Bioinformatics, 37(7), 2021, 1024–1025
doi: 10.1093/bioinformatics/btaa710
Advance Access Publication Date: 10 August 2020
Applications Note

Sequence analysis

A public website for the automated assessment and validation of SARS-CoV-2 diagnostic PCR assays

Po-E Li [†], Adán Myers y Gutiérrez [†], Karen Davenport, Mark Flynn, Bin Hu, Chien-Chi Lo, Elais Player Jackson, Migun Shakya, Yan Xu, Jason D. Gans * and Patrick S. G. Chain *

Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Arne Elofsson
Received on May 22, 2020; revised on July 20, 2020; editorial decision on July 30, 2020; accepted on August 4, 2020

Abstract

Summary: Polymerase chain reaction-based assays are the current gold standard for detecting and diagnosing SARS-CoV-2. However, as SARS-CoV-2 mutates, we need to constantly assess whether existing PCR-based assays will continue to detect all known viral strains. To enable the continuous monitoring of SARS-CoV-2 assays, we have developed a web-based assay validation algorithm that checks existing PCR-based assays against the ever-expanding genome databases for SARS-CoV-2 using both thermodynamic and edit-distance metrics. The assay-screening results are displayed as a heatmap, showing the number of mismatches between each detection and each SARS-CoV-2 genome sequence. Using a mismatch threshold to define detection failure, assay performance is summarized with the true-positive rate (recall) to simplify assay comparisons.

Availability and implementation: The assay evaluation website and supporting software are Open Source and freely available at https://covid19.edgebioinformatics.org/#/assayValidation, https://github.com/jgans/thermonucleotideBLAST and https://github.com/LANL-Bioinformatics/assay_validation.

Contact: jgans@lanl.gov or pchain@lanl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Many aspects of the control, management and treatment responses to the global COVID-19 pandemic require accurate detection of its causative agent, SARS-CoV-2. To address this challenge, research groups around the world have developed polymerase chain reaction (PCR)-based assays to detect SARS-CoV-2 genomic RNA (Supplementary Table S1).

The impact of SARS-CoV-2 genetic drift on the ability of PCR-based assays to successfully detect target sequences is a concern. To address this concern, we have developed a web-based application that monitors existing SARS-CoV-2 PCR-based assays that are in use around the world and provides a visual summary of assay performance. Both the acquisition of new genomes and the assay validation process are automated, so that assays are checked and displayed daily to give near real-time results.

2 Implementation

The core of the validation algorithm is the ThermonucleotideBLAST (Gans and Wolinsky, 2008) *in silico* PCR screening tool. Publicly available assays are used as queries in ThermonucleotideBLAST and searched against a target database of SARS-CoV-2 genomes from the Global Initiative on Sharing All Influenza Data (GISAID) (Shu and McCauley, 2017) and GenBank (Clark *et al.*, 2016).

Sequences are accessed daily from these databases and filtered to exclude any that are less than 29 kilobases or are pangolin-SARS and bat-SARS. For sequences found in both databases, only the GISAID version is retained. Predicted false negatives are defined as assay/target combinations that have either (i) one or more oligo/target pairwise alignments with three or more mismatches, (ii) one or more predicted oligo/target melting temperatures $<40°C$ or (iii) one or more mismatches in the last two 3′ positions of a primer that are reported by Li *et al.* (2004) to inhibit detection by *increasing* detection $Ct \geq 2$. True positives are defined as any assay–target combination not predicted to be a false negative. Since all of the included assays are intended to detect SARS-CoV-2 and false positives are not predicted, assay performance is quantified by the recall (defined as the number of true positives divided by the sum of true positives and predicted false negatives).

Per-assay recall values are summarized in Figure 1A. The assays with the best recall rates are shown in a bar chart, which also displays detailed mismatch counts. The total mismatch and failure

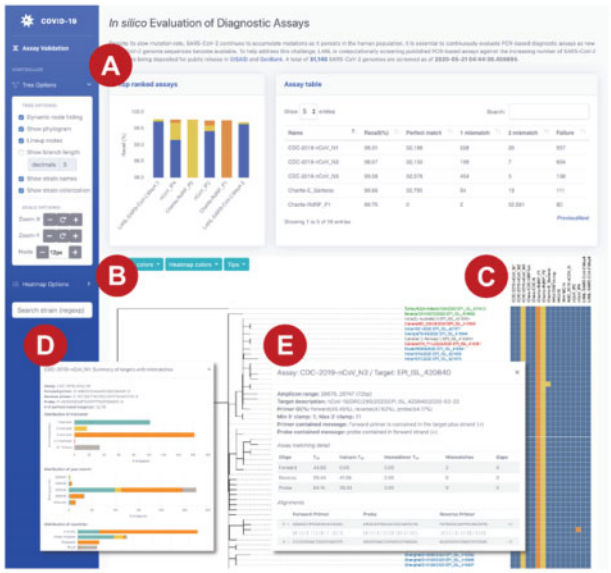

Fig. 1. Visualization of *in silico* evaluation of diagnostic assays. (**A**) Dashboard including a bar chart and table with per-assay recall and mismatch counts; (**B**) phylogenetic tree created from high-quality genomes color-labeled by continent; (**C**) a heatmap display of assay assessment per assay per genome; (**D**) assay details and statistics of genomes with mismatches; and (**E**) detailed assay evaluation results, including alignments and thermodynamic information

results are summarized in the per-assay table of aggregated data. Selecting any bar in the chart or assay in the table will display additional information on the distribution of targets with mismatches (Fig. 1D).

