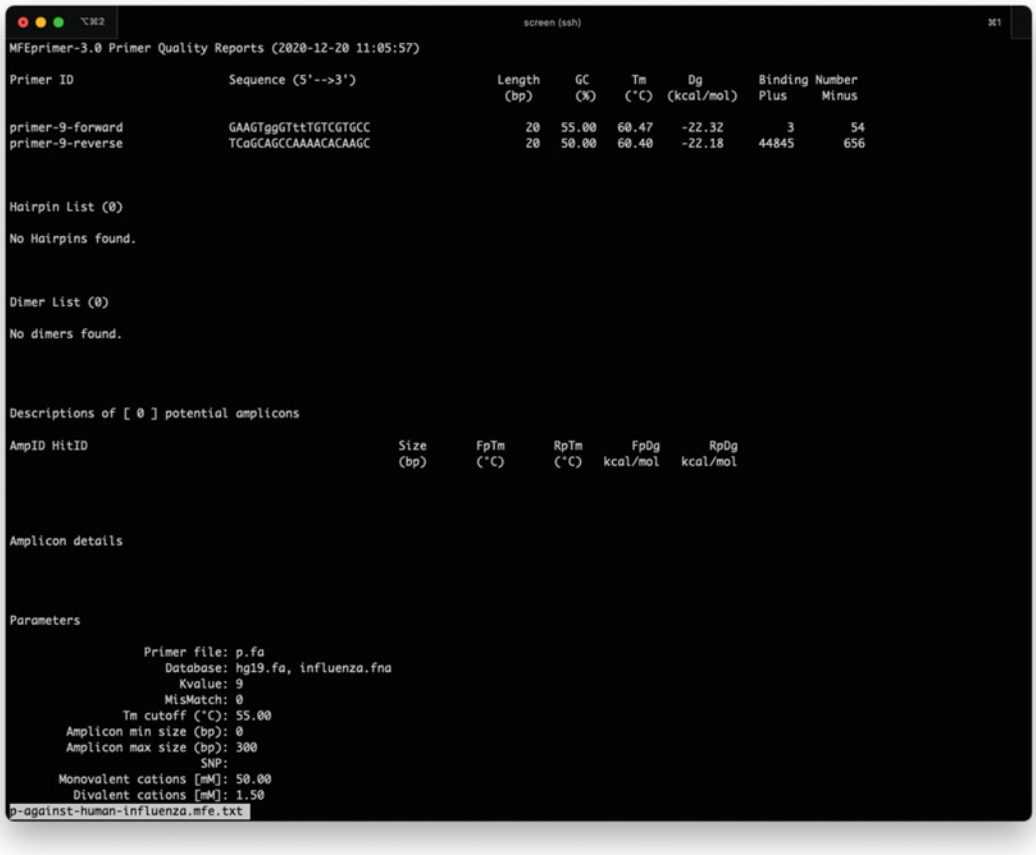

Fig. 8 Primer specificity analysis against human and influenza virus database

number has arisen to 26,456, and the coverage rate is still larger than 99%. However, the missed sequences now have 105 records. Repeat **step 6** in Subheading 3.3 to check the failure reason. If not N bases but new mutations occur, redesign primers may be needed. The redesign process starts with identifying the mutations (Subheading 3.1) for the newly added sequences and ends in this step.

4 Notes

1. Linux is an open-source Unix-like operating system. Distributions like Cent OS and Ubuntu are popular in the bioinformatics field.
2. FASTA format is a text-based format for representing nucleotide sequences, in which base pairs are represented using single-letter codes. A sequence in FASTA format begins with a single-line description with the first letter ">", followed by lines of