The phylogenetic tree (Fig. 1B) is created using PhaME (Shakya *et al.*, 2020) and 'high-quality' GISAID genomes (<1% *N*s and <0.05% unique mutations). The leaves on the tree are represented by the genome labels and color-coded by geographic location. Mousing over the genome labels displays metadata associated with the sample. Identical SARS-CoV-2 sequences are clustered and represented as collapsed branches in the tree. The heatmap (Fig. 1C), color-coded to indicate the number of mismatches, shows analysis of every combination of assay and SARS-CoV-2 genome sequence. Selecting an individual cell of the heatmap displays detailed pairwise alignment information (Fig. 1E). This visualization is rendered using a custom PhyD3 phylogenetic tree viewer (Kreft *et al.*, 2017).

3 Discussion

Few other public resources exist for assessing the performance of PCR-based SARS-CoV-2 assays. GISAID, one of the primary repositories for SARS-CoV-2 genomes, provides a high-level summary of PCR-based assay performance for registered users. However, this information is provided in the form of a static image with only a limited amount of information. The virological.org website provides static tables summarizing the high-level performance of PCR assays that have been periodically uploaded (Holland *et al.*, 2020). Unlike these resources, the web-based application presented here provides a more detailed and interactive view of molecular assay performance that is updated regularly with recently deposited genomes (>66K as of July 15, 2020).

The heatmap-phylogeny view reveals patterns in predicted assay performance, including mismatches for the Charité RdRP assays (Corman *et al.*, 2020; Vogels *et al.*, 2020) that were originally developed for testing SARS and/or SARS-related bat coronaviruses (Fig. 1C and Supplementary Fig. S1). A different pattern, previously noted by Vogels *et al.* (2020), is seen within a subset of phylogenetically related strains due to a mismatch in the USA CDC N3 assay (CDC, 2020) (Supplementary Fig. S1). As genomics continues to be used for understanding pathogen outbreaks, resources, such as the one provided by this website, may help in the early identification of potential assay concerns, and provide guidance on alternate assay designs early on, to mitigate current assays that may be eroding.

Acknowledgements

We acknowledge the authors and originators of sequences from the submitting laboratories who have contributed to the GISAID database.

Funding

This research was supported by Los Alamos National Laboratory [20200732ER], by Defense Threat Reduction Agency [CB10152, CB10623]; and by the Department of Energy Office of Science [KP160101], through the National Virtual Biotechnology Laboratory, a consortium of Department of Energy national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act. Hosting of edgebioinformatics.org is provided by CyVerse, which is supported by the National Science Foundation [DBI-0735191, DBI-1265383 and DBI-1743442].

Conflict of Interest: none declared.

References

CDC (2020) *2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel for Emergency Use Only: Instructions for Use*. Centers for Disease Control and Prevention. https://www.fda.gov/media/134922/download, (11 May 2020, date last accessed).

Clark,K. *et al.* (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.

Corman,V. *et al.* (2020) *Diagnostic Detection of 2019-nCoV by Real-Time RT-PCR. Protocol.* https://www.who.int/docs/default-source/coronaviruse/protocol-v2-1.pdf (11 May 2020, date last accessed).

Gans,J.D. and Wolinsky,M. (2008) Improved assay-dependent searching of nucleic acid sequence databases. *Nucleic Acids Res.*, **36**, 1–5.

Holland,M. *et al.* (2020) BioLaboro: a bioinformatics system for detecting molecular assay signature erosion and designing new assays in response to emerging and reemerging pathogens. https://doi.org/10.1101/2020.04.08.031963.

Kreft,Ł. *et al.* (2017) PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, **33**, 2946–2947.

Li,B. *et al.* (2004) Genotyping with TaqMAMA. *Genomics*, **83**, 311–320.

Shakya,M. *et al.* (2020) Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Sci. Rep.*, **10**, 1723.

Shu,Y. and McCauley,J. (2017) GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.*, **22**, 30494.

Vogels,C.B. *et al.* (2020) Analytical sensitivity and efficiency comparisons of SARS-COV-2 qRT-PCR assays. primer-probe sets. medRxiv 2020.03.30.20048108; doi: 10.1101/2020.03.30.20048108.