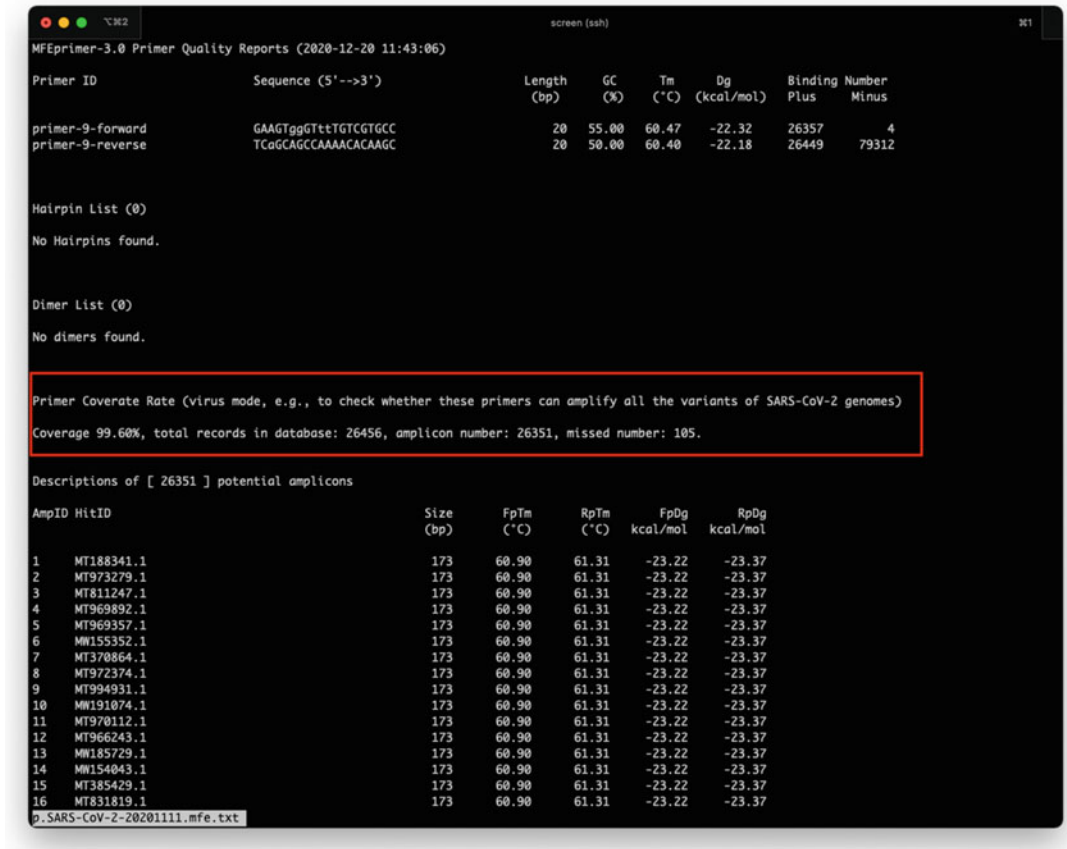

Fig. 9 Primer coverage analysis against SARS-CoV-2 database downloaded on November 11, 2020

sequence data. This format can be easily manipulated by computer languages like Python (https://www.python.org/) and Go (https://golang.org/).

3. Bioconda (https://bioconda.github.io/) is a channel for the conda (https://conda.io/en/latest/index.html) package manager specializing in bioinformatics software. For example, installing bcftools from conda is simply running the command: conda install -c bioconda bcftools (https://anaconda.org/bioconda/bcftools).
4. The variant call format (VCF) is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions, and structural variants, together with rich annotations.
5. Primer3 option PRIMER_ LOWERCASE _MASKING will reject candidate primers with lowercase letter exactly at 3′ end.
6. The SARS-CoV-2 is an RNA virus and is prone to degradation, so the amplified product should not exceed 200 bp.

7. Lowercase means there is a point mutation here. Mutations in the primers can cause PCR to fail. The mutation at the 3′ end is much worse than the mutation at the 5′ end. Therefore, try to choose primers that do not have lowercase bases. If it is not possible, choose primers that do not have lowercase bases for at least the last five bases.
8. The background database refers to the DNA sequences other than the target DNA. For example, if we design primers for SARS-CoV-2, then the human genome DNA in the sample is the background database.

Acknowledgments

This work was supported by the Research Foundation of iGeneTech [2019SX001, 2020SX001].

References

1. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS (2011) Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol 8 (2):270–279. https://doi.org/10.4161/rna.8.2.15013
2. Islam MR, Hoque MN, Rahman MS, Alam A, Akther M, Puspo JA, Akter S, Sultana M, Crandall KA, Hossain MA (2020) Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci Rep 10(1):14004. https://doi.org/10.1038/s41598-020-70812-6
3. Xing Y, Li X, Gao X, Dong Q (2020) MicroGMT: a mutation tracker for SARS-CoV-2 and other microbial genome sequences. Front Microbiol 11:1502. https://doi.org/10.3389/fmicb.2020.01502
4. Qu W, Zhang C (2015) Selecting specific PCR primers with MFEprimer. Methods Mol Biol 1275:201–213. https://doi.org/10.1007/978-1-4939-2365-6_15
5. Jalandra R, Yadav AK, Verma D, Dalal N, Sharma M, Singh R, Kumar A, Solanki PR (2020) Strategies and perspectives to develop SARS-CoV-2 detection methods and diagnostics. Biomed Pharmacother 129:110446. https://doi.org/10.1016/j.biopha.2020.110446
6. Nagy A, Jirinec T, Cernikova L, Jirincova H, Havlickova M (2015) Large-scale nucleotide sequence alignment and sequence variability assessment to identify the evolutionarily highly conserved regions for universal screening PCR assay design: an example of influenza A virus. Methods Mol Biol 1275:57–72. https://doi.org/10.1007/978-1-4939-2365-6_4
7. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, Schaffer AA, Brister JR (2017) Virus variation resource—improved response to emergent viral outbreaks. Nucleic Acids Res 45(D1): D482–D490. https://doi.org/10.1093/nar/gkw1065
8. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324
9. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34 (18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191
10. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40(15):e115. https://doi.org/10.1093/nar/gks596
11. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35:W71–W74. https://doi.org/10.1093/nar/gkm306. |ISSN 0305-1048
12. Wang K, Li H, Xu Y, Shao Q, Yi J, Wang R, Cai W, Hang X, Zhang C, Cai H, Qu W (2019) MFEprimer-3.0: quality control for PCR primers. Nucleic Acids Res 47(W1):

W610–W613. https://doi.org/10.1093/nar/gkz351

13. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12(6):996–1006. https://doi.org/10.1101/gr.229102
14. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D (2008) The influenza virus resource at the National Center for Biotechnology Information. J Virol 82(2):596–601. https://doi.org/10.1128/JVI.02005-07
15. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993. https://doi.org/10.1093/bioinformatics/btr509
16. Shen W, Le S, Li Y, Hu F (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11 (10):e0163962. https://doi.org/10.1371/journal.pone.0163962
17. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033
18. Quinlan AR (2014) BEDTools: The Swiss-Army tool for genome feature analysis. Curr Protoc Bioinformatics 47:11-12–11-34. https://doi.org/10.1002/0471250953.bi1112s47
19. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12(4):656–664. https://doi.org/10.1101/gr.229202
20. Gruning B, Dale R, Sjodin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Koster J, Bioconda T (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 15 (7):475–476. https://doi.org/10.1038/s41592-018-0046-7
21. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis G (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330